# Probability and Geometry on Groups

## Lecture notes for a graduate course

Gábor Pete
Technical University of Budapest
http://www.math.bme.hu/~gabor

April 3, 2024
Work in progress. Comments are welcome.

**Abstract**

These notes have grown (and are still growing) out of two graduate courses I gave at the University of Toronto. The main goal is to give a self-contained introduction to several interrelated topics of current research interests: the connections between

1) coarse geometric properties of Cayley graphs of infinite groups;
2) the algebraic properties of these groups; and
3) the behaviour of probabilistic processes (most importantly, random walks, harmonic functions, and percolation) on these Cayley graphs.

I try to be as little abstract as possible, emphasizing examples rather than presenting theorems in their most general forms. I also try to provide guidance to recent research literature. In particular, there are presently over 250 exercises and many open problems that might be accessible to PhD students. It is also hoped that researchers working either in probability or in geometric group theory will find these notes useful to enter the other field.

# Contents

# Preface

These notes have grown (and are still growing) out of two graduate courses I gave at the University of Toronto: *Probability and Geometry on Groups* in the Fall of 2009, and *Percolation in the plane, on $\mathbb{Z}^d$, and beyond* in the Spring of 2011. I am still adding material and polishing the existing parts, so at the end I expect it to be enough for two semesters, or even more. Large portions of the first drafts were written up by the nine students who took the first course for credit: Eric Hart, Siyu Liu, Kostya Matveev, Jim McGarva, Ben Rifkind, Andrew Stewart, Kyle Thompson, Lluís Vena, and Jeremy Voltz — I am very grateful to them. That first course was completely introductory: some students had not really seen probability before this, and only few had seen geometric group theory. Here is the course description:

> Probability is one of the fastest developing areas of mathematics today, finding new connections to other branches constantly. One example is the rich interplay between large-scale geometric properties of a space and the behaviour of stochastic processes (like random walks and percolation) on the space. The obvious best source of discrete metric spaces are the Cayley graphs of finitely generated groups, especially that their large-scale geometric (and hence, probabilistic) properties reflect the algebraic properties. A famous example is the construction of expander graphs using group representations, another one is Gromov's theorem on the equivalence between a group being almost nilpotent and the polynomial volume growth of its Cayley graphs. The course will contain a large variety of interrelated topics in this area, with an emphasis on open problems.

What I had originally planned to cover turned out to be ridiculously much, so a lot had to be dropped, which is also visible in the present state of these notes. The main topics that are still missing are Gromov-hyperbolic groups and their applications to the construction of interesting groups, metric embeddings of groups in Hilbert spaces, more on the construction and applications of expander graphs, more on critical spatial processes in the plane and their scaling limits, and a more thorough study of Uniform Spanning Forests and $\ell^2$-Betti numbers — I am planning to improve the notes regarding these issues soon. Besides research papers I like, my primary sources were [DrKa09], [dlHar00] for geometric group theory and [LyPer16], [Per04], [Woe00] for probability. I did not use more of [HooLW06], [Lub94], [Wil09] only because of the time constraints. There are proofs or even sections that follow rather closely one of these books, but there are always differences in the details, and the devil might be in those. Also, since I was a graduate student of Yuval Peres not too long ago, several parts of these notes are strongly influenced by his lectures. In particular, Chapter 9 contains paragraphs that are almost just copied from some unpublished notes of his that I was once editing. There is one more recent book, [Gri10], whose first few chapters have considerable overlap with the more introductory parts of these notes, although I did not look at that book before having finished most of these notes. Anyway, the group theoretical point of view is missing from that book entirely.

With all these books available, what is the point in writing these notes? An obvious reason is that it is rather uncomfortable for the students to go to several different books and start reading them somewhere from their middle. Moreover, these books are usually for a bit more specialized audience, so either nilpotent groups or martingales are not explained carefully. So, I wanted to add my favourite explanations and examples to everything, and include proofs I have not seen elsewhere in the literature. And there was a very important goal I had: presenting the material in constant conversation between the probabilistic and geometric group theoretical ideas. I hope this will help not only students, but also researchers from either field get interested and enter the other territory.

There are presently over 200 exercises, in several categories of difficulty: the ones without any stars should be doable by everyone who follows the notes, though they do not have to be entirely trivial; * means it should be a challenge for the reader; ** means that I think I would be able to do it, but it would be a challenge for me; *** means it is an open problem. Part of the grading scheme in the course (adopted from Yuval Peres) was to submit exercise solutions worth 8 points, where each exercise was worth $2^{\#\text{ of stars}}$.

> IT IS VERY IMPORTANT THAT YOU AT LEAST *READ* ALL THE EXERCISES, AND PREFERABLY YOU ALSO TRY TO *SOLVE* THEM. IF I WRITE DOWN A PROOF AND YOU READ IT, EVEN IF YOU UNDERSTAND IT, YOU WILL FORGET. IF YOU DO IT YOURSELF, YOU WILL REMEMBER.

There are also conjectures and questions in the notes — the difference compared to the *** exercises is that, according to my knowledge or feeling, the *** exercises have not been worked on yet thoroughly enough, so I want to encourage the reader to try and attack them. Of course, this does not necessarily mean that all conjectures are hard, neither that any of the *** exercises are doable.... Anyway, for a PhD topic, I would personally suggest starting with the *** exercises.

Besides my students and the books mentioned above, I am grateful to Ági Backhausz, Alex Bloemendal, Damien Gaboriau, Aranka Hrušková, Gady Kozma, Russ Lyons, Sébastien Martineau, Péter Mester, Yuval Peres, Mark Sapir, Andreas Thom, Ádám Timár, László Márton Tóth, Todor Tsankov and Bálint Virág for conversations and comments.

## Notation

If $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$ are two sequences of positive numbers, then, as $n \to \infty$:

$$f_n \ll g_n \quad \Longleftrightarrow \quad f_n = o(g_n) \quad \Longleftrightarrow \quad \frac{f_n}{g_n} \to 0\,,$$

$$f_n \asymp g_n \quad \Longleftrightarrow \quad f_n = O(g_n) \text{ and } g_n = O(f_n) \quad \Longleftrightarrow \quad 0 < c < \frac{f_n}{g_n} < C < \infty \text{ for all } n\,,$$

$$f_n \sim g_n \quad \Longleftrightarrow \quad \frac{f_n}{g_n} \to 1\,.$$

# 1 Basic examples of random walks

{s.1}

Random walks on infinite graphs is one of the most classical examples of how the geometry of the underlying space influences the behavior of stochastic processes on that space. And these mathematical ideas have wide applicability, since random walks serve as models for a wide range of diffusion processes: heat or electric conductance in physical materials, genetic or epidemic spreading in biological networks, information diffusion in social networks, and so on.

## 1.1 $\mathbb{Z}^d$ and $\mathbb{T}_d$, recurrence and transience, Green's function and spectral radius

{ss.ZdTd}

Consider a connected, infinite graph $G = (V, E)$ that is locally finite: i.e., each vertex $x \in V(G)$ has a finite number of neighbours, known as the **degree** $\deg(x)$; all the graphs in this book will be locally finite. In **(discrete time) simple random walk (SRW)** on $G$, we take a starting vertex $X_0 = o \in V(G)$, and then, given all the locations $X_0, X_1, \dots, X_n \in V(G)$ by time $n$, the next step is taken to one of the neighbours of $X_n$, uniformly at random, independently of all previous steps. Written in a formula,

$$\mathbf{P}_o[\, X_{n+1} = y \mid X_0, \dots, X_n \,] = \frac{\mathbf{1}_{\{(X_n,y)\in E(G)\}}}{\deg(X_n)}\,,$$

where the subscript $o$ shows where the walk was started.

One of the most basic properties that needs to be studied is the following.

**Definition 1.1.** *A random walk on a graph is called* **recurrent** *if the starting vertex is visited again, and hence visited infinitely often, with probability one. That is, using the positive hitting time $\tau_x^+ := \inf\{n \geq 1 : X_n = x\}$:*

$$\mathbf{P}_o\big[\tau_o^+ < \infty\big] = \mathbf{P}_o\big[X_n = o \text{ infinitely often}\big] = 1\,.$$

*Otherwise, the walk is called* **transient***, and $o$ is almost surely visited only finitely many times.*

Let us remark here that if we want to prove rigorously the intuitively obvious identity in the definition, namely, that coming back once almost surely implies that we also come back infinitely often, then we need the so-called **strong Markov property** for simple random walk: given the present $X_\tau$, the future $X_{\tau+1}, X_{\tau+2}, \dots$ is conditionally independent of the past $X_0, X_1, \dots, X_{\tau-1}$ for any stopping time $\tau$, not just for deterministic times. This is intuitively obvious, and is easy to prove by conditioning on the countable number of possibilities $\tau$ can take. For complicated continuous time and continuum state space Markov processes this is less trivial, sometimes even false.

{ex.transfinit

▷ **Exercise 1.1.** *Show that if $G(V, E)$ is a connected graph, and simple random walk started at some $o \in V$ visits $o$ infinitely often almost surely, then the walk started at any $x \in V$ visits any given $y \in V$ infinitely often, almost surely. Consequently, recurrence is a property solely of the graph. In a transient graph, the walk visits any given finite set only finitely many times.*

The result that started the area of random walks on groups is the following:

{t.Polya}

**Theorem 1.2** (Pólya 1921)**.** *Simple random walk on $\mathbb{Z}^d$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$. In fact, the so-called "on-diagonal heat-kernel decay" is $p_{2n}(o, o) = \mathbf{P}_o[\, X_{2n} = o \,] \asymp C_d n^{-d/2}$.*

*Proof.* Let us first see how the quantitative second sentence implies the qualitative first sentence. The trick is that

$$\mathbf{E}_o\big[\,\#\text{ of visits to } o\,\big] = \mathbf{E}_o\Big[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n = o\}}\Big] = \sum_{n=0}^{\infty} p_n(o, o)\,,$$

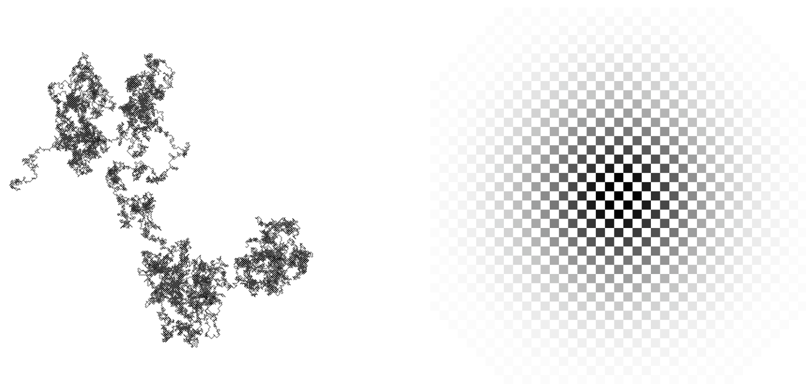


Figure 1.1: A SRW trajectory on $\mathbb{Z}^2$ and the distribution of the particle after 100 steps. {f.2dSRW}

where the interchange of the two integrals (the infinite sum and the expectation) was possible by Fubini, since the indicator variables are non-negative functions. Thus, the quantitative bound says that

$$\mathbf{E}_o\big[\,\#\text{ of visits to } o\,\big] < \infty \quad \text{if and only if} \quad d > 2.$$

And why is this equivalent to transience? Well, if the walk is recurrent, then the number of visits to $o$ is almost surely infinite, hence the expectation is also infinite. On the other hand, if the walk is transient, $\mathbf{P}_o[\,\tau_o^+ = \infty\,] = q > 0$, then

$$\mathbf{P}_o[\,\#\text{ of visits to } o = k\,] = (1-q)^{k-1}q \qquad \text{for } k = 1, 2, \dots,$$

a geometric random variable, with expectation $\mathbf{E}_o[\,\#\text{ of visits to } o\,] = 1/q$, which is finite.

For the proof of the quantitative bound, the key first step is the $d = 1$ case, which we state as a separate lemma. It is straightforward to deduce it from Stirling's formula $n! \sim \sqrt{2\pi n}\,(n/e)^n$, or alternatively, it is a consequence of the Bernoulli Local Central Limit Theorem; see [Fel68, Sections II.9 and VII.3] and [Dur10, Sections 3.1 and 3.5] for these theorems and their relationship with each other.

{l.Stirling}

**Lemma 1.3.** *For the one-dimensional simple random walk* $0 = Y_0, Y_1, Y_2, \dots$*, we have*

$$\mathbf{P}[\,Y_{2n} = 0\,] = \frac{1}{2^{2n}}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}\,.$$

On $\mathbb{Z}^d$ for general $d$, Pólya used Fourier analysis. Another approach is to use Stirling's formula to estimate multinomial coefficients and execute careful summations. However, in this book we prefer more robust techniques, hence we are proposing the following rather intuitive strategy:

1) being back at the origin is equivalent to being back in all the $d$ coordinates;
2) after $2n$ steps, the number of steps in any given coordinate, by the Law of Large Numbers, is $(1+o(1))\,2n/d$ with high probability, and it should be even with probability $\sim 1/2$, hence the return probability in any given coordinate, by Lemma 1.3, should be $\sim \frac{1}{2}\sqrt{\frac{d}{2\pi n}}$;
3) the moves in the different coordinates are independent, hence we can just multiply these $d$ probabilities together to get $\sim c_d\, n^{-d/2}$.

There are several places where this naive strategy needs serious work. Firstly, in step 2), the Law of Large Numbers has a probability error (whose size depends on the precision in the $1+o(1)$ factor), which then appears in the overall return probability. This probability error certainly has to be $o(1/\sqrt{n})$ in order to have $\sqrt{\frac{d}{2\pi n}}$ as the main contribution. Similarly, the guess $1/2+o(1)$ for the probability of having an even number of steps is very intuitive, but with what error term $o(1)$ can we prove it? Lastly, in step 3), the independence is false: once we know how many steps are made in each coordinate, the steps themselves are indeed independent, but the number of steps in each coordinate are not, since they have to sum up to $2n$. This introduces dependencies: if we know only that we are back at 0 in the first coordinate, that slightly decreases the likely number of steps in that coordinate (since the fewer steps we have made, the more likely it is to be back), which increases the number of steps in the other coordinates, making it a bit less likely to be back at 0 in those.

We will make this strategy rigorous in two different ways. The first approach introduces a new process, continuous time random walk $X(t)_{t\geq 0}$, for which the problem with 3) simply disappears and the problems with 2) become doable; thus we can easily get $\mathbf{P}_o\big[\,X(t)=o\,\big] \asymp t^{-d/2}$. Unfortunately, this does not automatically yield $p_{2n}(o,o)\asymp n^{-d/2}$, but is nevertheless enough to decide about recurrence versus transience even for the original SRW. The second approach uses brute force to deal with all the problems together.

In **continuous time random walk (CTRW)** on a graph $G(V,E)$, each edge has an independent exponential clock of rate 1 (i.e., a rate 1 Poisson process of clock rings), and the walker always immediately takes the edge emanating from its current location whose clock first rings. Note that, after each step, it does not matter if we restart the clocks emanating from the new vertex or not, due to the memoryless property of the exponential distribution. This also implies that the sequence of vertices visited by this random walker follows just discrete time SRW on the graph, since, among $k$ iid exponential variables, each has probability $1/k$ to be the smallest. That is, CTRW is just SRW with a random time change: an independent waiting time at each vertex $v\in V$ of distribution $\mathsf{Expon}(\deg(v))$, the minimum of $\deg(v)$ many iid $\mathsf{Expon}(1)$ variables.

Now, the point of considering CTRW is that the $d$ coordinates of CTRW on $\mathbb{Z}^d$ are just iid CTRW's on $\mathbb{Z}$; i.e., the dependency problem between the coordinates is now gone. The reason for this is a basic property of Poisson point processes: if we take a PPP of intensity $2d$, and label each point by one of the $d$ coordinates and one of the two directions, $\pm e_i$, $i=1,\dots,d$, with iid uniform choices, then we get, in distribution, the union of $d$ independent PPP processes of intensity 2, in each of which the points are labeled $\pm$ in an iid uniform manner.

Therefore, for this CTRW process $(X(t))_{t\geq 0}=\big(X_1(t),\dots,X_d(t)\big)_{t\geq 0}$ on $\mathbb{Z}^d$, started at $X(0)=o=\underline{0}$, we have

$$\mathbf{P}_o\big[\,X(t)=o\,\big]=\prod_{i=1}^{d}\mathbf{P}_0[\,X_i(t)=0\,]=\mathbf{P}_0[\,X_1(t)=0\,]^d\,. \qquad (1.1) \quad \texttt{\{e.prod\}}$$

We now want to express the 1-dimensional CTRW return probability $\mathbf{P}_0[\,X_1(t)=0\,]$ using the SRW result we had in Lemma 1.3. For this, we need the random variable $N_1(t)$, the number of steps made in the first coordinate by time $t$. This has a distribution $\mathsf{Poisson}(2t)$, with mean and variance $2t$, hence Chebyshev's inequality tells us that

$$\mathbf{P}\big[\,N_1(t)\notin(t,3t)\,\big]=\mathbf{P}\big[\,|N_1(t)-\mathbf{E}N_1(t)|\geq t\,\big]\leq\frac{2t}{t^2}=\frac{2}{t}\,. \qquad (1.2) \quad \texttt{\{e.ChebyN\}}$$

Moreover, for any $t\geq 0$,

$$\mathbf{P}\big[\,N_1(t)\text{ is even}\,\big]=1/2+e^{-4t}/2. \qquad (1.3) \quad \texttt{\{e.eveN\}}$$

Convergence to 1/2 is intuitively quite clear, and our exact formula can be verified by a standard continuous time Markov chain argument, as follows. Let $p(t)$ denote the probability in question. We have

$$p(t+\epsilon) = (1-2\epsilon)\,p(t) + 2\epsilon\,(1-p(t)) + O(\epsilon^2),$$

for small $\epsilon > 0$, since $N_1(t+\epsilon)$ can be even in several ways: either $N_1(t)$ was even and there was no jump in $(t, t+\epsilon)$, which altogether happens with probability $p(t)\,e^{-2\epsilon}$, or $N_1(t)$ was odd and there was a single jump in $(t, t+\epsilon)$, or there were at least two jumps, which has probability $O(\epsilon^2)$. Letting $\epsilon \to 0$, we find that $p(t)$ satisfies the differential equation

$$p'(t) = 2 - 4p(t)\,,$$

with initial condition $p(0) = 1$, which gives (1.3).

Combining (1.2) and (1.3), we get that $\mathbf{P}\big[\, N_1(t) \in (t, 3t) \cap 2\mathbb{N} \,\big] = 1/2 - O(1/t)$ as $t \to \infty$. The importance of this, of course, is that by Lemma 1.3 we have $\mathbf{P}_0[\, X_1(t) = 0 \mid N_1(t) \,] \asymp 1/\sqrt{N_1(t)}$ if $N_1(t)$ is even and larger than 0. Thus, for $t > 1$,

$$\begin{aligned}
\mathbf{P}_0\big[\, X_1(t) = 0 \,\big] &= \mathbf{P}_0\big[\, X_1(t) = 0,\ N_1(t) \in (t, 3t) \cap 2\mathbb{N} \,\big] + \mathbf{P}_0\big[\, X_1(t) = 0,\ N_1(t) \notin (t, 3t) \,\big] \\
&= \mathbf{P}_0\big[\, X_1(t) = 0 \;\big|\; N_1(t) \in (t, 3t) \cap 2\mathbb{N} \,\big]\, (1/2 - O(1/t)) + O(1/t) \\
&\asymp \frac{1}{\sqrt{t}} + O(1/t)\,.
\end{aligned}$$

Substituting this back to (1.1), we get that $\mathbf{P}_o\big[\, X(t) = o \,\big] \asymp t^{-d/2}$ for $t > 1$. For $t \le 1$, we obviously have $\mathbf{P}_o[\, X(t) = o \,] \in (0, 1)$. As in the discrete time case, we can now calculate an expectation: for $t > 1$,

$$\mathbf{E}_o\big[\, \text{Lebesgue measure of time in } [0,t] \text{ spent at } o \,\big] = \int_0^t \mathbf{P}_o[\, X(s) = o \,]\, ds \asymp \begin{cases} \sqrt{t} & \text{if } d = 1\,, \\ \log t & \text{if } d = 2\,, \\ 1 & \text{if } d > 2\,. \end{cases} \tag{1.4}$$ {e.spent}

We still have to translate these bounds back to the recurrence of the SRW. Let $N = N(\infty)$ be the total number of visits to $o$. At the $i^{\text{th}}$ visit, the time spent at $o$ is a random variable $\xi_i \sim \mathsf{Expon}(2d)$, independent from each other and from $N$. Hence, the total expected time spent at $o$ is

$$\mathbf{E}_o\Big[ \sum_{i=1}^{N} \xi_i \Big] = \mathbf{E}_o\Big[ \mathbf{E}_o\Big[ \sum_{i=1}^{N} \xi_i \;\Big|\; N \Big] \Big] = \mathbf{E}_o\Big[ N \,\frac{1}{2d} \Big] = \mathbf{E}_o[\, N \,]\, \frac{1}{2d}\,.$$

Thus, the expected amount of time spent at $o$ is finite if and only if the expected number of visits is finite, which is equivalent to transience. Hence (1.4) shows that SRW on $\mathbb{Z}^d$ is transient iff $d > 2$.

For the second approach, we will use something similar to (1.2), but stronger. The proof of this lemma will be discussed later, in Section 1.2.

{l.LD}

**Lemma 1.4.** *The following estimate holds for a d-dimensional lattice:*

$$\mathbf{P}\Big[\, \#\text{ of steps among first } n \text{ that are in the } i^{th} \text{ coordinate} \notin \Big[\frac{n}{2d}, \frac{3n}{2d}\Big] \,\Big] < C \exp(-c_d n)\,.$$

As before, we let $o = \underline{0} \in \mathbb{Z}^d$ denote the origin, $X_n = (X_n^1, \ldots, X_n^d)$ the walk, and $\underline{n} = (n_1, \ldots, n_d)$ a multi-index, with $|\underline{n}| := n_1 + \cdots + n_d$. Furthermore, let $N_n^i = N^i(X_1, \ldots, X_n)$ denote the number of steps taken in the $i^{\text{th}}$ coordinate, $i = 1, \ldots, d$. Then

$$\mathbf{P}_o[\, X_{2n} = o \,] = \mathbf{P}_o[\, X_{2n}^i = 0\ \forall i \,] = \sum_{\underline{n} : |\underline{n}| = n} \mathbf{P}_o\big[\, Y_{2n_i}^i = 0\ \forall i \;\big|\; N_{2n}^i = 2n_i\ \forall i \,\big]\, \mathbf{P}[\, N_{2n}^i = 2n_i\ \forall i \,]\,,$$

which we got by using the Law of Total Probability, and where $Y_0^i, Y_1^i, \dots$ is the $i^{\text{th}}$ coordinate of the sequence $X_0, X_1, \dots$ with the null moves deleted. In other words, $Y^i_{N^i_n} := X^i_n$, a one-dimensional SRW.

Using the independence of the steps taken and Lemma 1.3, the last formula becomes

$$\begin{aligned}\mathbf{P}_o[\,X_{2n}=o\,] &\asymp \sum_{|\underline{n}|=n}(n_1\cdots n_d)^{-1/2}\mathbf{P}[\,N^i_{2n}=2n_i\ \forall i\,]\\ &\asymp \sum_{\exists n_i\notin[\frac{n}{2d},\frac{3n}{2d}]}\epsilon(\underline{n})\cdot\mathbf{P}[\,N^i_{2n}=2n_i\ \forall i\,]\ +\sum_{\forall n_i\in[\frac{n}{2d},\frac{3n}{2d}]}C_d\cdot\ n^{-d/2}\cdot\mathbf{P}[\,N^i_{2n}=2n_i\ \forall i\,]\,,\end{aligned}\tag{1.5}$$ {e.twosums}

where $\epsilon(\underline{n})\in[0,1]$. Now, by Lemma 1.4,

$$\sum_{\exists n_i\notin[\frac{n}{2d},\frac{3n}{2d}]}\mathbf{P}[\,N^i_{2n}=2n_i\ \forall i\,]\le C\cdot d\cdot\exp(-c_d n)\,,$$

hence the first term in (1.5) is exponentially small, while the second term is polynomially large, so we get

$$\mathbf{P}_o[\,X_{2n}=o\,]\asymp C_d n^{-d/2}\,,\tag{1.6}$$ {e.Zdret}

as claimed. □

We will now look at SRW on a very different graph: the $k$-regular tree $\mathbb{T}_k$ with $k\ge 3$. Also the walk has a very different behaviour, which can be most easily seen on the **rate of escape**.

In $\mathbb{Z}$, or any $\mathbb{Z}^d$, $\mathbf{E}[\,\mathrm{dist}(X_n,X_0)\,]\asymp\sqrt{n}$; see Exercise 1.2 below.

For the $k$-regular tree $\mathbb{T}_k$, $k\ge 3$, let $\mathrm{dist}(X_n,X_0)=D_n$. Then

$$\begin{aligned} &\mathbf{P}\big[\,D_{n+1}=D_n+1\ \big|\ D_n\neq 0\,\big]=\frac{k-1}{k}\\ \text{and}\quad &\mathbf{P}\big[\,D_{n+1}=D_n-1\ \big|\ D_n\neq 0\,\big]=\frac{1}{k}\,,\\ \text{hence}\quad &\mathbf{E}\big[\,D_{n+1}-D_n\ \big|\ D_n\neq 0\,\big]=\frac{k-1}{k}-\frac{1}{k}\,.\end{aligned}$$

On the other hand, $\mathbf{E}\big[\,D_{n+1}-D_n\ \big|\ D_n=0\,\big]=1$. Altogether,

$$\mathbf{E}[\,D_n\,]\ge\frac{k-2}{k}n\,.$$

One can even couple the process $(D_n)_{n\ge 0}$ to an asymmetric random walk $(Y_n)_{n\ge 0}$ on $\mathbb{Z}$, which always goes to the right neighbour with probability $(k-1)/k$ and to the left neighbour with probability $1/k$, in such a way that $D_n\ge Y_n$ holds almost surely for all $n\ge 0$. Since the Strong Law of Large Numbers tells us that $Y_n/n\to(k-2)/k$ almost surely, we get that $Y_n\le 0$ only finitely many times, hence $D_n=0$ only finitely many times. That is, SRW on $\mathbb{T}_k$, with $k\ge 3$, is transient.

{ex.basicspeed

▷ **Exercise 1.2.** *Let $D_n:=\mathrm{dist}(X_n,X_0)$ be the distance of SRW from the starting point.*

**(a)** *On any $\mathbb{Z}^d$, using the Central Limit Theorem, prove that $\mathbf{E}[\,D_n\,]\asymp\sqrt{n}$.*

**(b)** *On $\mathbb{T}_k$, $k\ge 3$, using the transience of the SRW deduced above, show that $D_n/n\to\frac{k-2}{k}$ almost surely, and $\mathbf{E}[\,D_n\,]\sim\frac{k-2}{k}n$, as $n\to\infty$.*

In summary, the random walk escapes much faster on $\mathbb{T}_k$, $k\ge 3$, than on any $\mathbb{Z}^d$. This big difference is also visible in the return probabilities. Using a Large Deviations theorem from Section 1.2 below, one could easily show that $p_{2n}(o,o)$ is exponentially small in $n$ on $\mathbb{T}_k$. We will now actually compute the exact exponential rate of decay.

{d.GU}

**Definition 1.5.** *For simple random walk on a graph $G = (V, E)$, let* **Green's function** *be defined as*

$$G(x, y|z) = \sum_{n=0}^{\infty} p_n(x, y) z^n, \qquad x, y \in V(G) \text{ and } z \in \mathbb{C}\,.$$

*In particular,*

$$G(x, y|1) = \mathbf{E}_x\big[\, \#\text{ of visits to } y\,\big]\,.$$

*Let us also define*

$$U(x, y|z) = \sum_{n=1}^{\infty} \mathbf{P}_x[\tau_y^+ = n] z^n,$$

*where, as before, $\tau_y^+$ is the first positive time when the walk hits $y$.*

*Since any probability is at most 1, the power series $G(x, y|z)$ and $U(x, y|z)$ certainly converge for $|z| < 1$. But one can also think of them as just formal power series, encoding the sequences $p_n(x, y)$ and $\mathbf{P}_x[\tau_y^+ = n]$. Furthermore, for $0 \le z \le 1$ we also have a probabilistic interpretation: $G(x, y|z)$ is the expected number of visits to $y$, while $U(x, y|z)$ is the probability of ever reaching $y$, in the* **random walk with killing***, where we die during each step with probability $1 - z$.*

Note that, for $n \ge 1$, we have $p_n(x, x) = \sum_{k=1}^{n} \mathbf{P}_x[\tau_x^+ = k]\, p_{n-k}(x, x)$, from which we get

$$\begin{aligned} p_n(x, x) z^n &= \sum_{k=1}^{n} \mathbf{P}_x[\tau_x^+ = k]\, z^k\, p_{n-k}(x, x)\, z^{n-k}\,, \\ \sum_{n=1}^{\infty} p_n(x, x) z^n &= U(x, x|z)\, G(x, x|z)\,, \\ G(x, x|z) - 1 &= U(x, x|z)\, G(x, x|z)\,, \\ G(x, x|z) &= \frac{1}{1 - U(x, x|z)}\,, \end{aligned} \tag{1.7}$$ {e.GU}

provided that $G(x, x|z)$ and $U(x, x|z)$ are both finite. For instance, as we proved above, on $\mathbb{Z}^d$ with $d = 1, 2$, we have $G(x, x|1) = \mathbf{E}_x[\,\#\text{ of visits to } x\,] = \infty$, and also $U(x, x|1) = \mathbf{P}_x[\tau_x^+ < \infty] = 1$, so, in fact, (1.7) holds even in this degenerate case.

Let us denote the **radius of convergence** of $G(x, y|z)$ by $\mathsf{rad}(x, y)$. By the Cauchy-Hadamard criterion,

$$\mathsf{rad}(x, y) = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{p_n(x, x)}} \ge 1\,,$$

hence $\mathsf{rad}(x, y)$ determines the exponential rate of decay of the sequence $p_n(x, y)$. In determining the radius of convergence, a useful classical theorem is the following:

{t.Pringsheim}

**Theorem 1.6** (Pringsheim's theorem)**.** *If $f(z) = \sum_n a_n z^n$ with $a_n \ge 0$, then the radius $\mathsf{rad}(f)$ of convergence is the smallest positive singularity of $f(z)$.*

This theorem is not very hard, but we are not going to prove it here. The intuition is that, since $\big|\sum_n a_n z^n\big| \le \sum_n a_n |z|^n$, one can only make the series "more divergent" by replacing $z$ by $|z|$. In particular, for $r = \mathsf{rad}(f)$, if we have $\sum_n a_n r^n < \infty$, then $\sum_n a_n z^n$ is absolutely convergent for any $|z| \le r$. However, this does not exclude the possibility that we have a singularity at some $|z| = r$ with $z \ne$, hence $r$ still could be the radius of convergence without $r$ itself being a singularity, so this argument was far from being a complete proof.

The following is quite similar to Exercise 1.1:

{ex.Gxyvw}

▷ **Exercise 1.3.** *Prove that for simple random walk on a connected graph, for real $z > 0$,*

$$G(x, y|z) < \infty \;\Leftrightarrow\; G(v, w|z) < \infty\,.$$

*Therefore, by Theorem 1.6, we have that* $\mathsf{rad}(x, y)$ *is independent of* $x, y$.

By this exercise, we can define

$$\rho := \frac{1}{\mathsf{rad}(x, y)} = \limsup_{n\to\infty} \sqrt[n]{p_n(x, x)} \le 1\,, \tag{1.8}$$ {e.rho}

which is independent of $x, y$, and is called the **spectral radius** of the simple random walk on the graph. We will see in Section 7.1 the reason for this name.

Now, an obvious fundamental question is when $\rho$ is strictly smaller than 1. I.e., on what graphs are the return probabilities exponentially small? We have seen that they are polynomial on $\mathbb{Z}^d$.

{t.treerho}

**Theorem 1.7.** *The spectral radius of the* $k$*-regular tree* $\mathbb{T}_k$ *is* $\rho(\mathbb{T}_k) = \frac{2\sqrt{k-1}}{k}$.

We first give a proof using the generating functions introduced above, then will sketch a completely probabilistic proof.

*Proof.* With the generating function of Definition 1.5, consider $U(z) = U(x, y|z)$, where $x$ is a neighbour of $y$ (which we will often write as $x \sim y$); this is the probability of ever reaching $y$ when started at $x$. By taking a step on $\mathbb{T}_k$, we see that either we immediately hit $y$ and survive the killing, with probability $\frac{1}{k}z$, or we move to another neighbour of $x$ and survive, with probability $\frac{k-1}{k}z$, in which case, in order to hit $y$, we have to first return to $x$ and then hit $y$. So, by the symmetries of $\mathbb{T}_k$,

$$U(z) = \frac{1}{k}z + \frac{k-1}{k} z U(z)^2\,,$$

which gives

$$U(z) = \frac{k \pm \sqrt{k^2 - 4(k-1)z^2}}{2(k-1)z}\,.$$

From the definition of $U(z)$ as a power series it is clear that $U(0) = 0$. Furthermore, $U(z)$ is easily seen to be continuous for $|z| < 1$, hence we get $U(z) = \frac{k-\sqrt{k^2-4(k-1)z^2}}{2(k-1)z}$. Then,

$$U(x, x|z) = zU(x, y|z) = \frac{k - \sqrt{k^2 - 4(k-1)z^2}}{2(k-1)}\,,$$

and thus, at least for $|z| < 1$,

$$\begin{aligned} G(x, x|z) &= \frac{1}{1 - U(x, x|z)} \\ &= \frac{2(k-1)}{k - 2 + \sqrt{k^2 - 4(k-1)z^2}}\,. \end{aligned} \tag{1.9}$$ {e.GreenTree}

The smallest positive singularity of the right hand side is $z = \frac{k}{2\sqrt{k-1}}$, so Pringheim's Theorem 1.6 gives that this is its radius of convergence. However, we are not quite done yet. We know only that $G(x, x|z)$, whenever exists, is given by the right hand side, but maybe its radius of convergence could be even smaller than that of the right hand side. Nevertheless, this can easily be excluded: the right hand side has a Taylor series expansion for $|z| < \frac{k}{2\sqrt{k-1}}$, while we already know that the Taylor series of $G(x, x|z)$ converges for $|z| < 1$; the values given by these two series coincide on the open set $|z| < 1$, hence the coefficients themselves and the radii of convergence must coincide, too. Therefore, $\rho(\mathbb{T}_k) = \frac{2\sqrt{k-1}}{k}$. □

We have only proved so far that $\limsup_{n\to\infty} \sqrt[n]{p_n(o,o)} = \rho(\mathbb{T}_k)$. Of course, $p_{2n+1}(o,o) = 0$, but is it true that $\lim_{n\to\infty} \sqrt[2n]{p_{2n}(o,o)} = \rho(\mathbb{T}_k)$? The answer is yes. Note that

$$p_{2n+2m}(o,o) \geq p_{2n}(o,o)\, p_{2m}(o,o)\,, \tag{1.10}$$ {e.Fekete}

since a particular way to be back after $2n+2m$ steps is to be back also after $2n$ steps. Hence $a_n := -\log p_{2n}(o,o) \in [0,\infty)$ is a subadditive sequence: $a_{n+m} \leq a_n + a_m$. Fekete's lemma (prove it yourself or look it up in [LyPer16, Exercise 3.9] if you have not seen it) says that $\lim_{n\to\infty} a_n/n$ exists and equals $\inf_n a_n/n$, hence $\lim_{n\to\infty} \sqrt[2n]{p_{2n}(o,o)} = \lim_{n\to\infty} \exp(-a_n/(2n))$ also exists, and equals $\rho(\mathbb{T}_k)$, of course.

▷ **Exercise 1.4.** *Let $k,\ell \geq 2$. Compute $\rho(\mathbb{T}_{k,\ell})$, where $\mathbb{T}_{k,\ell}$ is a tree such that if $v_n \in \mathbb{T}_{k,\ell}$ is a vertex at distance $n$ from the root, then*

$$\deg v_n = \begin{cases} k, & \text{if } n \text{ is even} \\ \ell, & \text{if } n \text{ is odd.} \end{cases}$$

*(Hint: you can either follow the strategy above, or note that two steps on $\mathbb{T}_{k,\ell}$ give a lazy random walk on $\mathbb{T}_{k(\ell-1)}$, where laziness means that in every step there is a certain probability of staying put. This laziness changes, of course, the spectral radius, and the final result will be symmetric in $k$ and $\ell$.)*

The next three exercises provide a probabilistic proof of Theorem 1.7, in a more precise form. (But note that the correction factor $n^{-3/2}$ of Exercise 1.7 can also be obtained by analyzing the singularities of the generating functions.) This probabilistic strategy might be known to a lot of people, but I do not know any reference — I found it based on my own work [GarPS10a, Section 4] and some conversations with Bálint Virág.

{ex.nozeros}

▷ **Exercise 1.5.** *Show that for biased SRW on $\mathbb{Z}$, i.e., $\mathbf{P}[\,X_{n+1} = j+1 \mid X_n = j\,] = p$,*

$$\mathbf{P}_0\big[\,X_i > 0 \text{ for } 0 < i < n \;\big|\; X_n = 0\,\big] \asymp \frac{1}{n}\,,$$

*with constants independent of $p \in (0,1)$. (Hint: first show, using the reflection principle [Dur10, Section 4.3], that for symmetric simple random walk, $\mathbf{P}_0[\,X_i > 0$ for all $0 < i \leq 2n\,] = \frac{1}{2}\mathbf{P}_0[\,X_{2n} = 0\,]$.)*

{ex.zeros}

▷ **Exercise 1.6.** * *For biased SRW on $\mathbb{Z}$, show that there is a subexponentially growing function $g(m) = \exp(o(m))$ such that*

$$\mathbf{P}_0\big[\,\#\{i : X_i = 0 \text{ for } 0 < i < n\} = m \;\big|\; X_n = 0\,\big] \leq g(m)\frac{1}{n}\,.$$

*(Hint: count all possible $m$-element zero sets, together with a good bound on the occurrence of each.)*

{ex.tree3/2}

▷ **Exercise 1.7.** *Note that for SRW on the $k$-regular tree $\mathbb{T}_k$, the distance process $D_n = \mathrm{dist}(X_0, X_n)$ is a biased SRW on $\mathbb{Z}$ reflected at 0.*

**(a)** *Using this and Exercise 1.5, prove that the return probabilities on $\mathbb{T}_k$ satisfy*

$$c_1\, n^{-3/2} \rho^n \leq p_n(x,x) \leq c_2\, n^{-1/2} \rho^n,$$

*for some constants $c_i$ depending only on $k$, where $\rho = \rho(\mathbb{T}_k)$ is given by Theorem 1.7.*

**(b)** *Using Exercise 1.6, improve this to*

$$p_n(x,x) \asymp n^{-3/2} \rho^n\,,$$

*with constants depending only on $k$.*

## 1.2 Large deviations: Azuma-Hoeffding and relative entropy

{ss.LD}

We discuss now a result needed for the random walk estimates in the previous section.

{p.Azuma}

**Proposition 1.8** (Azuma-Hoeffding inequality)**.** *Let $X_1, X_2, \dots$ be random variables satisfying the following criteria:*

- $\mathbf{E}\big[\, X_i \,\big] = 0 \;\forall i$.
- *More generally,* $\mathbf{E}\big[\, X_{i_1} \cdots X_{i_k} \,\big] = 0$ *for any* $k \in \mathbb{Z}_+$ *and* $i_1 < i_2 < \cdots < i_k$.
- $\|X_i\|_\infty < \infty \;\forall i$ .

*Then, for any* $L > 0$,

$$\mathbf{P}\big[\, X_1 + \cdots + X_n > L \,\big] \le e^{-L^2/(2\sum_{i=1}^n \|X_i\|_\infty^2)}.$$

*By symmetry, the same bound holds for* $\mathbf{P}\big[\, X_1 + \cdots + X_n < -L \,\big]$.

Before starting the proof, note that the simplest setup where the conditions of the proposition are satisfied is when the $\{X_i\}$ is an i.i.d. sequence of bounded variables, which was the case also in the previous section. For an i.i.d. sequence, with $\mathbf{E}\big[\, X_i \,\big] = \mu$ and $\|X_i - \mu\|_\infty = \gamma$, for any $\alpha > \mu$ we get

$$\mathbf{P}\big[\, S_n > \alpha n \,\big] \le \exp\left( - \left( \frac{\alpha - \mu}{2\gamma} \right)^2 n \right) ,$$

which immediately implies Lemma 1.4, where we have $\mu = 1/d$ and $\gamma < 1$. In fact, for an i.i.d. Bernoulli sequence, $X_i \sim \mathsf{Ber}(p)$, this exponential bound is among the most basic tools in discrete probability, usually called Chernoff's inequality; see, e.g., [AloS00, Appendix A.1].

However, we would not have stated Proposition 1.8 in such a general form if its only use was for i.i.d. sequences. Indeed, the uncorrelatedness conditions for the $X_i$'s are also fulfilled by martingale differences $X_i = M_i - M_{i-1}$, where $\{M_i\}_{i=0}^\infty$ is a martingale sequence. See Section 6.3 for the definition of a martingale and examples, in case you do not know them already.

{ex.MGdiff}

▷ **Exercise 1.8.** *Show that if* $\{M_i\}_{i=0}^\infty$ *is a martingale, then the differences* $X_i = M_i - M_{i-1}$ *satisfy the uncorrelatedness condition* $\mathbf{E}\big[\, X_{i_1} \cdots X_{i_k} \,\big] = 0$, *for any* $k \in \mathbb{Z}_+$ *and* $i_1 < i_2 < \cdots < i_k$.

*Proof of Proposition 1.8.* Define $S_n := X_1 + \cdots + X_n$. Choose any $t > 0$. We have

$$\begin{aligned} \mathbf{P}\big[\, X_1 + \cdots + X_n > L \,\big] &\le \mathbf{P}\big[\, e^{tS_n} > e^{tL} \,\big] \\ &\le e^{-tL}\mathbf{E}\big[\, e^{tS_n} \,\big] , \quad \text{by Markov's inequality.} \end{aligned}$$

By the convexity of $e^{tx}$, for $|x| \le a$, we have

$$e^{tx} \le e^{at}\frac{a+x}{2a} + e^{-at}\frac{a-x}{2a} = \cosh(at) + \frac{x}{a}\sinh(at) \,.$$

We now apply this for $x := X_i$ and $a = a_i := \|X_i\|_\infty$. The point is that the bound is linear in $X_i$, hence, after taking the product over all $i$ and taking expectations, we can use our condition that $\mathbf{E}\big[\, X_{i_1} \cdots X_{i_k} \,\big] = 0$, and get

$$\mathbf{E}\big[\, e^{tS_n} \,\big] = \mathbf{E}\Big[\, \prod_{i=1}^n e^{tX_i} \,\Big] \le \mathbf{E}\Big[\, \prod_{i=1}^n \Big( \cosh(a_i t) + \frac{X_i}{a_i}\sinh(a_i t) \Big) \,\Big] = \prod_{i=1}^n \cosh(a_i t) \,.$$

Since

$$\cosh(a_i t) = \sum_{k=0}^\infty \frac{(a_i t)^{2k}}{(2k)!} \le \sum_{k=0}^\infty \frac{(a_i t)^{2k}}{2^k k!} = e^{a_i^2 t^2/2} ,$$

we see that

$$\mathbf{P}\big[\, S_n > L \,\big] \le e^{-tL} e^{\sum_{i=1}^n \|X_i\|_\infty^2 t^2/2}.$$

We optimize the bound by setting $t = L/\sum_{i=1}^n \|X_i\|_\infty^2$, proving the proposition. $\square$

Sometimes, the martingale differences are bounded, but not at all symmetric: e.g., in the i.i.d. case, we might have $X_i \sim \mathsf{Ber}(p) - p$, with some $p$ close to 0 or 1. A general way of expressing this is that the conditional variance $\mathrm{Var}[M_{i+1} \,|\, M_i]$ is small. In such cases, the following version is more useful.

{p.AsymmAzuma}

**Proposition 1.9** (Azuma-Hoeffding for small variance martingales)**.** *Let $M_0, M_1, \dots$ be a martingale w.r.t. a filtration $(\mathscr{F}_i)_{i\ge 0}$, satisfying $\|M_i - M_{i-1}\|_\infty \le D < \infty$ and $\sigma_i^2 := \mathrm{Var}[M_{i+1} \,|\, \mathscr{F}_i] < \infty$ for all $i$. Then, for any $L > 0$,*

$$\mathbf{P}\big[\, M_n - M_0 > L \,\big] \le e^{-L^2/\left(2DL/3 + 2\sum_{i=1}^n \sigma_i^2\right)},$$

*and the same for the lower tail.*

*In particular, if $S_n = X_1 + \cdots + X_n$ is a sum of independent indicator variables $X_i \sim \mathsf{Ber}(p_i)$, with total mean $\mu = \sum_{i=1}^n p_i$, then for any $\delta > 0$ there exists $c_\delta > 0$ such that*

$$\mathbf{P}\big[\, |S_n - \mu| > \delta\mu \,\big] \le 2e^{-c_\delta \mu}\,. \qquad (1.11)$$ {e.iidAA}

*Sketch of proof.* The argument is somewhat similar to the previous one. The extra key trick is to use $|M_i - M_{i-1}|^k \le |M_i - M_{i-1}|^2 D^{k-2}$ in the Taylor expansion of $e^{t(M_i - M_{i-1})}$ to obtain that

$$\mathbf{E}\big[\, e^{tM_i} \mid \mathscr{F}_{i-1} \,\big] \le e^{t^2 g(tD)\sigma_i^2} e^{tM_{i-1}}\,,$$

with $g(y) := \sum_{k=2}^\infty \frac{y^{k-2}}{k!}$. Proceed inductively to get $\mathbf{E}\big[\, e^{tM_n} \,\big] \le e^{t^2 g(tD)\sum_{i=1}^n \sigma_i^2} e^{tM_0}$, then choose $t = L/\big(DL/3 + 2\sum_{i=1}^n \sigma_i^2\big)$ in the exponential Markov's inequality, and use that $g(y) < \frac{1/2}{1-y/3}$ for $y < 3$. See [McDi98, Section 3] or [ChuL06, Theorems 18 and 22] for the details. The i.i.d. case is also proved in [AloS00, Corollary A.1.14]. $\square$

In Section 6.3, we will see applications of Proposition 1.8 to martingales. An application of Proposition 1.9 is in Exercise 14.4 part (b), while an application of (1.11) is in the proof of Theorem 12.23. In the rest of the present section, we will discuss some precise general results for i.i.d. sequences.

First of all, for any i.i.d. sequence and any $\alpha \in \mathbb{R}$, the limit

$$\lim_{n\to\infty} \frac{\log \mathbf{P}\big[\, S_n > \alpha n \,\big]}{n} = -\gamma(\alpha) \qquad (1.12)$$ {e.LDiid}

exists. The reason is that

$$\mathbf{P}\big[\, S_n > \alpha n \,\big]\, \mathbf{P}\big[\, S_m > \alpha m \,\big] \le \mathbf{P}\big[\, S_{n+m} > \alpha(n+m) \,\big]\,,$$

hence $\gamma_n := -\log \mathbf{P}\big[\, S_n > \alpha n \,\big]$ is a subadditive sequence, and $\lim_n \gamma_n/n = \inf_n \gamma_n/n$ holds by Fekete's lemma, as for (1.10) in the previous section. The function $\gamma(\alpha)$ is called the **large deviation rate function** (associated with the distribution of $X_i$). In order to ensure that $\gamma(\alpha) > 0$ holds for $\alpha > \mu = \mathbf{E}\big[\, X_i \,\big]$, instead of boundedness, it is enough that the moment generating function $\mathbf{E}\big[\, e^{tX} \,\big]$ is finite for some $t > 0$. The rate function can be computed in terms of the moment generating function, and, e.g., for $X_i \sim \mathsf{Ber}(p)$ it is

$$\gamma_p(\alpha) = \alpha \log \frac{\alpha}{p} + (1-\alpha)\log\frac{1-\alpha}{1-p}\,, \qquad \text{for } \alpha > p\,. \qquad (1.13)$$ {e.LDBinomial}

The proofs of these results involving the moment generating function can be found, e.g., in [Dur10, Section 2.6] and [Bil86, Section 1.9]. We will here sketch a proof of formula (1.13), and give an interpretation of it in terms of relative entropy.

To start with, note that $\gamma_p(\alpha) > 0$ holds for all $\alpha \neq p$. What is its interpretation for $\alpha < p$? It turns out that

$$\lim_{n\to\infty} \frac{\log \mathbf{P}\big[\, \mathsf{Binom}(n,p) = \alpha_n n \,\big]}{n} = -\gamma_p(\alpha) \tag{1.14}$$ {e.LDiideq}

holds for any $\alpha \in \mathbb{R}$ and any sequence $\alpha_n \to \alpha$ such that $\alpha_n n \in \mathbb{N}$ for every $n$. This is an easy consequence of Stirling's formula (see the next exercise). That is, the main claim in the large deviations bounds is that for $\alpha > p$, the exponential rate of $\mathbf{P}\big[\, S_n \geq \alpha_n n \,\big]$ is kind of comparable to being exactly $\alpha_n n$, which is not that surprising (see the next exercise again): it is very hard for $\mathsf{Binom}(n,p)$ to be so large, so it typically fulfils this task by going over $\alpha_n n$ just a little bit, hence it is not that unlikely that it is exactly $\alpha_n n$.

{ex.StirlingLD

▷ **Exercise 1.9.** *Using Stirling's formula, prove first (1.14) then (1.12), with $\gamma_p(\alpha)$ given by (1.13).*

More generally, consider any probability measure $\pi$ on a finite set $S$ such that $\pi(x) > 0$ for all $x \in S$. For i.i.d. samples $\underline{X} = (X_1, \dots, X_n)$ from $\pi$, let $L_n(\underline{X}) := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the associated empirical distribution (where $\delta_x$ is the unit mass at $x$). Furthermore, let $\mu_n \to \mu$ be probability measures on $S$ such that $n\mu_n(x) \in \mathbb{N}$ for all $n$ and $x \in S$. Then, we may ask how hard it is for i.i.d. samples from $\pi$ to behave as if they were samples from $\mu$, and the answer is the following:

$$\lim_{n\to\infty} \frac{1}{n} \log \pi^n \Big( \underline{X} \in S^n : L_n(\underline{X}) = \mu_n \Big) = -D(\mu \,\|\, \pi)\,, \tag{1.15}$$ {e.pimu}

where $D(\mu \,\|\, \pi)$ is called the relative entropy of $\mu$ with respect to $\pi$, which we now define.

{d.entropy}

**Definition 1.10.** *The* **Shannon entropy** *$H(\mu)$ of a probability measure $\mu$ on a countable set $S$ is*

$$H(\mu) := -\sum_{x\in S} \mu(x) \log \mu(x)\,.$$

*Similarly, for a random variable $X$ with values in a countable set, $H(X)$ is the entropy of the measure $\mu(x) = \mathbf{P}[\, X = x \,]$. The* **relative entropy** *between the random variables $X$ and $Y$, also called the* **Kullback-Leibler divergence**, *is*

$$D(X \,||\, Y) = \sum_{x\in S} \mathbf{P}\big[\, X = x \,\big] \, \log \frac{\mathbf{P}[\, X = x \,]}{\mathbf{P}[\, Y = x \,]}\,.$$

*Note that this depends only on the distributions of $X$ and $Y$, not on the joint distribution. Of course, when $S$ is infinite, these quantities might be infinite.*

Note that $H(\mu) \leq \log |\operatorname{supp} \mu|$ by the concavity of $-x \log x$ on $x \in [0,1]$ and Jensen's inequality, with equality for the uniform measure. Here is a certain converse:

▷ **Exercise 1.10.** *If $\mu(x) < \epsilon$ for all $x \in S$ in a finite or countably infinite support $S$ for a probability measure $\mu$, then $H(\mu) \geq \log \frac{1}{\epsilon}$.*

A rough interpretation of $H(\mu)$ is the number of bits needed to encode the amount of randomness in $\mu$, or in other words, the amount of information contained in a random sample from $\mu$. We will see a precise formulation of this in (1.16) below. (The interpretation with bits works the best, of course, if we are using $\log_2$.) A quote from Claude Shannon (from Wikipedia):

"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

Entropy plays an important role in many parts of probability theory and mathematics in general. We will meet it again in Section 5.4, Chapter 9, and Section 13.2. For a nice summary of several probabilistic aspects of entropy see [Geo03], which we partly follow here.

In the special case that $Y = U$ is uniform on $|S| = k$ elements:

$$D(X \,\|\, U) = \sum_{x \in S} \mathbf{P}\big[\, X = x \,\big] \, \log \mathbf{P}[\, X = x \,] + \log k = H(U) - H(X)\,.$$

In this case, using the uniform measure as $\pi$ and noting that the number of all possible sequences $\underline{X} \in S^n$ is $k^n$, the limit (1.15) says that

$$\lim_{n\to\infty} \frac{1}{n} \log \Big| \Big\{ \underline{X} \in S^n : L_n(\underline{X}) = \mu_n \Big\} \Big| = H(\mu)\,. \tag{1.16}$$ {e.piunifmu}

This gives the nice interpretation to $H(\mu)$ that it measures the uncertainty inherent in the measure. (Or in other words, the new information in a sample from $\mu$.) And now it is actually quite easy to prove (1.15):

▷ **Exercise 1.11.**

**(a)** *Using multinomial coefficients and Stirling's formula, prove (1.16).*

**(b)** *Show that, for any probability measure $\pi$ on a finite set $S$, and any $\underline{X} \in S^n$,*

$$-\frac{1}{n} \log \pi^n(\underline{X}) = D\big(L_n(\underline{X}) \,\big\|\, \pi\big) + H\big(L_n(\underline{X})\big).$$

**(c)** *From parts (a) and (b), deduce (1.15).*

Now, given formula (1.15), it is not surprising that the following large deviations result holds. For a proof, see [DemZ98, Theorem 2.1.10]:

**Theorem 1.11** (Sanov's large deviations theorem)**.** *Let $\pi$ be a probability measure and $\mathcal{G}$ a set of probability measures on a finite set $S$, and let $\mathcal{G}^\circ$ and $\overline{\mathcal{G}}$ denote its interior and closure, respectively, as subsets of $\mathbb{R}^{|S|}$. Then,*

$$\begin{aligned} -\inf_{\mu \in \mathcal{G}^\circ} D(\mu \,\|\, \pi) &\le \liminf_{n\to\infty} \frac{1}{n} \log \pi^n \Big( \underline{X} \in S^n : L_n(\underline{X}) \in \mathcal{G} \Big) \\ &\le \limsup_{n\to\infty} \frac{1}{n} \log \pi^n \Big( \underline{X} \in S^n : L_n(\underline{X}) \in \mathcal{G} \Big) \le -\inf_{\mu \in \overline{\mathcal{G}}} D(\mu \,\|\, \pi)\,. \end{aligned}$$

In many natural examples (e.g., the set of $\mathsf{Ber}(q)$ measures with $q > \alpha$), the infima over $\mathcal{G}^\circ$ and $\overline{\mathcal{G}}$ are equal, hence this theorem gives an exact answer. Also, there are several extensions beyond the case of a finite $S$, but then it starts to matter what topology we use on the set of probability measures on $S$, and there are different versions depending on that. See [DemZ98] for more information.

We close this section with the following nice exercise from [GácsL09, Chapter 6]:

{ex.ApproxAlgo

▷ **Exercise 1.12.** *We want to compute a real quantity $a$. Suppose that we have a randomized algorithm that computes an approximation $A$ (which is a random variable) such that the probability that $|A - a| > 1$ is at most $1/20$.*

**(a)** *Show that by calling the algorithm $t$ times, you can compute an approximation $B$ such that the probability that $|B - a| > 1$ is at most $1/2^t$.*

**(b)** *Show by examples that the* $1/2$ *versus* $1/20$ *above is not a typo: if both numbers are* $1/2$ *or both are* $1/20$*, then the claim can fail, e.g., for* $t = 2$.

# 2 Free groups and presentations

{s.groups}

The group structure of $\mathbb{R}^d$ and $\mathbb{Z}^d$ is familiar to everyone from high school, even if it was not defined what a group is. In this chapter, we study free groups, the grandmothers of all groups.

## 2.1 Introduction

{ss.groupintro

**Definition 2.1** ($F_k$, the free group on $k$ generators)**.** *Let* $a_1, a_2, \dots, a_k$*,* $a_1^{-1}, a_2^{-1}, \dots, a_k^{-1}$ *be symbols. The elements of the free group generated by* $\{a_1, \dots, a_k\}$ *are the finite reduced words: remove any* $a_i a_i^{-1}$ *or* $a_i^{-1} a_i$ *repeatedly until there is none. Group multiplication is concatenation of words, followed by reduction if needed. The unit element is the empty word.*

The associativity of this multiplication follows from the next lemma:

**Lemma 2.2.** *Every word has a unique reduced word.*

*Proof.* We use induction on the length of the word $w$ to be reduced. If there is at most one immediate reduction to make in $w$, the induction is obvious. If there are two different reductions, resulting in $w_1$ and $w_2$, note that the reduction leading from $w$ to $w_i$ is still available in $w_{3-i}$, and after making this second reduction we get the same word $w_{12}$ both from $w_1$ and $w_2$. The induction hypothesis for $w_1$, $w_2$, $w_{12}$ now gives that the unique final reduced words from $w_1$ and $w_2$ both must be equal to the final reduced word from $w_{12}$, proving uniqueness for $w$. □

**Proposition 2.3.** *If* $S$ *is a set and* $F_S$ *is the free group generated by* $S$*, and* $\Gamma$ *is any group, then for any map* $f : S \longrightarrow \Gamma$ *there is a group homomorphism* $\hat{f} : F_S \longrightarrow \Gamma$ *extending* $f$*.*

*Proof.* Define $\hat{f}(s_1^{i_1} \cdots s_k^{i_k}) = f(s_1)^{i_1} \cdots f(s_k)^{i_k}$, then check that this is a homomorphism. □

**Corollary 2.4.** *Every group is a quotient of a free group.*

*Proof.* Lazy solution: take $S = \Gamma$, then there is an onto map $F_S \longrightarrow \Gamma$. A less lazy solution is to take a generating set, $\Gamma = \langle S \rangle$. Then, by the proposition, there is a surjective map $\hat{f} : F_S \longrightarrow \Gamma$ with $\ker(\hat{f}) \triangleleft F_S$, hence $F_S / \ker(\hat{f}) \simeq \Gamma$. □

**Definition 2.5.** *Given a generating set* $S$ *and a set of relations* $R$ *of elements of* $S$*, a* **presentation** *of* $\Gamma$ *is given by* $\langle S \rangle$ *mod the relations in* $R$*. This is written* $\Gamma = \langle S | R \rangle$*. More formally,* $\Gamma \simeq \langle S \rangle / \langle\langle R \rangle\rangle$ *where* $R \subset F_S$ *and* $\langle\langle R \rangle\rangle$ *is the smallest normal subgroup containing* $R$*. A group is called* **finitely presented** *if both* $R$ *and* $S$ *can be chosen to be finite sets. The minimal number of generators is called the* **rank** *of the group.*

**Example**: Consider the group $\Gamma = \langle a_1, \dots, a_d \mid [a_i, a_j]\ \forall i, j \rangle$, where $[a_i, a_j] = a_i a_j a_i^{-1} a_j^{-1}$. We wish to show that this is isomorphic to the group $\mathbb{Z}^d$. It is clear that $\Gamma$ is commutive — if we have $a_i a_j$ in a word, we can insert the commutator $[a_i a_j]$ to reverse the order — so every word in $\Gamma$ can be written in the form

$$v = a_1^{n_1} a_2^{n_2} \cdots a_k^{n_k}.$$

Define $\phi : G \longrightarrow \mathbb{Z}^d$ by $\phi(v) = (n_1, n_2, \dots, n_k)$. It is now obvious that $\phi$ is an isomorphism.

{d.Cayley}

**Definition 2.6.** *Let $\Gamma$ be a finitely generated group, $\langle S \rangle = \Gamma$. Then the* **right Cayley graph** $G(\Gamma, S)$ *is the graph with vertex set $V(G) = \Gamma$ and edge set $E(G) = \{(g, gs) : g \in \Gamma, s \in S\}$. The* **left Cayley graph** *is defined using left multiplications by the generators. These graphs are often considered to have directed edges, labeled with the generators, and then they are sometimes called* **Cayley diagrams**. *However, if $S$ is symmetric ($\forall s \in S,\ s^{-1} \in S$), then $G$ is naturally undirected and $|S|$-regular (even if $S$ has order 2 elements).*

**Example**: The $\mathbb{Z}^d$ lattice is the Cayley graph of $\mathbb{Z}^d$ with the $2d$ generators $\{e_i, -e_i\}_{i=1}^d$.
**Example**: The Cayley graph of $F_2$ with generators $\{a, b, a^{-1}, b^{-1}\}$ is a 4-regular tree, see Figure 2.1.

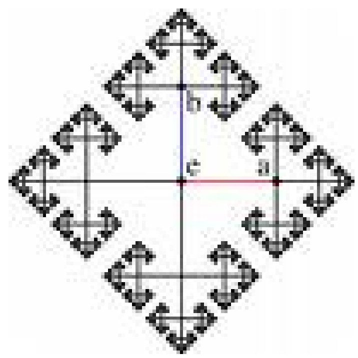



Figure 2.1: The Cayley graph of $F_2$. {f.F2}

Let $\Gamma$ be a group with right Cayley graph $G = G(\Gamma, S)$. Then $\Gamma$ acts on $G$ by multiplication from the left as follows: for $h \in \Gamma$, an element $g \in V(G)$ maps to $hg$, and an element $(g, gs) \in E(G)$ maps to $(hg, hgs)$. This shows that every Cayley graph is transitive.

Finitely presentedness has important consequences for the geometry of the Cayley graphs, see Sections 2.4 and 3.1, and the discussion around Proposition 12.8. Classical examples of finitely generated non-finitely presented groups are the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$, which will be defined in Section 5.1, and studied in Chapter 9 from a random walk point of view. The non-finitely presentedness of $\mathbb{Z}_2 \wr \mathbb{Z}$ is proved in Exercise 12.22.

We have seen two effects of $\mathbb{T}_k$ being much more "spacious" than $\mathbb{Z}^d$ on the behaviour of simple random walk: the escape speed is much larger, and the return probabilities are much smaller. It looks intuitively clear that $\mathbb{Z}$ and $\mathbb{T}_k$ should be the extremes, and that there should be a large variety of possible behaviours in between. It is indeed relatively easy to show that $\mathbb{T}_k$ is one extreme among $2k$-regular Cayley-graphs, but, from the other end, it is only a recent theorem that the expected rate of escape is at least $\mathbf{E}\big[\operatorname{dist}(X_0, X_n)\big] \geq c\sqrt{n}$ on any group: see Section 10.2. One reason for this not being obvious is that not every infinite group contains $\mathbb{Z}$ as a subgroup: there are finitely generated infinite groups with a finite exponent $n$: the $n$th power of any element is the identity. Groups with such strange properties are called Tarksi monsters. But even on groups with a $\mathbb{Z}$ subgroup, there does not seem to be an easy proof.

We will also see that constructing groups with intermediate behaviours is not always easy. One reason for this is that the only general way that we have seen so far to construct groups is via presentations, but there are famous undecidability results here: there is no general algorithm to decide whether a word can be reduced in a given presentation to the empty word, and there is no general algorithm to decide if a group given by a presentation is isomorphic to another group, even to the trivial group. So, we will need other means of constructing groups.

## 2.2 Digression to topology: the fundamental group and covering spaces

{ss.topology}

Several results on free groups and presentations become much simpler in a topological language. The present section discusses the necessary background.

We will need the concept of a CW-complex. The simplest way to define an **$n$-dimensional CW-complex** is to do it recursively:

- A 0-complex is a countable (finite or infinite) union of points, with the discrete topology.
- To get an $n$-complex, we can glue $n$-cells to an $n-1$-complex, i.e., we add homeomorphic images of the $n$-balls such that each boundary is mapped continuously onto a union of $n-1$-cells.

We will always assume that our topological spaces are connected CW-complexes.

Consider a space $X$ and a fixed point $x \in X$. Consider two loops $\alpha : [0,1] \longrightarrow X$ and $\beta : [0,1] \longrightarrow X$ starting at $x$. I.e., $\alpha(0) = \alpha(1) = x = \beta(0) = \beta(1)$. We say that $\alpha$ and $\beta$ are **homotopic**, denoted by $\alpha \sim \beta$, if there is a continuous function $f : [0,1] \times [0,1] \longrightarrow X$ satisfying

$$f(t,0) = \alpha(t), \qquad f(t,1) = \beta(t), \qquad f(0,s) = f(1,s) = x \quad \forall s \in [0,1].$$

We are ready to define the **fundamental group** of $X$. Let $\pi_1(X,x)$ be the set of equivalence classes of paths starting and ending at $x$. The group operation on $\pi_1$ is induced by concatenation of paths:

$$\alpha\beta(t) = \begin{cases} \alpha(2t) & t \in [0, \frac{1}{2}] \\ \beta(2t-1) & t \in [\frac{1}{2}, 1] \end{cases}.$$

While it seems from the definition that the fundamental group would depend on the point $x$, this is not true. To find an isomorphism between $\pi_1(X,x)$ and $\pi_1(X,y)$, map any loop $\gamma$ starting at $x$ to a path from $y$ to $x$ concatenated with $\gamma$ and the same path from $x$ back to $y$.

The spaces $X$ and $Y$ are **homotopy equivalent**, denoted by $X \sim Y$, if there exist continuous functions $f : X \longrightarrow Y$ and $g : Y \longrightarrow X$ such that

$$f \circ g \sim \mathrm{id}_Y, \qquad g \circ f \sim \mathrm{id}_X.$$

A basic result (with a simple proof that we omit) is the following:

{t.fundhomotop

**Theorem 2.7.** *If $X \sim Y$ then $\pi_1(X) \cong \pi_1(Y)$.*

{t.fundRose}

**Theorem 2.8.** *Consider the CW-complex with a single point and $k$ loops from this point. Denote this CW-complex by $\mathrm{Rose}_k$, a rose with $k$ petals. Then $\pi_1(\mathrm{Rose}_k) = F_k$.*

The proof of this theorem uses the Seifert-van Kampen Theorem 3.2. We do not discuss this here, but we believe the statement is intuitively obvious.

{c.fundgraph}

**Corollary 2.9.** *The fundamental group of any (connected) graph is free.*

*Proof.* For any finite connected graph with $n$ vertices and $l$ edges, consider a spanning tree $T$. Then $T$ has $n-1$ edges. Contract $T$ to a point $x$. There are $k = l-n+1$ edges left over, and after the contraction each begins and ends at $x$. Contraction of a spanning tree to a point is a homotopy equivalence, so the graph is homotopy equivalent to $\mathrm{Rose}_k$. Hence, the fundamental group is free by Theorems 2.7 and 2.8. □

▷ **Exercise 2.1.** *If $\Gamma$ is a topological group then $\pi_1(\Gamma)$ is commutative. (Recall that a group $\Gamma$ is a topological group if it is also a topological space such that the functions $\Gamma\times\Gamma \longrightarrow \Gamma : (x,y) \mapsto xy$ and $\Gamma \longrightarrow \Gamma : x \mapsto x^{-1}$ are continuous.)*

We now introduce another basic notion of geometry:

{d.covering}

**Definition 2.10.** *We say that $X' \overset{p}{\twoheadrightarrow} X$ is a* **covering map** *if for every $x \in X$, there is an open neighbourhood $U \ni x$ such that each connected component of $p^{-1}(U)$ is homeomorphic to $U$ by $p$.*

{p.uniquelift}

**Proposition 2.11.** *Let $\gamma \subset X$ be any path starting at $x$. Then for every $x^* \in p^{-1}(x)$ there is a unique $\gamma^*$ starting at $x^*$ with $p(\gamma^*) = \gamma$.*

*Sketch of proof.* Because of the local homeomorphisms, there is always a unique way to continue the lifted path. □

{t.monodromy}

**Theorem 2.12** (Monodromy Theorem)**.** *If $\gamma$ and $\delta$ are homotopic with fixed endpoints $x$ and $y$, then the lifts $\gamma^*$ and $\delta^*$ starting from the same $x^*$ have the same endpoints and $\gamma^* \sim \delta^*$.*

*Sketch of proof.* The homotopies can be lifted through the local homeomorphisms. □

The following results can be easily proved using Proposition 2.11 and Theorem 2.12:

{l.graphcover}

**Lemma 2.13.** *Any covering space of a graph is a graph.*

{l.k-fold}

**Lemma 2.14.** *Let $x \in X$ and $U \ni x$ a neighbourhood as in Definition 2.10. Then the number $k$ of connected components of $p^{-1}(U)$ is independent of $x$ and $U$, and the covering is called $k$-fold. (In fact, the lifts of any path $\gamma$ between $x, y \in X$ give a pairing between the preimages of $x$ and $y$.)*

Let $X$ be a topological space. We say that $\widehat{X}$ is a **universal cover** of $X$ if:

- $\widehat{X}$ is a cover.
- $\widehat{X}$ is connected.
- $\widehat{X}$ is simply connected, i.e., $\pi_1(\widehat{X}) = 1$.

The existence of a universal cover is guaranteed by the following theorem:

{t.univcover}

**Theorem 2.15.** *Every connected CW complex $X$ has a universal cover $\widehat{X}$.*

*Sketch of proof.* Let the set of points of $\widehat{X}$ be the fixed endpoint homotopy classes of paths starting from a fixed $x \in X$. The topology on $\widehat{X}$ is defined by thinking of a class of paths $[\gamma]$ to be close to $[\delta]$ if there are representatives $\gamma, \delta$ such that $\delta$ is just $\gamma$ concatenated with a short piece of path. □

▷ **Exercise 2.2.** *Write down the above definition of the topology on $\widehat{X}$ properly and the proof that, with this topology, $\pi_1(\widehat{X}) = 1$.*

The fundamental group $\pi_1(X)$ acts on $\widehat{X}$ with continuous maps, as follows.

Let $\gamma \in \pi_1(X,x)$ be a loop, $p$ the surjective map defined by the covering $\widehat{X}$, and $x^* \in p^{-1}(x)$ a point above $x$. Using Proposition 2.11 and Theorem 2.12, there exists a unique homeomorphic curve $\gamma^*$ (not depending on the representative for $\gamma$) with starting point $x^*$ and some ending point $\overline{x^*}$ , which also belongs to $p^{-1}(x)$. The action $f_\gamma$ on $x^*$ is now defined by $f_\gamma(x^*) = \overline{x^*}$. As we mentioned above, this action does not depend on the representative for $\gamma$, and it is clear that $f_\delta \circ f_\gamma = f_{\gamma\delta}$ for $\gamma, \delta \in \pi_1(X,x)$, so we indeed have an action of $\pi_1(X,x)$ on the fibre $p^{-1}(x)$.

We need to define the action also on any $y^* \in p^{-1}(y)$, $y \in X$. We take a path $\delta^*$ in $\widehat{X}$ from $y^*$ to $x^*$, then $\delta = p(\delta^*)$ is a path from $y$ to $x$, and $\delta\gamma\delta^{-1}$ is a path from $y$ to itself. Since $\pi_1(\widehat{X}) = 1$, all possible choices of $\delta^*$ are homotopic to each other, hence all resulting $\delta$ and all $\delta\gamma\delta^{-1}$ curves are homotopic. By Proposition 2.11 and Theorem 2.12, there is a unique lift of $\delta\gamma\delta^{-1}$ starting from $y^*$, and its endpoint $\overline{y^*} \in p^{-1}(y)$ does not depend on the choice of $\delta^*$. Hence we indeed get an action of $\pi_1(X,x)$ on the entire $\widehat{X}$.

This action permutes points inside each fibre, and it is easy to see that the action is free (i.e., only the identity has fixed points). If we make the quotient space of $\widehat{X}$ by this group action (i.e., each point of $\widehat{X}$ is mapped to its orbit under the group action), we will obtain $X$, and the quotient map is exactly the covering map $p$:

$$\widehat{X}/\pi_1(X,x) = X\,. \tag{2.1}$$ {e.pi1factor}

**Example:** The fundamental group of the torus $\mathbb{T}^2$ is $\mathbb{Z}^2$. We have $\mathbb{R}^2/\mathbb{Z}^2 = \mathbb{T}^2$, where $\mathbb{R}^2$ is the usual covering made out of copies of the square $[0,1) \times [0,1)$.

{ex.qtransTree

▷ **Exercise 2.3.**

**(a)** *Show that the universal covering tree of any finite graph is quasi-transitive.*

**(b)** *Give an example of a quasi-transitive infinite tree that is not the universal covering tree of any finite graph. (Hint: specify that there are $k$ orbits in the tree, and for $1 \le i,j \le k$, each vertex of type $i$ has $n(i,j)$ neighbours of type $j$.)*

For each subgroup $H \le \pi_1(X)$, we can consider $X_H = \widehat{X}/H$, which is still a covering space of $X$: there is a natural surjective map from $X_H$ to $X$, by taking the $\pi_1(X)$-orbit containing any given $H$-orbit. On the other hand, $\widehat{X}$ is the universal cover of $X_H$, too, and we have $\pi_1(X_H) \cong H$.

## 2.3 The main results on free groups

{ss.freemain}

A probably unsurprising result is the following:

{t.freeisom}

**Theorem 2.16.** *$F_k \cong F_l$ only if $k = l$ .*

▷ **Exercise 2.4.** *Prove Theorem 2.16. (Hint 1: how many index 2 subgroups are there? Or, hint 2: What is $F_k/[F_k,F_k]$?)*

The next two results can be proved also with combinatorial arguments, but the topological language of the previous section makes things much more transparent.

{t.NieSch}

**Theorem 2.17** (Nielsen-Schreier)**.** *Every subgroup of a free group is free.*

{t.Schind}

**Theorem 2.18** (Schreier's index formula)**.** *If $F_k$ is free and $F_l \le F_k$ such that $[F_k : F_l] = r < \infty$, then*

$$l - 1 = (k-1)r.$$

*Proof of Theorem 2.17.* Let the free group be $F_k$. By Theorem 2.8, we have $F_k \cong \pi_1(\mathrm{Rose}_k)$. Let $G$ be the universal cover of $\mathrm{Rose}_k$. By Lemma 2.13, it is a graph; in fact, it is the $2k$-regular tree $T_{2k}$.

Now take $H \leq F_k$. As discussed at the end of the previous section, $G_H = G/H$ is a covering of $\mathrm{Rose}_k$ and is covered by $T_{2k}$. Again by Lemma 2.13, it is a graph, and $\pi_1(G_H) = H$. By Corollary 2.9, the fundamental group of a graph is free, hence $H$ is free, proving Theorem 2.17. □

*Proof of Theorem 2.18.* The Euler characteristic of a graph is the difference between the number of vertices and the number of edges:

$$\chi(G) = |V(G)| - |E(G)| .$$

Homotopies of graphs increase or decrease the vertex and the edge sets by the same amount, hence $\chi$ is a homotopy invariant of graphs. Furthermore, if $G'$ is an $r$-fold covering of $G$, then $r\chi(G) = \chi(G')$.

Since the graph $\mathrm{Rose}_k$ has one vertex and $k$ edges, $\chi(\mathrm{Rose}_k) = 1 - k$. As the index of $H = F_l$ in $F_k$ is $r$, we see that $T_{2k}/H$ is an $r$-fold covering of $\mathrm{Rose}_k$. Thus, $\chi(T_{2k}/H) = r\chi(\mathrm{Rose}_k) = r(1-k)$.

On the other hand, $T_{2k}/H$ is a graph with $\pi_1(T_{2k}/H) = H = F_l$. Since any graph is homotopic to a rose, and, by Theorems 2.8 and 2.16, different roses have non-isomorphic fundamental groups, $\pi_1(\mathrm{Rose}_l) = F_l$, we get that $T_{2k}/H$ must be homotopic to $\mathrm{Rose}_l$, and thus $\chi(T_{2k}/H) = 1 - l$. Therefore, we obtain $r(1-k) = 1-l$. □

Notice that, if $2 \leq r < \infty$ then $r(1-k) \neq 1-k$.

▷ **Exercise 2.5.** *Prove that $F_k$ has itself as a subgroup of infinite index.*

▷ **Exercise 2.6.** * *A finitely generated group acts on a tree freely if and only if the group is free. (The action is by graph automorphisms of the tree $T$, and a free action means that $\mathrm{Stab}_G(x) = \{1\}$ for any $x \in V(T) \cup E(T)$.)*

*Hint: separate the cases where there is an element of order 2 in $\Gamma$, and where there are no such elements (in which case there is a fixed vertex, as it turns out).*

We will now show that a finitely generated free group is linear. For instance, $F_2 \leq \mathsf{SL}_2(\mathbb{Z})$ (integer $2 \times 2$ matrices with determinant 1). Indeed:

{l.pingpong}

**Lemma 2.19** (Ping-Pong Lemma)**.** *Let $\Gamma$ be a group acting on some set $X$. Let $\Gamma_1$, $\Gamma_2$ be subgroups of $\Gamma$, with $|\Gamma_1| \geq 3$, and let $\Gamma^* = \langle \Gamma_1, \Gamma_2 \rangle$. Assume that there exist non-empty sets $X_1, X_2 \subseteq X$ with $X_2 \nsubseteq X_1$ and*

$$\gamma(X_2) \subseteq X_1 \ \forall \gamma \in \Gamma_1 \setminus \{1\} ,$$
$$\gamma(X_1) \subseteq X_2 \ \forall \gamma \in \Gamma_2 \setminus \{1\} .$$

*Then $\Gamma_1 * \Gamma_2 = \Gamma^*$.*

The product $*$ denotes the free product. If $\Gamma_1 = \langle S_1 | R_1 \rangle$ and $\Gamma_2 = \langle S_2 | R_2 \rangle$, then $\Gamma_1 * \Gamma_2 = \langle S_1, S_2 | R_1, R_2 \rangle$. In particular, $F_k = \mathbb{Z} * \mathbb{Z} * \cdots * \mathbb{Z}$, $k$ times.

▷ **Exercise 2.7.** *Prove the Ping-Pong Lemma.*

Now let

$$\Gamma_1 = \left\{ \begin{pmatrix} 1 & 2n \\ 0 & 1 \end{pmatrix}, \ n \in \mathbb{Z} \right\}, \qquad \Gamma_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 2n & 1 \end{pmatrix}, \ n \in \mathbb{Z} \right\},$$

with

$$X_1 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \ |x| > |y| \right\}, \qquad X_2 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \ |x| < |y| \right\},$$

with the usual (linear) action of matrices in $\mathsf{SL}_2(\mathbb{Z})$ on $\mathbb{R}^2$. Observe that the hypotheses of the Ping-Pong Lemma are fulfilled. Therefore, the matrices $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ generate a free group.

{ex.resfin}

▷ **Exercise 2.8.** *Show that the following properties of a group $\Gamma$ are equivalent.*

**(i)** *For every $\gamma \neq 1 \in \Gamma$ there is a homomorphism $\pi$ onto a finite group $F$ such that $\pi(\gamma) \neq 1$.*

**(ii)** *The intersection of all its subgroups of finite index is trivial.*

**(iii)** *The intersection of all its normal subgroups of finite index is trivial.*

*Such groups are called* **residually finite***.*

For instance, $\mathbb{Z}^n$, $\mathsf{SL}_n(\mathbb{Z})$ and $\mathsf{GL}_n(\mathbb{Z})$ are residually finite, because of the natural (mod $p$) homomorphisms onto $(\mathbb{Z}/p\mathbb{Z})^n$, $\mathsf{SL}_n(\mathbb{Z}/p\mathbb{Z})$ and $\mathsf{GL}_n(\mathbb{Z}/p\mathbb{Z})$.

{ex.Fresfin}

▷ **Exercise 2.9.** *Show that any subgroup of a residually finite group is also residually finite. Conclude that the free groups $F_k$ are residually finite.*

▷ **Exercise 2.10.**

**(i)** *Show that if $\Gamma$ is residually finite and $\phi : \Gamma \longrightarrow \Gamma$ is a surjective homomorphism, then it is a bijection. (Hint: assume that* $\ker\phi$ *is non-trivial, show that the number of index $k$ subgroups of $\Gamma \simeq \Gamma/\ker\phi$ is finite, then arrive at a contradiction. Alternatively, show that any homomorphism $\pi$ onto a finite group $F$ must annihilate* $\ker\phi$*.)*

**(ii)** *Conclude that if $k$ distinct elements generate the free group $F_k$, then they generate it freely.*

## 2.4 Presentations and Cayley graphs

{ss.presentatio

Fix some subset of (oriented) loops in a directed Cayley graph $G(\Gamma, S)$ of a group $\Gamma = \langle S|R\rangle$, which will be called the *basic loops*. A *geometric conjugate* of a loop can be:

- "rotate" the loop, i.e., choose another starting point in the same loop;
- translate the loop in $G$ by a group element.

A *combination* of basic loops is a loop made of a sequence of loops $(\ell_i)$, where the $i$-th is a geometric conjugate of a basic loop, and starts in a vertex contained in some $\ell_j, j \in [1, i-1]$.

To any loop $\ell = (g, gs_1, gs_1s_2, \ldots, gs_1s_2\cdots s_k = g)$, with each $s_i \in S$, we can associate the word $w(\ell) = s_1s_2\cdots s_k$. So, translated copies of the same loop have the same word associated to them.

{p.loops}

**Proposition 2.20.** *Any loop in $G(\Gamma, S)$ is homotopy equivalent to a combination of basic loops if and only if the words given by the basic loops generate $\langle\langle R\rangle\rangle$ as a normal subgroup.*

Recall that $\langle\langle R\rangle\rangle$ denotes the normal subgroup of the relations, and $\Gamma = F_S/\langle\langle R\rangle\rangle$.

*Proof.* Let the set of all loops in $G$ be $\mathcal{L}$, the set of basic loops $\mathcal{B}$, and the set of loops produced from $\mathcal{B}$ by combining them and applying homotopies $\mathcal{C}$. The statement of the proposition is that $\mathcal{C} = \mathcal{L}$ if and only if $\langle\langle w(\mathcal{B})\rangle\rangle = \langle\langle R\rangle\rangle$.

We first show that $\langle\langle w(\mathcal{B})\rangle\rangle = w(\mathcal{C})$. Let us start with the direction $\supseteq$.

What happens to the associated word when we rotate a loop? From the word $s_1s_2\cdots s_k \in w(\mathcal{B})$, rotating by one edge, say, we get $s_2\cdots s_ks_1$. This new word is in fact a conjugate of the old word: $s_1^{-1}(s_1s_2\cdots s_k)s_1$, hence it is an element of $\langle\langle w(\mathcal{B})\rangle\rangle$.

Next, when we combine two loops, $(g, gs_1, \ldots, gs_1 \cdots s_n = g)$ and $(h, ht_1, \ldots, ht_1 \cdots t_m = h)$ into

$$(g, \ldots, gS_1, gS_1t_1, \ldots, gS_1T, gS_1Ts_{k+1}, gS_1TS_2 = g),$$

where $S_1 = s_1 \cdots s_k$, $S_2 = s_{k+1} \cdots s_n$ and $T = t_1 \cdots t_m$, then the new associated word $S_1TS_2$ equals $SS_2^{-1}TS_2$. Thus, combining loops does not take us out from $\langle\langle w(\mathcal{B})\rangle\rangle$. See Figure 2.2 (a).

Finally, if two loops are homotopic to each other in $G$, then we can get one from the other by combining in a sequence of contractible loops, which are just "contour paths" of subtrees in $G$, i.e., themselves are combinations of trivial $s_i s_i^{-1}$ loops for some generators $s_i \in S$. See Figure 2.2 (b). The effect of these geometric combinations on the corresponding word is just plugging in trivially reducible words, wich again does not take us out from $\langle\langle w(\mathcal{B})\rangle\rangle$.

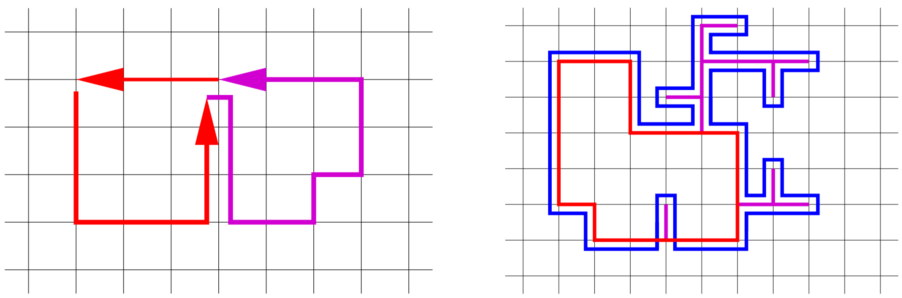


Figure 2.2: (a) Combining two loops. (b) Homotopic loops inside $\mathbb{Z}^2$. {f.loops}

So, we have proved $\langle\langle w(\mathcal{B})\rangle\rangle \supseteq w(\mathcal{C})$. But $\subseteq$ is now also clear: we have encountered the geometric counterpart of all the operations on the words in $w(\mathcal{B})$ that together produce $\langle\langle w(\mathcal{B})\rangle\rangle$.

On the other hand, we obviously have $w(\mathcal{L}) = \langle\langle R\rangle\rangle$, since a word on $S$ is a loop in $G$ iff the word represents 1 in $\Gamma$ iff the word is in the kernel $\langle\langle R\rangle\rangle$ of the factorization from the free group $F_S$.

So, the last step we need is that $\mathcal{L} = \mathcal{C}$ if and only if $w(\mathcal{L}) = w(\mathcal{C})$. Since both $\mathcal{L}$ and $\mathcal{C}$ are closed under translations, this is clear. $\square$

We now state an elegant topological definition of the Cayley graph $G(\Gamma, S)$: {d.CayleyCW}

**Definition 2.21.** *Take the 1-dimensional CW-complex $Rose_{|S|}$, and for any word in $R$, glue a 2-cell on it with boundary given by the word. Let $X(S,R)$ denote the resulting 2-dimensional CW complex. Now take the universal cover $\widehat{X}(S,R)$. This is called the* **Cayley complex** *corresponding to $S$ and $R$. The 1-skeleton of $\widehat{X}(S,R)$ is the $2|S|$-regular Cayley graph $G(\Gamma, S)$ of $\Gamma = \langle S|R\rangle$.*

For instance, when $S = \{a, b\}$ and $R = \{aba^{-1}b^{-1}\}$, we glue a single 2-cell, resulting in a torus, whose universal cover will be homeomorphic to $\mathbb{R}^2$, with a $\mathbb{Z}^2$ lattice as its 1-skeleton.

The proof that this definition of the Cayley graph coincides with the usual one has roughly the same math content as our previous proposition. But here is a sketch of the story told in the language of covering surfaces: It is intuitively clear, and can be proved using the Seifert – van Kampen theorem, that $\pi_1(X(S,R)) = \langle S|R\rangle$. Then, as we observed in (2.1), $\widehat{X}(S,R)/\Gamma = X(S,R)$. We can now consider the **right Schreier graph** $G(\Gamma, \widehat{X}, S)$ of the action of $\Gamma$ on $\widehat{X} = \widehat{X}(S,R)$ with generating set $S$: the vertices are the vertices of the CW complex $\widehat{X}$ (this is just one $\Gamma$-orbit), and the edges are $(x^*, x^*s)$, where $x^*$ runs through the vertices and $s$ runs through $S$. Clearly, this is exactly the 1-skeleton of $\widehat{X}$. On the other hand, since the action is free, this Schreier graph is just the Cayley graph $G(\Gamma, S)$, and we are done.

So, if we factorize $F_k$ by normal subgroups, we get all possible groups $\Gamma$, and the Schreier graph of the action by $\Gamma$ on the Cayley complex will be a $2k$-regular Cayley graph. Besides the Schreier graph of a group action, there is another usual meaning of a Schreier graph, which is just a special case: if $H \leq \Gamma$, and $S$ is a generating set of $\Gamma$, then the set $H\backslash\Gamma$ of right cosets $\{Hg\}$ supports a natural graph structure, $Hg \sim Hgs$ for $s \in S$. This graph is denoted by $G(\Gamma, H, S)$.

{ex.Schreier}

▷ **Exercise 2.11.*** *Show that any $2k$-regular graph is the Schreier graph of $F_k$ with respect to some subgroup $H$. On the other hand, the 3-regular Petersen graph is not a Schreier graph.*

The above results show that being finitely presented is a property with a strong topological flavour. Indeed, it is clearly equivalent to the existence of some $r < \infty$ such that the $\mathrm{Rips}_r(G(\Gamma, S))$ complex is simply connected, with the following definition:

{d.Rips}

**Definition 2.22.** *If $X$ is a CW-complex and $r > 0$, then the* **Rips complex** $\mathrm{Rips}_r(X)$ *is given by adding all simplices (of arbitrary dimension) of diameter at most $r$.*

# 3 The asymptotic geometry of groups

{s.asymptotic}

## 3.1 Quasi-isometries. Ends. The fundamental observation of geometric group theory

{ss.qisom}

We now define what we mean by two metric spaces having the same geometry on the large scale. It is a weakening of bi-Lipschitz equivalence:

{d.Quasi-isome

**Definition 3.1.** *Suppose $(X_1, d_1)$ and $(X_2, d_2)$ are metric spaces. A map $f : X_1 \longrightarrow X_2$ is called a* **quasi-isometry** *if $\exists\, C > 0$ such that the following two conditions are met:*

1. *For all $p, q \in X_1$, we have $\frac{d_1(p,q)}{C} - C \leq d_2(f(p), f(q)) \leq C\, d_1(p,q) + C$.*
2. *For each $x_2 \in X_2$, there is some $x_1 \in X_1$ with $d_2(x_2, f(x_1)) < C$.*

*Informally speaking, the first condition means that $f$ does not distort the metric too much (it is coarsely bi-Lipschitz), while the second states that $f(X_1)$ is relatively dense in the target space $X_2$.*

*We also say that $X_1$ is* **quasi-isometric** *to $X_2$, denoted by $X_1 \simeq_q X_2$, if there exists a quasi-isometry $f : X_1 \longrightarrow X_2$. The notation suggests that this is an equivalence relation — this is shown in the next exercise.*

{ex.Qisom-equi

▷ **Exercise 3.1.** *Verify that being quasi-isometric is an equivalence relation. (Hint: given $f : X_1 \longrightarrow X_2$, construct a quasi-inverse $g : X_2 \longrightarrow X_1$, a quasi-isometry with the property that both $g \circ f$ and $f \circ g$ are at a bounded distance from the identity, i.e., $\sup_{x \in X_1} d_1(x, g(f(x)))$ and $\sup_{y \in X_2} d_2(y, f(g(y)))$ are finite.)*

For example, $\mathbb{Z}^2$ with the graph metric (the $\ell^1$ or taxi-cab metric) is quasi-isometric to $\mathbb{R}^2$ with the Euclidean metric. Also, $\mathbb{Z} \times \mathbb{Z}_2$ with the graph metric is quasi-isometric to $\mathbb{Z}$: $(n, i) \mapsto n$ for $n \in \mathbb{Z}$, $i \in \{0, 1\}$ is a non-injective quasi-isometry that is "almost an isometry on the large scale", while $(n, i) \mapsto 2n + i$ is an injective quasi-isometry with Lipschitz constant 2. Finally, and maybe most importantly to us, if $\Gamma$ is a finitely-generated group, its Cayley graph depends on choosing the symmetric finite generating set, but the good thing is that any two such graphs are quasi-isometric: each generator $s_i$ in the first generating set is a product of some finite number $k_i$ of generators from the second generating set, hence any path of length $\ell$ in the first Cayley graph can be replaced by a path of length at most $\ell \max\{k_i : s_i \in S\}$ in the second Cayley graph.

▷ **Exercise 3.2.** *Show that the regular trees $\mathbb{T}_k$ and $\mathbb{T}_\ell$ for $k, \ell \geq 3$ are quasi-isometric to each other, by giving explicit quasi-isometries.*

For a transitive connected graph $G$ with finite degrees of vertices, we can define the **volume growth function** $v_G(n) = |B_n(o)|$, where $o$ is some vertex of $G$ and $B_n(o)$ is the closed ball of radius $n$ (in the graph metric on $G$) with center $o$. We will sometimes call two functions $v_1, v_2$ from $\mathbb{N}$ to $\mathbb{N}$ **equivalent** if $\exists C > 0$ such that

$$v_1\left(\frac{r}{C}\right)/C < v_2(r) < Cv_1(Cr) \qquad \text{for all } r > 0. \tag{3.1}$$ {e.growthequiv

It is almost obvious that quasi-isometric transitive graphs have equivalent volume growth functions. (For this, note that for any quasi-isometry of locally finite graphs, the number of preimages of any point in the target space is bounded from above by some constant.)

Another quasi-isometry invariant of groups is the property of being finitely presented, as it follows easily from our earlier remark on Rips complexes just before Definition 2.22.

Our next quasi-isometry invariant will be the number and topological structure of the different ways one can disappear into infinity. To start with, call a sequence $\{x_n\}_{n\geq 1} \subseteq V(G)$ of vertices **end-convergent** if, for every finite $K \subset V$, all but finitely many $x_n$'s are in the same connected component of $G \setminus K$.

{ex.Mysterious

▷ **Exercise 3.3.** *Define in a reasonable way the* **space of ends** *of a graph as a topological space, knowing that $\mathbb{Z}$ has two ends, $\mathbb{Z}^d$ has one end for $d \geq 2$, while the $k$-regular tree $\mathbb{T}_k$ has a continuum of ends. (In particular, for every end-convergent sequence of vertices defined above, there should exist exactly one end where the sequence converges to.) Prove that any quasi-isometry of graphs induces naturally a homeomorphism of their spaces of ends. Thus, the number of ends is a quasi-isometry invariant of the graph.*

By invariance under quasi-isometries, we can define the space of ends of a finitely generated group to be the space of ends of any of its Cayley graphs.

▷ **Exercise 3.4** (Hopf 1944)**.** * {ex.2and3ends}

**(a)** *Show that a group has two ends iff it has $\mathbb{Z}$ as a finite index subgroup.*

**(b)** *Show that if a f.g. group has at least 3 ends, then it has continuum many.*

{ex.EndsExampl

▷ **Exercise 3.5.**

**(a)** *Show that if $G_1, G_2$ are two infinite graphs, then the direct product graph $G_1 \times G_2$ has one end.*

**(b)** *Show that if $|\Gamma_1| \geq 2$ and $|\Gamma_2| \geq 3$ are two finitely generated groups, then the free product $\Gamma_1 * \Gamma_2$ has a continuum number of ends.*

An extension of the notion of free product is the **amalgamated free product**: if $\Gamma_i = \langle S_i \mid R_i \rangle$, $i = 1, 2$, are finitely generated groups, each with a subgroup isomorphic to some $H$, with an embedding $\varphi_i : H \longrightarrow \Gamma_i$, then

$$\Gamma_1 *_H \Gamma_2 = \Gamma_1 *_{H,\varphi_1,\varphi_2} \Gamma_2 := \left\langle S_1 \cup S_2 \;\middle|\; R_1 \cup R_2 \cup \{\varphi_1(h)\varphi_2(h)^{-1} : h \in H\} \right\rangle.$$

If both $\Gamma_i$ are finitely presented, then $\Gamma_1 *_H \Gamma_2$ is of course also finitely presentable. There is an important topological way of how such products arise:

{t.SvK}

**Theorem 3.2** (Seifert-van Kampen)**.** *If $X = X_1 \cup X_2$ is a decomposition of a CW-complex into connected subcomplexes with $Y = X_1 \cap X_2$ connected, and $y \in Y$, then*

$$\pi_1(X, y) = \pi_1(X_1, y) *_{\pi_1(Y,y)} \pi_1(X_2, y),$$

*with the natural embeddings of $\pi_1(Y, y)$ into $\pi_1(X_i, x)$.*

{ex.amalgamsem

▷ **Exercise 3.6.**

**(a)** *What is the Cayley graph of $\mathbb{Z} *_{2\mathbb{Z}} \mathbb{Z}$ (with the obvious embedding $2\mathbb{Z} < \mathbb{Z}$), with one generator for each $\mathbb{Z}$ factor? Identify the group as a semidirect product $\mathbb{Z}^2 \rtimes \mathbb{Z}_4$, see Section 4.2.*

**(b)** *Do there exist CW-complexes realizing this amalgamated free product as an application of Seifert-van Kampen?*

The reason for talking about amalgamated free products here is the following theorem. See the references in [DrKa09, Section 2.2] for proofs. (There is one using harmonic functions [Kap14]; I might say something about that in a later version of these notes.)

{t.Stallings}

**Theorem 3.3** (Stallings [Sta68], [Bergm68])**.** *Recall from Exercise 3.4 that a group has 0, 1, 2 or a continuum of ends. The last case occurs iff the group is a free product amalgamated over a finite subgroup.*

The following result is sometimes called "the fundamental observation of geometric group theory". For instance, it connects two usual "definitions" of geometric group theory: 1. it is the study of group theoretical properties that are invariant under quasi-isometries; 2. it is the study of groups using their actions on geometric spaces. We start with some definitions:

{d.spaces}

**Definition 3.4.** *A metric space $X$ is called* **geodesic** *if for all $p, q \in X$ there exists $a, b \in \mathbb{R}$ and an isometry $g : [a, b] \longrightarrow X$ with $g(a) = p$, $g(b) = q$.*

*A metric space is called* **proper** *if all closed balls of finite radius are compact.*

{d.actions}

**Definition 3.5.** *An action of a group $\Gamma$ on a metric space $X$ by isometries is called* **properly discontinuous** *if for every compact $K \subset X$, $|\{g \in \Gamma : g(K) \cap K \neq \emptyset\}|$ is finite.*

*Any group action defines an equivalence relation on $X$: the decomposition into orbits. The set of equivalence classes, equipped with the factor topology coming from $X$, is denoted by $X/\Gamma$. The action is called* **co-compact** *if $X/\Gamma$ is compact.*

{l.Milnor-Schw

**Lemma 3.6** (Milnor-Schwarz)**.** *Let $X$ be a proper geodesic metric space, and suppose that a group $\Gamma$ acts on it by isometries properly discontinuously and co-compactly. Then $\Gamma$ is finitely generated, and for any fixed $x \in X$, the map $J_x : \Gamma \longrightarrow X$ defined by $J_x(g) = g(x)$ is a quasi-isometry (on each Cayley graph).*

*Proof.* Pick an arbitrary point $x \in X$, and consider the projection $\pi : X \longrightarrow X/\Gamma$. By compactness of $X/\Gamma$, there is an $R < \infty$ such that the $\Gamma$-translates of the closed ball $B = B_R(x)$ cover $X$, or in other words, $\pi(B) = X/\Gamma$. (Why exactly? Since each element of $\Gamma$ is invertible, its action on $X$ is a bijective isometry, hence a homeomorphism. So, for any open $U \subset X$ and $g \in \Gamma$, we have that $g(U)$ and $\pi^{-1}(\pi(U)) = \bigcup_{g\in\Gamma} g(U)$ are open in $X$, and therefore $\pi(U)$ is open in $X/\Gamma$. In particular, the images $\pi(B_r^\circ(x))$ of open balls are open. Since $B_r^\circ \nearrow X$ as $r \to \infty$, we also have $\pi(B_r^\circ(x)) \nearrow X/\Gamma$, and, by compactness, there exists an $R < \infty$ such that $X/\Gamma = \pi(B_R^\circ(x))$.)

Since the action of $\Gamma$ is properly discontinuous, there are only finitely many elements $s_i \in \Gamma \setminus \{1\}$ such that $B \cap s_i(B) \neq \emptyset$. Let $S$ be the subset of $\Gamma$ consisting of these elements $s_i$. Since each $s_i$ is an isometry, $s_i^{-1}$ belongs to $S$ iff $s_i$ does. Let

$$r := \inf \big\{ d(B, g(B)) : g \in \Gamma \setminus (S \cup \{1\}) \big\}.$$

Observe that $r > 0$. Indeed, if we denote by $B'$ closed ball with center $x_0$ and radius $R + 1$, then for all but finitely many $g \in \Gamma$ we will have $B' \cap g(B) = \emptyset$, hence $d(B, g(B)) \geq 1$, therefore the infimum above is a minimum of finitely many positive numbers, so is positive. The claim now is that $S$ generates $\Gamma$, and for each $g \in \Gamma$,

$$\frac{d(x,g(x))}{2R} \le \|g\|_S \le \frac{d(x,g(x))}{r} + 1\,, \qquad (3.2)$$ {e.Jqisom}

where $\|g\|_S = d_S(1,g)$ is the word norm on $\Gamma$ with respect to the generating set $S$. Since, in the right Cayley graph w.r.t. $S$, we have $d_S(h,hg) = d_S(1,g)$ for any $g,h \in \Gamma$, this will mean that $J_x(g) := g(x)$ is coarsely bi-Lipschitz. Furthermore, since $\pi(B) = X/\Gamma$, for each $y \in X$ there is some $g \in \Gamma$ with $d(y, g(x)) \le R$, hence $J_x$ will indeed be a quasi-isometry from $\Gamma$ to $X$.

Let $g \in \Gamma$, and connect $x$ to $g(x)$ by a geodesic $\gamma$. Let $m$ be the unique nonnegative integer with

$$(m-1)r + R \le d(x,g(x)) < mr + R\,.$$

Choose points $x_0 = x, x_1, \dots, x_{m+1} = g(x) \in \gamma$ such that $x_1 \in B$ and $d(x_j, x_{j+1}) < r$ for $1 \le j \le m$. Each $x_j$ belongs to some $g_j(B)$ for some $g_j \in \Gamma$ (and take $g_{m+1} = g$ and $g_0 = 1$). Observe that

$$d(B, g_j^{-1}(g_{j+1}(B))) = d(g_j(B), g_{j+1}(B)) \le d(x_j, x_{j+1}) < r\,,$$

hence the balls $B$ and $g_j^{-1}(g_{j+1}(B))$ intersect, and $g_{j+1} = g_j s_{i(j)}$ for some $s_{i(j)} \in S \cup \{1\}$. Therefore,

$$g = s_{i(1)} s_{i(2)} \cdots s_{i(m)}\,,$$

and $S$ is a generating set for $\Gamma$. Moreover,

$$\|g\|_S \le m \le \frac{d(x,g(x)) - R}{r} + 1 < \frac{d(x),g(x))}{r} + 1\,.$$

On the other hand, the triangle inequality implies that

$$d(x,g(x)) \le 2R\,\|g\|_S,$$

because, for any $y \in X$ and $s \in S$, we have $B_R(y) \cap B_R(s(y)) \neq \emptyset$ and hence $d(y,s(y)) \le 2R$. This finishes the proof of (3.2). □

This lemma implies that if $\Gamma_1$ and $\Gamma_2$ both act nicely (cocompactly, etc.) on a nice metric space, then they are quasi-isometric. A small generalization and a converse are provided by the following characterization noted in [Gro93]:

{p.qisomcomm}

**Proposition 3.7.** *Two f.g. groups $\Gamma_1$ and $\Gamma_2$ are quasi-isometric iff there is a locally compact topological space $X$ where they both act properly discontinuously and cocompactly, moreover, the two actions commute with each other.*

{ex.qisomcomm}

▷ **Exercise 3.7.** *Prove the above proposition. (Hint: given the quasi-isometric groups, the space $X$ can be constructed using the set of all quasi-isometries between $\Gamma_1$ and $\Gamma_2$.)*

An important consequence of Lemma 3.6 is the following:

{cor.FiniteInd

**Corollary 3.8.** *A finite index subgroup $H$ of a finitely generated group $\Gamma$ is itself finitely generated and quasi-isometric to $\Gamma$. The same conclusions hold if $H$ is a factor of $\Gamma$ with a finite kernel.*

*Proof.* For the case $H \leq \Gamma$, we consider an "extended" Cayley graph of $\Gamma$ with respect to some generating set. This is a metric space which contains not only vertices of the Cayley graph but also points on edges, so that each edge is a geodesic of length 1. $H$ naturally acts on it by isometries and clearly this action satisfies all properties in the statement of the Milnor-Schwarz lemma $\Rightarrow$ we can just apply it.

For the case when $H$ is a factor of $\Gamma$, the image of any generating set of $\Gamma$ generates $H$, so it is finitely generated, and the action of $\Gamma$ on the Cayley graph of $H$ satisfies the conditions of the Milnor-Schwarz lemma. □

Based on this corollary, we will usually be interested in group properties that hold up to moving to a finite index subgroup or factor group or group extension. In particular, we say that two groups, $\Gamma_1$ and $\Gamma_2$, are **virtually** or **almost isomorphic**, if there are finite index subgroups $H_i \leq \Gamma_i$ and finite normal subgroups $F_i \triangleleft H_i$ such that $H_1/F_1 \simeq H_2/F_2$. Correspondingly, if a group $\Gamma$ is virtually isomorphic to a group that has some property $\mathcal{P}$, we will say that $\Gamma$ **almost** or **virtually** has $\mathcal{P}$.

Proposition 3.7 above is the geometric analogue of the following characterization of virtual isomorphisms:

{ex.isomcomm}

▷ **Exercise 3.8.** *Show that two f.g. groups, $\Gamma_1$ and $\Gamma_2$, are virtually isomorphic iff they admit commuting actions on a set $X$ such that the factors $X/\Gamma_i$ and all the stabilizers $\mathrm{St}_{\Gamma_i}(x)$, $x \in X$, are finite.*

A key example of geometric group theory is the action of $\mathsf{SL}_2(\mathbb{R})$ on the hyperbolic plane $\mathbb{H}^2 = \{z \in \mathbb{C} : \Im z > 0\}$ by **Möbius transformations**:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : \; z \mapsto \frac{az+b}{cz+d} \,. \tag{3.3}$$ {e.Moebius}

It is immediate to check that this is indeed a group action, and that these are isometries of $\mathbb{H}^2$. In fact, these are all the isometries of $\mathbb{H}^2$; more precisely, the kernel of this map is $\{\pm I\}$, and the isometry group of $\mathbb{H}^2$ is $\mathsf{PSL}_2(\mathbb{R}) = \mathsf{SL}_2(\mathbb{R})/\{\pm I\}$. The stabilizer of $i \in \mathbb{H}^2$ in $\mathsf{SL}_2(\mathbb{R})$ is $\mathsf{SO}(2)$, a compact subgroup, and we get $\mathbb{H}^2$ as a quotient of Lie groups, $\mathsf{SL}_2(\mathbb{R})/\mathsf{SO}(2)$. An analogue of the Milnor-Schwarz Lemma 3.6 says that the Lie group $\mathsf{SL}_2(\mathbb{R})$ is quasi-isometric to $\mathbb{H}^2$.

An important discrete subgroup is the **modular group** $\mathsf{PSL}_2(\mathbb{Z}) = \mathsf{SL}_2(\mathbb{Z})/\{\pm I\}$. It is generated, for instance, by an inversion and a translation:

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} : \; z \mapsto -1/z \qquad \text{and} \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : \; z \mapsto z+1 \,.$$

{ex.modular}

▷ **Exercise 3.9** (Alperin 1993)**.** *Using the Ping-Pong Lemma 2.19, show that $\mathsf{PSL}_2(\mathbb{Z}) = \mathbb{Z}_2 * \mathbb{Z}_3$, with $\mathbb{Z}_2$ generated by $S$ and $\mathbb{Z}_3$ by $ST$.*

A fundamental domain for the action of $\mathsf{PSL}_2(\mathbb{Z})$ is $D = \{z \in \mathbb{H}^2 : |z| > 1, |\Re z| < 1/2\}$; see Figure 3.1. This domain has finite volume in the volume measure generated by the Riemannian hyperbolic metric, but it is not compact. Thus, the Milnor-Schwarz Lemma 3.6 does not apply, and indeed, $\mathsf{PSL}_2(\mathbb{Z})$ and $\mathsf{SL}_2(\mathbb{Z})$ are not quasi-isometric to $\mathsf{PSL}_2(\mathbb{R})$ and $\mathbb{H}^2$. A simple way to see this is that a version Exercise 3.3 implies that any group quasi-isometric to $\mathbb{H}^2$ has one end, while Exercise 3.9 implies that $\mathsf{SL}_2(\mathbb{Z})$ is quasi-isometric to the free group $F_2$, with continuum many ends.

The discrete subgroups of $\mathsf{PSL}_2(\mathbb{R})$ are called **Fuchsian groups**; these are the crystallographic groups of the hyperbolic plane. For an introduction in hyperbolic geometry and Fuchsian groups, see [Bea83].

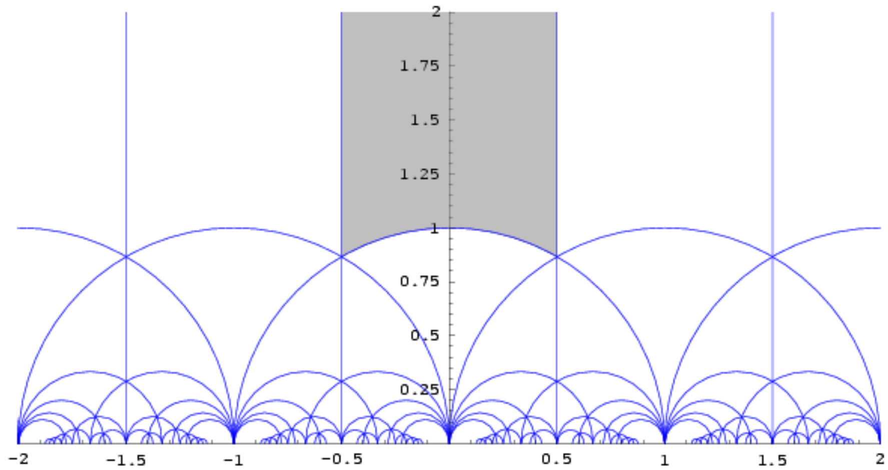



Figure 3.1: A fundamental domain of the action of the modular group $\mathsf{PSL}_2(\mathbb{Z})$ on $\mathbb{H}^2$. Picture by Fropuff, via Wikimedia Commons. {f.Modular}

Gromov proposed in [Gro93] the long-term project of quasi-isometric classification of all f.g. groups. This is a huge research area, with connections to a lot of parts of mathematics, including probability theory, as we will see.

Here is one more instance of realizing a group theoretical notion via geometric properties of group actions. Recall the notion of residually finite groups from Exercise 2.8.

▷ **Exercise 3.10** ([dlHar00]).*

**(a)** *Show that a group $\Gamma$ is residually finite iff it has a faithful* **chaotic action** *on a Hausdorff topological space $X$ by homeomorphisms, meaning the following two properties:*

(i) *the union of all finite orbits is dense in $X$;*

(ii) *the action is topologically transitive: for any nonempty open $U, V \subseteq X$ there is $\gamma \in \Gamma$ such that $\gamma(U) \cap V \neq \emptyset$.*

*(A rather obvious hint: start by looking at finite groups.)*

**(b)** *Construct such a chaotic action of $\mathsf{SL}(n, \mathbb{Z})$ on the torus $\mathbb{T}^n$.*

### 3.2 Gromov-hyperbolic spaces and groups

{ss.hyperbolic

### 3.3 Asymptotic cones

{ss.cones}

## 4 Nilpotent and solvable groups

{s.nilpsolv}

We have studied the free groups, which are the "largest groups" from most points of view. We have also seen that it is relatively hard to produce new groups from them: their subgroups are all free, as well, while taking quotients, i.e., defining groups using presentations, has the disadvantage that we might not know what the group is that we have just defined. On the other hand, the "smallest" infinite group is certainly $\mathbb{Z}$, and it is more than natural to start building new groups from it. Recall that a finitely generated group is Abelian if and only if it is a direct product of cyclic groups, and the number of (free) cyclic factors is called the (free) rank of the group. We understand these examples very well, so we should now go beyond commutativity.

## 4.1 The basics {ss.nilpsolvba

Recall that the commutator is defined by $[g,h] = ghg^{-1}h^{-1}$. {d.nilpsolv}

**Definition 4.1.** *A group $\Gamma$ is called* **nilpotent** *if the* **lower central series** $\Gamma_0 = \Gamma$, $\Gamma_{n+1} = [\Gamma_n, \Gamma]$ *terminates at $\Gamma_s = \{1\}$ in finite steps. If $s$ is the smallest such index, $\Gamma$ is called $s$-step nilpotent.*
*A group $\Gamma$ is called* **solvable** *if the* **derived series** $\Gamma^{(0)} = \Gamma$, $\Gamma^{(n+1)} = [\Gamma^{(n)}, \Gamma^{(n)}]$ *terminates at $\Gamma^{(s)} = \{1\}$ in finite steps. If $s$ is the smallest such index, $\Gamma$ is called $s$-step solvable.*

Clearly, nilpotent implies solvable. Before discussing any further properties of such groups, let us give the simplest non-Abelian nilpotent example.

The **3-dimensional discrete Heisenberg group** is the matrix group

$$\Gamma = \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : x, y, z \in \mathbb{Z} \right\}.$$

If we denote by $X, Y, Z$ the matrices given by the three permutations of the entries $1, 0, 0$ for $x, y, z$, then $\Gamma$ is given by the presentation

$$\big\langle X, Y, Z \,\big|\, [X,Z] = 1, [Y,Z] = 1, [X,Y] = Z \big\rangle.$$

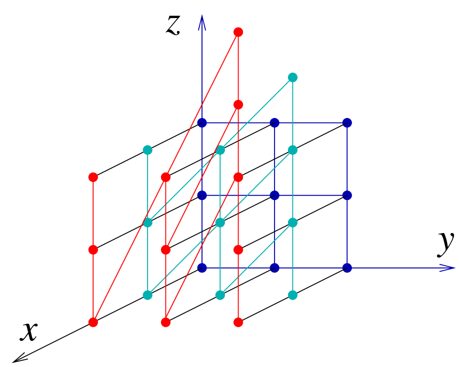



Figure 4.1: The Cayley graph of the Heisenberg group with generators $X, Y, Z$. {f.Heisenberg}

Clearly, it is also generated by just $X$ and $Y$. Note that $[X^m, Y^n] = Z^{mn}$. Furthermore, $[\Gamma, \Gamma] = \langle Z \rangle$, and the center (the subgroup of elements that commute with everything) is $Z(\Gamma) = \langle Z \rangle$, hence $\Gamma$ is 2-step nilpotent.

{ex.HeisenGrow

▷ **Exercise 4.1.** *Show that the Heisenberg group has 4-dimensional volume growth.*

Whenever $\Gamma^* \triangleleft \Gamma$, we have that $[\Gamma^*, \Gamma]$ is a subgroup of $\Gamma^*$ that is normal in $\Gamma$: if $g, h \in \Gamma$ and $\gamma \in \Gamma^*$, then $[\gamma, g] = \gamma(g\gamma^{-1}g^{-1}) \in \Gamma^*$ and $h^{-1}[\gamma, g]h = (h^{-1}\gamma h)(h^{-1}gh)(h^{-1}\gamma^{-1}h)(h^{-1}g^{-1}h) \in [\Gamma^*, \Gamma]$. Therefore, in the nilpotent case, by induction, $\Gamma_{n+1} \leq \Gamma_n \triangleleft \Gamma$, while in the solvable case, $\Gamma^{(n+1)} \triangleleft \Gamma^{(n)} \triangleleft \Gamma$. The factor $\Gamma^{\mathrm{A}b} := \Gamma/[\Gamma, \Gamma]$ is called the **Abelianization** of $\Gamma$, the largest Abelian factor of $\Gamma$:

{ex.solvable}

▷ **Exercise 4.2.**

**(a)** *Prove that if $f : \Gamma \longrightarrow A$ is a surjective homomorphism with $A$ being Abelian, then $f$ can be written as a composition of factor maps $\Gamma \longrightarrow \Gamma^{\mathrm{A}b} \longrightarrow A$.*

**(b)** *A group $\Gamma$ is solvable iff there is a finite sequence of subgroups $\Gamma = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_k = \{1\}$ such that each factor $H_i/H_{i+1}$ is Abelian.*

It is not true that, for any finitely generated group $\Gamma$, the subgroup $[\Gamma,\Gamma]$ is generated by the commutators of the generators of $\Gamma$. For instance, we will define in Section 5.1 the lamplighter group $\Gamma = \mathbb{Z}_2 \wr \mathbb{Z}$, a finitely generated solvable group, for which $[\Gamma,\Gamma] = \oplus_{\mathbb{Z}}\mathbb{Z}_2$ is not finitely generated. However, if $\Gamma$ is nilpotent, then, by simple word manipulations one can show that each $\Gamma_n$ is finitely generated by iterated commutators of the generators of $\Gamma$, as follows. In any commutator of two words on the generators of $\Gamma$, there is an equal number of $x$ and $x^{-1}$ letters for each generator $x$. By introducing commutators, we can move such pairs of letters towards each other until they become neighbours and annihilate each other. So, at the end, we are left with a product of iterated commutators. For instance,

$$\begin{aligned}
[a,bc] &= abca^{-1}c^{-1}b^{-1} = ab[c,a^{-1}]a^{-1}cc^{-1}b^{-1} = [a,b]ba[c,a^{-1}]a^{-1}b^{-1} \\
&= [a,b]b\big[a,[c,a^{-1}]\big][c,a^{-1}]aa^{-1}b^{-1} = [a,b]\Big[b,\big[a,[c,a^{-1}]\big]\Big]\big[a,[c,a^{-1}]\big]bb^{-1}[c,a^{-1}]\big[[c,a^{-1}]^{-1},b\big] \\
&= [a,b]\Big[b,\big[a,[c,a^{-1}]\big]\Big]\big[a,[c,a^{-1}]\big][c,a^{-1}]\big[[a^{-1},c],b\big]\,.
\end{aligned}$$

In the resulting word, the iteration depths of the commutators depend on the two words we started with, but since $\Gamma$ is $s$-step nilpotent, we can just delete from the word any iteration of depth larger than $s$. This way we get a finite set of generators for $[\Gamma,\Gamma]$, and then for each $\Gamma_n$, as well.

The fact that for any finitely generated nilpotent group $\Gamma$, each $\Gamma_n$ is finitely generated, together with the next lemma, implies that $\Gamma$ is **polycyclic**, i.e., there is a subgroup sequence $\Gamma = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_k = \{1\}$ such that each factor $H_i/H_{i+1}$ is cyclic:

{l.polycyc}

**Lemma 4.2.** *A solvable group is polycyclic iff all subgroups are finitely generated.*

*Proof.* If a group $\Gamma$ is polycyclic, then it is solvable by Exercise 4.2 (b) above, and finitely generated by Exercise 4.3 below. Furthermore, any subgroup $\Gamma' \leq \Gamma$ is also polycyclic: just consider the chain $H_i' := \Gamma' \cap H_i$. Thus $\Gamma'$ is also finitely generated, and we are done with the forward direction.

If $\Gamma$ is finitely generated, then the factor $\Gamma/[\Gamma,\Gamma]$ is finitely generated and Abelian, hence is a finite direct product of cyclic groups, $A_1 \oplus \cdots \oplus A_k$. By projecting further to the first coordinate, we get a factor map $\Gamma \longrightarrow A_1$, whose kernel $H_{1,1}$ projects onto $A_2 \oplus \cdots \oplus A_k$. By iterating this, we obtain a chain $\Gamma \triangleright H_{1,1} \triangleright \cdots \triangleright H_{1,k} = \Gamma^{(1)}$, with $H_{1,i}/H_{1,i+1}$ being cyclic. By hypothesis, $\Gamma^{(1)}$ is also finitely generated, hence we can now do the same for $\Gamma^{(2)} = [\Gamma^{(1)},\Gamma^{(1)}]$, and so on. Since $\Gamma$ is solvable, in a finite number of steps we get to $H_{s,k_s} = \{1\}$, showing that $\Gamma$ is polycyclic. □

For future reference, note that the above proof also shows that if $\Gamma$ is a finitely generated group and the factor $\Gamma/[\Gamma,\Gamma]$ is infinite, then in its decomposition into a finite direct product of cyclic groups one of the factors must be infinite, and hence we obtain a surjection of $\Gamma$ onto $\mathbb{Z}$.

{ex.finpres}

▷ **Exercise 4.3.**

**(a)** *Assume that for some $H \triangleleft \Gamma$, both $H$ and $\Gamma/H$ are finitely generated. Show that $\Gamma$ is also finitely generated. Same for finitely presented.*

**(b)** *Show that any finitely generated almost-nilpotent group is finitely presented.*

## 4.2 Semidirect products

{ss.semidirect

We do not yet have any general procedure to build non-commutative nilpotent and solvable groups. Such a procedure is given by the **semidirect product**.

For any group $N$, let $\mathrm{Aut}(N)$ be the group of its group-automorphisms, acting on the right, i.e., $f$ acts by $a^f = f(a)$ and $fg$ acts by $a^{fg} = (a^f)^g = g(f(a))$. Further, let $\varphi : H \longrightarrow \mathrm{Aut}(N)$ be some group homomorphism. Then define the semidirect product $N \rtimes_\varphi H$ by having the elements of $N \times H$, and group multiplication

$$(a, g)\,(b, h) = (a^{\varphi(h)} b, gh)\,, \qquad a, b \in N \text{ and } g, h \in H\,.$$

It is easy to check that this is indeed a group; in particular, $(a, f)^{-1} = (f^{-1}(a^{-1}), f^{-1})$. Furthermore, $N \lhd \Gamma := N \rtimes_\varphi H$, with the inclusion $a \mapsto (a, \mathrm{id})$, and $H \le \Gamma$ with the inclusion $f \mapsto (1, f)$. Then $fa = (1, f)(a, \mathrm{id}) = (a, f)$, so one might prefer writing $\Gamma = H\,{}_\varphi\!\ltimes N = \{(h, a) : h \in H, a \in N\}$. Anyway, we have $HN = \Gamma$ and $H \cap N = \{1_G\}$. Conversely, whenever we have a group $\Gamma$ with subgroups $N \lhd \Gamma$ and $H \le \Gamma$ satisfying $HN = \Gamma$ and $H \cap N = \{1\}$, then $H$ acts on $N$ by conjugations, $a^{\varphi(h)} := h^{-1}ah$, and it is easy to check that $\Gamma = H\,{}_\varphi\!\ltimes N$.

Of course, for the trivial homomorphism $\varphi : H \longrightarrow \{\mathrm{id}\} \subset \mathrm{Aut}(N)$ we get the direct product of $N$ and $H$. The archetypical non-trivial examples are the affine group $\mathbb{R}^d \rtimes \mathsf{GL}_d(\mathbb{R}) = \{v \mapsto vA + b : A \in \mathsf{GL}_d(\mathbb{R}), b \in \mathbb{R}^d\}$ and the group of Euclidean isometries $\mathbb{R}^d \rtimes \mathsf{O}(d)$, with the product being just composition, $v \mapsto vA_1 + b_1 \mapsto vA_1A_2 + b_1A_2 + b_2$. Similar examples are possible with $\mathbb{Z}^d$ instead of $\mathbb{R}^d$, but note that $\mathrm{Aut}(\mathbb{Z}^d) = \mathsf{GL}_d(\mathbb{Z})$ is just $\pm\mathsf{SL}_d(\mathbb{Z})$. One more small example of a semidirect product is in Exercise 3.6.

{ex.cyclicsemi

▷ **Exercise 4.4.** *For what primes $p, q$ is there a semidirect product $\mathbb{Z}_p \rtimes \mathbb{Z}_q$ that is not a direct product? (Hint: you do not have to know what exactly the group $\mathrm{Aut}(\mathbb{Z}_p)$ is; it is enough to use Cauchy's theorem on having an element of order $q$ in any group of size $n$, for any prime $q|n$. But, in fact, $\mathrm{Aut}(\mathbb{Z}_p)$ is always cyclic, because there exist primitive roots modulo any prime.)*

A fancy way of saying that $N \lhd \Gamma$ and $\Gamma/N \simeq F$ is to write down the **short exact sequence**

$$1 \longrightarrow N \xrightarrow{\alpha} \Gamma \xrightarrow{\beta} F \longrightarrow 1\,,$$

which just means that the image of each map indicated by an arrow is exactly the kernel of the next map. (When I was a Part III student in Cambridge, UK, a famous number theory professor, Sir Swinnerton-Dyer, wrote down short exact sequences just to prove he was not hopelessly old and old-fashioned, he said.) If, in addition, there is an injective homomorphism $\gamma : F \longrightarrow \Gamma$ with $\beta \circ \gamma = \mathrm{id}_F$, then we say that the short exact sequence **splits**. In this case, $\gamma(F) \cap \alpha(N) = \{1\}$ and $\gamma(F)\alpha(N) = \Gamma$, hence $\Gamma \simeq N \rtimes_\varphi F$ is a semidirect product with $\varphi : F \longrightarrow \mathrm{Aut}(N)$ given by conjugation: $a^{\varphi(f)} := \alpha^{-1}\big(\gamma(f)^{-1}\alpha(a)\gamma(f)\big)$. This splitting does not always happen; for instance, the exact sequence

$$1 \longrightarrow \mathbb{Z}_2 \longrightarrow \mathbb{Z}_4 \longrightarrow \mathbb{Z}_2 \longrightarrow 1$$

defined by $\{0, 2\} \lhd \{0, 1, 2, 3\}$ does not split, since $\mathrm{Aut}(\mathbb{Z}_2) = \{\mathrm{id}\}$ means that all semidirect products are actually direct products, but $\mathbb{Z}_2 \times \mathbb{Z}_2 \not\simeq \mathbb{Z}_4$. (Or, the only $\mathbb{Z}_2$ subgroup of $\mathbb{Z}_4$ does not have trivial intersection with the given normal subgroup $\{0, 2\}$, hence there is no good $\gamma$.) Similarly,

$$1 \longrightarrow [\Gamma, \Gamma] \longrightarrow \Gamma \longrightarrow \Gamma^{Ab} \longrightarrow 1$$

does not always split. Nevertheless, a lot of solvable and nilpotent groups can be produced using semidirect products, as we will see in a second. It will also be important to us that any sequence

$$1 \longrightarrow N \longrightarrow \Gamma \xrightarrow{\pi} \mathbb{Z} \longrightarrow 1 \tag{4.1}$$

{e.NGZ}

does split. This is because if $x \in \pi^{-1}(1)$, then $\pi(x^k) = k \neq 0$ for all $k \in \mathbb{Z} \setminus \{0\}$, while $\ker(\pi) = N$, hence for $H := \langle x \rangle \simeq \mathbb{Z}$ we have $H \cap N = \{1\}$ and $HN = \Gamma$.

{ex.semidirect

▷ **Exercise 4.5.**

**(a)** *Show that*

$$(a,g)(b,h)(a,g)^{-1}(b,h)^{-1} = \left(a^{\varphi(hg^{-1}h^{-1})}b^{\varphi(g^{-1}h^{-1})}(a^{-1})^{\varphi(g^{-1}h^{-1})}(b^{-1})^{\varphi(h^{-1})}, ghg^{-1}h^{-1}\right),$$

*and using this, show that if $N$ and $H$ are solvable, then $N \rtimes_\varphi H$ is also.*

**(b)** *Given $N \rtimes \langle g \rangle$, $g \in \mathrm{Aut}(N)$, show that*

$$N \rtimes \langle g^k \rangle \quad k \in \mathbb{Z} \setminus \{0\}$$

*is a finite index subgroup of $N \rtimes \langle g \rangle$.*

{ex.almostnilp

▷ **Exercise 4.6.** *Show that if $\Gamma/K \simeq N$ with $K$ a finite normal subgroup and $N$ nilpotent, then $\Gamma$ has a finite index nilpotent subgroup. Also, obviously, any subgroup or factor group of a nilpotent group is also nilpotent. Therefore, an almost nilpotent group in the sense given at the end of Section 3.1 has a finite index nilpotent subgroup. (Hint: use the $\mathbb{Z}$ factor from the remark after Lemma 4.2 and the splitting in (4.1).)*

Now, given $\Gamma = \mathbb{Z}^d \rtimes_M \mathbb{Z}$, $M \in \mathsf{GL}_d(\mathbb{Z})$, we will prove the following:

{p.NilpOrExp}

**Proposition 4.3.** *If $M$ has only absolute value 1 eigenvalues, then $\Gamma$ is almost nilpotent. If not, then $\Gamma$ has exponential volume growth.*

In fact, this is true in larger generality. The proof of the following theorem will be given in the next section (with a few details omitted):

{t.MilnorWolf}

**Theorem 4.4** (Milnor-Wolf [Mil68a, Wol68, Mil68b])**.** *A finitely generated almost-solvable group is of polynomial growth if and only if it is almost-nilpotent, and is of exponential growth otherwise.*

Moreover, the **Bass-Guivarch formula** (which we will not prove) states that if $d_{i+1} := \text{free-rank}(\Gamma_i/\Gamma_{i+1})$ for the lower central series, then the volume growth of a nilpotent $\Gamma$ is

$$d(\Gamma) = \sum_i i d_i\,.$$

For example, for the Heisenberg group we have $d_1 = 2$ and $d_2 = 1$, hence $d(\Gamma) = 4$, agreeing with Exercise 4.1.

To prove Proposition 4.3, we shall first show the following lemma. I learnt it from [DrKa09], with its nice elementary proof coming from [Ros74].

{l.rootsofunit

**Lemma 4.5.** *If $M \in \mathsf{GL}_d(\mathbb{Z})$ has only eigenvalues with absolute value 1, then all of them are roots of unity.*

*Proof.* Let $\lambda_1, \dots \lambda_d$ be the set of eigenvalues of $M$. Then we have

$$\mathrm{Tr}(M^k) = \sum_{i=1}^d \lambda_i^k \in \mathbb{Z}.$$

Now consider the sequence $v_k = (\lambda_1^k, \dots, \lambda_d^k) \in (S^1)^d$ as $k = 0, 1, 2, \dots$. Since $(S_1)^d$ is a compact group, there exists a convergent subsequence $\{v_{k_\ell}\}$ of $\{v_k\}$, and hence $v_{k_{\ell+1}} v_{k_\ell}^{-1} \to 1$ as $\ell \to \infty$. Let $m_\ell = k_{\ell+1} - k_\ell$, then $(\lambda_1^{m_\ell}, \dots, \lambda_d^{m_\ell}) \to 1$ and $\sum_{i=1}^d \lambda_i^{m_\ell} \to d$ as $\ell \to \infty$. But $\sum_{i=1}^d \lambda_i^{m_\ell} \in \mathbb{Z}$. This implies that there exists $m_\ell$ such that $\sum_{i=1}^d \lambda_i^{m_\ell} = d$. But $|\lambda_i^{m_\ell}| = 1 \forall i$. Thus, $\lambda_i^{m_\ell} = 1 \forall i$. □

**Definition 4.6.** *A matrix $M$ is* **unipotent** *if all eigenvalues are equal to* 1 *and* **quasi-unipotent** *if all eigenvalues are roots of unity.*

Note that a matrix $M$ is quasi-unipotent if and only if there exists $m < \infty$ such that $M^m$ is unipotent.

Thus, by Exercise 4.5 part (b), for the first part of Proposition 4.3 it is enough to show that if $M$ is unipotent, then $\Gamma = \mathbb{Z}^d \rtimes_M \mathbb{Z}$ is nilpotent.

{ex.unipotent}

▷ **Exercise 4.7.** *If $M \in \mathsf{GL}_d(\mathbb{Z})$ is unipotent, then*

$$M = I + S^{-1}NS = S^{-1}(I+N)S$$

*for some strictly upper-triangular integer matrix $N$.*

Note that $N^d = 0$ in the exercise. Then, for any $m \in \mathbb{Z}_+$,

$$\begin{aligned} M^m &= S^{-1}(N^m + \binom{m}{1}N^{m-1} + \ldots + I)S \\ M^m - I &= S^{-1}(N^m + \binom{m}{1}N^{m-1} + \ldots + N)S, \end{aligned}$$

which implies that $(M^m - I)^d = 0$. Since the inverse of a unipotent matrix is again unipotent, the last result holds, in fact, for any $m \in \mathbb{Z}$.

What is now $\Gamma_1 = [\Gamma, \Gamma]$? By Exercise 4.5 part (a),

$$\begin{aligned} (v, M^i)(w, M^j)(v, M^i)^{-1}(w, M^j)^{-1} &= (vM^{-i} + wM^{-i-j} - vM^{-i-j} - wM^{-j}, I) \\ &= \big(vM^{-i-j}(M^j - I) + wM^{-i-j}(I - M^i), I\big)\,. \end{aligned}$$

So, the matrix part has got annihilated, while, since $M^i - I$ lowers dimension for any $i \in \mathbb{Z}$, the "vector part" has a smaller and smaller support:

$$\begin{aligned} \Gamma_1 &\subset (d-1)\,\mathrm{dim\,subspace} \\ \Gamma_2 = [\Gamma_1, \Gamma_1] &\subset (d-2)\,\mathrm{dim\,subspace} \\ &\vdots \\ \Gamma_d &= \{1\}\,, \end{aligned}$$

hence $\Gamma$ is at most $d$-step nilpotent, proving the first half of Proposition 4.3.

{ex.HeisenSemi

▷ **Exercise 4.8.** *Show that $\mathbb{Z}^2 \rtimes_M \mathbb{Z}$ with $M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is isomorphic to the Heisenberg group.*

For the second half of Proposition 4.3, the key step is the following (again learnt from [DrKa09]):

{l.alldifferen

**Lemma 4.7.** *If $M \in \mathsf{GL}_d(\mathbb{Z})$ has an eigenvalue $|\lambda| > 2$, then $\exists v \in \mathbb{Z}^d$ such that for any $k \in \mathbb{N}$ and $\epsilon_i \in \{0, 1\}$, the $2^{k+1}$ vectors*

$$\epsilon_0 v + \epsilon_1 vM + \ldots + \epsilon_k vM^k$$

*are all different.*

*Proof.* Note that $M^T$ also has the eigenvalue $\lambda$. Let $b \in \mathbb{C}^d$ a corresponding eigenvector, $bM^T = \lambda b$, or $Mb^T = \lambda b^T$. Define the linear form $\beta(v) = vb^T$. Then we have:

$$\beta(vM^k) = vM^k b^T = v\lambda^k b^T = \lambda^k \beta(v)$$

Since $\beta : \mathbb{C}^d \longrightarrow \mathbb{C}$ is linear and non-zero, its kernel is $(d-1)$-dimensional, hence there exists $v \in \mathbb{Z}^d$ such that $\beta(v) \neq 0$. Now suppose that

$$\sum_{i=0}^{k} \delta_i \lambda^i \beta(v) = \beta\Big(\sum_{i=0}^{k} \delta_i v M^i\Big) = \beta\Big(\sum_{i=0}^{k} \epsilon_i v M^i\Big) = \sum_{i=0}^{k} \epsilon_i \lambda^i \beta(v)$$

with $\epsilon_m - \delta_m \neq 0$ for some $m \leq k$, but $\epsilon_i = \delta_i$ for all $m < i \leq k$. Let $\eta_i = \epsilon_i - \delta_i$. Then

$$\sum_{i=0}^{m} \eta_i \lambda^i \beta(v) = 0\,.$$

Now consider $\big|\sum_{i=0}^{m-1} \eta_i \lambda^i\big|$. On one hand, we have

$$\Big|\sum_{i=0}^{m-1} \eta_i \lambda^i\Big| = |\eta_m \lambda^m| = |\lambda|^m\,.$$

On the other hand,

$$\Big|\sum_{i=0}^{m-1} \eta_i \lambda^i\Big| \leq \frac{|\lambda|^m - 1}{|\lambda| - 1} \leq |\lambda|^m - 1\,,$$

since $|\lambda| > 2$. This is a contradiction, hence we have found the desired $v$. □

Now, to finish the proof of Proposition 4.3, assume that $M$ has an eigenvalue with absolute value not 1. Since $|\det(M)| = 1$, there is an eigenvalue $|\lambda| > 1$. Then, there is an $m < \infty$ such that $M^m$ has the eigenvalue $|\lambda^m| > 2$. Consider the products

$$(\epsilon_k v, M^m)(\epsilon_{k-1} v, M^m) \cdots (\epsilon_0 v, M^m) = (\epsilon_k v M^{mk} + \epsilon_{k-1} v M^{m(k-1)} + \cdots + \epsilon_0 v, M^{m(k+1)}),$$

with $\epsilon_i \in \{0,1\}$ and the vector $v \in \mathbb{Z}^d$ given by Lemma 4.7. By that lemma, for a fixed $k$ these elements are all different. But, in the Cayley graph given by any generating set containing $(v, I)$ and $(0, M^m)$, these elements are inside the ball of radius $2(k+1)$, hence $\Gamma$ has exponential growth. □

The proposition has the following generalization:

{e.nore}

▷ **Exercise 4.9.*** *Let $N$ be a finitely generated almost nilpotent group. Then $N \rtimes_\varphi \mathbb{Z}$ is almost nilpotent or of exponential growth (and this can be easily detected from $\varphi$).*

## 4.3 The volume growth of nilpotent and solvable groups

{ss.volnilpsol

As promised in the previous section, we want to go beyond semidirect products, and prove the Milnor-Wolf Theorem 4.4. The exposition is mostly based on [DrKa09], with some differences in organization and in using Exercise 4.11 that I was somehow missing from [DrKa09].

The first step is for free: we have seen that an exact sequence (4.1) always splits, hence Exercise 4.9 can be reformulated as follows:

{p.gnore}

**Proposition 4.8.** *Assume that*

$$1 \longrightarrow N \longrightarrow \Gamma \longrightarrow \mathbb{Z} \longrightarrow 1$$

*is a short exact sequence with $N$ being a finitely generated almost nilpotent group. Then, if $\Gamma$ is not almost nilpotent, then it has exponential growth.*

We will need two more ingredients, Propositions 4.9 and 4.11. We will prove the first one, but not the second, although that is not hard, either: the key idea is just to analyze carefully the procedure we used in Section 4.1 to prove that the subgroups $\Gamma_n$ in the lower central series are finitely generated. See [DrKa09] for the details.

{p.subtfg}

**Proposition 4.9.** *Assume we have a short exact sequence*

$$1 \longrightarrow N \longrightarrow \Gamma \xrightarrow{\pi} \mathbb{Z} \longrightarrow 1\,.$$

**(i)** *If $\Gamma$ has sub-exponential growth, then $N$ is finitely generated and also has sub-exponential growth.*
**(ii)** *Moreover, if $\Gamma$ has growth $O(R^d)$, then $N$ has growth $O(R^{d-1})$.*

{d.polydist}

**Definition 4.10.** *Let $H \leq \Gamma$, and $S_H$ and $S_\Gamma$ finite generating sets of $H$ and $\Gamma$ respectively, with $S_\Gamma \supset S_H$. Note that for the distance in the corresponding Cayley graphs, $d_H(e,h) =: \|h\|_H \geq \|h\|_\Gamma$ $\forall h \in H$. We say that $H$ has* **polynomial distortion** *if there is a polynomial $p(x)$ so that $p(\|h\|_G) \geq \|h\|_H$ $\forall h \in H$.*

{p.polydist}

**Proposition 4.11.** *If $\Gamma$ is a finitely generated nilpotent group, then $[\Gamma,\Gamma]$ has polynomial distortion.*

In fact, any subgroup has polynomial distortion, but we will not need that result.

▷ **Exercise 4.10.*** *Without looking into [DrKa09], but using the hint given in the paragraph before Proposition 4.9, prove the last proposition.*

**Anti-example**: Consider the solvable Baumslag-Solitar group,

$$\mathsf{BS}(1,m) := \langle\, a, b \mid a^{-1}ba = b^m \,\rangle\,.$$

A concrete realization of this presentation is to take the additive group $H$ of the rationals of the form $x/m^y$, $x, y \in \mathbb{Z}$, and then $\mathsf{BS} \simeq H \rtimes \mathbb{Z}$, where $t \in \mathbb{Z}$ acts on $H$ by multiplication by $m^t$. Indeed, we can take the generators $a : u \mapsto um$ and $b : u \mapsto u+1$. One can check that $[\mathsf{BS},\mathsf{BS}] = \langle b \rangle$, hence the group is two-step solvable. Furthermore, $a^{-n}ba^n = b^{m^n}$. So $\|b^{m^n}\|_{\langle b \rangle} = m^n$ but $\|b^{m^n}\|_{\mathsf{BS}} = 2n+1$. So $\langle b \rangle$ as a subgroup of $\mathsf{BS}(1,m)$ does not have polynomial distortion.

{t.nilpoly}

**Theorem 4.12.** *Finitely generated almost nilpotent groups have polynomial growth.*

*Proof.* The proof will use induction on the nilpotent rank. First of all, by Corollary 3.8, we can assume that $\Gamma$ is nilpotent. So, we have the short exact sequence

$$1 \longrightarrow [\Gamma,\Gamma] \longrightarrow \Gamma \xrightarrow{\pi} \Gamma^{Ab} \longrightarrow 1\,,$$

where $\Gamma_1 = [\Gamma,\Gamma]$ is nilpotent of strictly smaller rank. Moreover, as we showed in Section 4.1, it is finitely generated. Take a finite generating set $S$ of it. Let the Abelian rank of $\Gamma^{Ab}$ be $r'$ and its free-rank be $r \leq r'$. Take a generating set $e_1, \dots e_{r'}$ for $\Gamma^{Ab}$, with $r$ free generators, and let the $\pi$-preimages be $T = \{t_1, \dots, t_{r'}\}$. Then $\Gamma$ is generated by $S \cup T$.

Consider any word $w$ of length at most $R$ in $S \cup T$. Move all letters of $T$ to the front of $w$: since $\Gamma_1 \triangleleft \Gamma$, for any $g \in \Gamma_1$ and $t \in T$ there is some $g' \in \Gamma_1$ with $gt = tg'$. We get a word $w'$ representing the same group element, but of the form $w_T w_S$, where $w_T$ is a word on $T$ of length equal to the number of $T$-letters in the original $w$, while $w_S$ is a word on $S$, but its length might be much longer than the original number of $S$-letters in $w$. Now, since $\pi(w_T)$ in $\Gamma^{Ab}$ can be written in the form $e_1^{k_1} \cdots e_{r'}^{k_{r'}}$, there is some element $h \in [\Gamma, \Gamma]$ such that $w_T w_S = t_1^{k_1} \cdots t_{r'}^{k_{r'}} h$ in $\Gamma$.

Since $\|w\|_{S\cup T} \le R$, we also have $\|t_1^{k_1} \cdots t_{r'}^{k_{r'}}\|_{S\cup T} \le k_1 + \cdots + k_{r'} \le R$, and hence, by the triangle inequality, $\|h\|_{S\cup T} \le 2R$. But, by Proposition 4.11, $\Gamma_1$ has polynomial distortion, so, if $D$ is the degree of this polynomial for the generating sets $S$ and $S \cup T$, then $\|h\|_S \le O(R^D)$. By the induction hypothesis, $\Gamma_1$ has polynomial growth of some degree $d$, hence $|\{h \in \Gamma_1 : \|h\|_{S\cup T} \le 2R\}| = O(R^{Dd})$. Since the number of different possible words $t_1^{k_1} \cdots t_{r'}^{k_{r'}}$ is $O(R^r)$, we altogether have $|B_R^{S\cup T}| = O(R^{dD+r})$, so $\Gamma$ has polynomial growth. □

{t.solvnilorex

**Theorem 4.13.**
**(i)** *Any finitely generated almost solvable group of polynomial growth is almost nilpotent.*
**(ii)** *The statement remains true if only subexponential growth is assumed.*

*Proof.* The beginning and the overall strategy of the proof of the two cases are the same. By Corollary 3.8, we may assume that $\Gamma$ is infinite and solvable. Then, there is a first index $j \ge 0$ in its derived series such that $\Gamma^{(j)}/[\Gamma^{(j)}, \Gamma^{(j)}]$ is infinite. Since $[\Gamma : \Gamma^{(j)}] < \infty$, we can further assume that $j = 0$. So, by the argument at the end of Section 4.1, there is a short exact sequence

$$1 \longrightarrow N \longrightarrow \Gamma \longrightarrow \mathbb{Z} \longrightarrow 1\,. \qquad (4.2)$$ {e.1KGZ1}

It is also clear from that argument that $[N, N] = [\Gamma, \Gamma]$, so $N$ is solvable. By Proposition 4.9, it is also finitely generated. Furthermore, we know from the argument at (4.1) that this exact sequence splits.

We now prove (i) by induction on the degree of the polynomial growth of $\Gamma$. For degree 0, i.e., when $\Gamma$ is finite the statement is trivial. Now, if $\Gamma$ has volume growth $O(R^d)$, then Proposition 4.9 says that $N$ has growth $O(R^{d-1})$, so, by induction, it is almost nilpotent, and then Proposition 4.8 says that $\Gamma$ is also almost nilpotent, and we are done.

In case (ii), we know only that $\Gamma$ has subexponential growth. By Proposition 4.9, also $N$ does. So, we can iterate (4.2), with $N$ in place of $\Gamma$. If this procedure stops after finitely many steps, then in the last step $N$ is a finite group, which is trivially almost nilpotent, hence we can apply Proposition 4.8 iteratively to the short exact sequences we got, and find that $\Gamma$ at the top was also almost nilpotent.

However, it is not obvious that the splitting procedure (4.2) terminates. For instance, for the finitely generated solvable lamplighter group $\Gamma = \mathbb{Z} \wr \mathbb{Z}$ that we mentioned earlier, in the first step we can have $N = [\Gamma, \Gamma] = \oplus_{\mathbb{Z}} \mathbb{Z}$, then we can continue splitting off a $\mathbb{Z}$ factor ad infinitum. Of course, this $\Gamma$ has exponential growth and $N = [\Gamma, \Gamma]$ is not finitely generated, which suggests that subexponential growth should crucially be used. This is done in the next exercise, finishing the proof. □

{ex.inftysplit

▷ **Exercise 4.11.** * *Assume that $\Gamma$ is finitely generated and has infinitely many subgroups $\Gamma_1, \Gamma_2, \dots$ with the properties that for each $i \ge 1$, $\Gamma_i \simeq \mathbb{Z}$, and $\Gamma_i \cap \langle \Gamma_{i+1}, \Gamma_{i+2}, \dots \rangle = \{1\}$. Show that $\Gamma$ has exponential growth.*

*Proof of Proposition 4.9.* Let $\Gamma = \langle f_1, f_2, \dots, f_k \rangle$. Take $\gamma \in \Gamma$ so that $\pi(\gamma) = 1 \in \mathbb{Z}$. Now, for each $f_i$ pick $s_i \in \mathbb{Z}$ so that $\pi(f_i \gamma^{s_i}) = 0$. Let $g_i = f_i \gamma^{s_i} \in N$. Then $\Gamma = \langle g_1, \dots, g_k, \gamma \rangle$. If we let

$$S = \left\{ \gamma_{m,i} := \gamma^m g_i \gamma^{-m} \,\middle|\, m \in \mathbb{Z},\ i = 1, \dots, k \right\},$$

then $\langle S\rangle = N$, since for any $f \in \ker(\pi)$,

$$f = f_{i_1}\cdots f_{i_l} = g_{i_1}\gamma^{-s_{i_1}}\cdots g_{i_l}\gamma^{-s_{i_l}}$$
$$= (g_{i_1})\cdot(\gamma^{-s_{i_1}}g_{i_2}\gamma^{s_{i_1}})\cdot(\gamma^{-s_{i_1}-s_{i_2}}g_{i_3}\gamma^{s_{i_1}+s_{i_2}})\cdots(\gamma^{-\sum_{k=1}^{l-1}s_{i_k}}g_{i_l}\gamma^{-s_{i_l}}),$$

where the last factor is actually also a conjugated $g_{i_l}$, since $\sum_{k=1}^{l} s_{i_k} = 0$, due to $f \in \ker(\pi)$. So indeed $f \in \langle S\rangle$.

Now consider a fixed $i$ and the collection of $2^m$ words (for $m > 0$)

$$\left\{\gamma g_i^{\epsilon_1}\gamma g_i^{\epsilon_2}\cdots\gamma g_m^{\epsilon_m}\,\middle|\,\epsilon_j \in \{0,1\}\right\},$$

each of length at most $m$ on the generators $g_i$ and $\gamma g_i$. So, the subexponential growth of $\Gamma$ implies that there must exist some $m$ and $\epsilon_m \neq \delta_m$ such that

$$\gamma g_i^{\epsilon_1}\cdots\gamma g_i^{\epsilon_m} = \gamma g_i^{\delta_1}\cdots\gamma g_i^{\delta_m}.$$

Now notice that, somewhat miraculously,

$$\gamma g_i^{\epsilon_1}\cdots\gamma g_i^{\epsilon_m} = \gamma_{1,i}^{\epsilon_1}\cdots\gamma_{m,i}^{\epsilon_m}\gamma^m.$$

Thus our relation becomes

$$\gamma_{1,i}^{\epsilon_1}\cdots\gamma_{m,i}^{\epsilon_m} = \gamma_{1,i}^{\delta_1}\cdots\gamma_{m,i}^{\delta_m}.$$

By $\epsilon_m - \delta_m \neq 0$, this yields $\gamma_{m,i} \in \langle\gamma_{0,i},\dots,\gamma_{m-1,i}\rangle$. Since $\gamma_{m+1,i} = \gamma\cdot\gamma_{m,i}\cdot\gamma^{-1}$, we also get

$$\gamma_{m+1,i} \in \langle\gamma_{1,i},\dots,\gamma_{m,i}\rangle \subset \langle\gamma_{0,i},\dots,\gamma_{m-1,i}\rangle.$$

We can do the same argument for $m < 0$, and so get that $\langle\gamma_{n,i} \mid n \in \mathbb{Z}\rangle$ is finitely generated, and doing this for every $i$ gives that $N$ is finitely generated.

Let $Y$ be this particular finite generating set of $N$. Then $\Gamma = \langle Y \cup \{\gamma\}\rangle$. Let $B_R^Y$ be the ball of radius $R$ in $N$ with this generating set. Suppose $|B_R^Y| = \{h_1,\dots,h_K\}$, and consider the elements

$$\{h_i\gamma^k \mid -R \le k \le R\}.$$

If $h_i\gamma^k = h_j\gamma^l$, then $\gamma^{l-k} = h_j^{-1}h_i$, but $\gamma \notin N$, so $l = k$ and $i = j$. That is, these elements are distinct and belong to $B_{R+1}^{Y\cup\{\gamma\}}$. There are $K(2R+1)$ of them. So

$$|B_R^Y| = K \le \frac{|B_{R+1}^{Y\cup\{\gamma\}}|}{2R+1}.$$

This shows both parts (i) and (ii) of the proposition. $\square$

## 4.4 Expanding maps. Polynomial and intermediate volume growth

{ss.franks}

Let us start with some definitions.

{d.expanding}

**Definition 4.14.** *Let $X$ and $Y$ be metric spaces. Then $\varphi : X \longrightarrow Y$ is an* **expanding map** *if $\exists\ \lambda > 1$ such that $\lambda d_X(x,y) \le d_Y(\varphi(x),\varphi(y)) \quad \forall\ x,y \in X$.*

Recall the definition of virtually isomorphic groups from the end of Section 3.1. In particular, a group homomorphism $\varphi : \Gamma_1 \longrightarrow \Gamma_2$ is an **expanding virtual isomorphism** if it is expanding (hence injective), and $[\Gamma_2 : \varphi(\Gamma_1)] < \infty$.

Our reason for considering expanding virtual isomorphisms is the following result. We will come back to its proof after looking at some examples.

{l.franks}

**Lemma 4.15** (Franks' lemma 1970)**.** *If a finitely generated group has an expanding virtual automorphism, then it has polynomial growth.*

▷ **Exercise 4.12.**

**(a)** *Given two different generating sets $S_1$ and $S_2$ for $\Gamma$, prove that if $\varphi$ is expanding in $G(\Gamma, S_1)$ then $\varphi^k$ is expanding in $G(\Gamma, S_2)$ for some $k$.*

**(b)** *Give an injective virtual automorphism of $\mathbb{Z}^2$ that is not expanding but whose square is.*

The standard examples of expanding virtual automorphisms are the following:

1. In $\mathbb{Z}^d$, the map $x \mapsto k \cdot x$ is an expanding virtual isomorphism, since $[\mathbb{Z}^d : k\mathbb{Z}^d] = k^d$.

2. For the Heisenberg group $H_3(\mathbb{Z})$, the map

$$\begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \overset{\varphi_{m,n}}{\longmapsto} \begin{pmatrix} 1 & mx & mnz \\ 0 & 1 & ny \\ 0 & 0 & 1 \end{pmatrix}$$

is an expanding virtual automorphism, with index $[H_3(\mathbb{Z}) : \varphi_{m,n}(H_3(\mathbb{Z}))] = m^2n^2$.

▷ **Exercise 4.13.** $^{**}$ *If $\mathbb{Z}^d \rtimes_M \mathbb{Z}$ is nilpotent, does it have an expanding virtual automorphism?*

There exist nilpotent groups with no isomorphic subgroup of finite index greater than one [Bel03]. Groups with this property are called co-Hopfian. (And groups having no isomorphic factors with non-trivial kernel are called Hopfian.)

Instead of proving Franks' Lemma 4.15, let us explore a geometric analogue:

{l.franks2}

**Lemma 4.16.** *Let $M$ be a Riemannian manifold. Assume $v(r) := \sup_{x \in M} \mathrm{vol}(B_r(x)) < \infty$. If $\varphi : M \to M$ is an expanding homeomorphism with Jacobian $Det(D\varphi) < K$, then $M$ has polynomial volume growth.*

*Proof.* From Definition 4.14 of expanding, we have $\varphi(B_r(x)) \supseteq B_{\lambda r}(\varphi(x))$, which gives

$$\mathrm{vol}(\varphi(B_r(x))) \geq \mathrm{vol}(B_{\lambda r}(\varphi(x))).$$

By the bound on the Jacobian of $\varphi$, we have that $Kv(r) \geq \mathrm{vol}(\varphi(B_r(x)) \geq \mathrm{vol}(B_{\lambda r}(\varphi(x)))$, and taking the supremum over $x$ on both sides gives $Kv(r) \geq v(\lambda r)$. This implies polynomial growth by the following: For a given $r$, set $j = \log_\lambda r$. Then

$$v(r) = v(\lambda^j) \leq v(\lambda^{\lceil j \rceil}) \leq Kv(\lambda^{\lceil j \rceil - 1}) \leq \cdots \leq K^{\lceil j \rceil} v(1).$$

Since $K^{\lceil j \rceil} \leq K^j \max(1, K)$ and $K^j = K^{\log_\lambda r} = r^{\log_\lambda K}$, we finally have $v(r) \leq Cr^d$, where $d = \log_\lambda K$ and $C = v(1) \max(1, K)$. □

▷ **Exercise 4.14.** *Prove the group version of Franks' Lemma. (Hint: Emulate the ideas of the proof of the geometric version, but in a discrete setting, where the bounded Jacobian is analogous to the finite index of the subgroup.)*

▷ **Exercise 4.15.** $^{***}$ *Assume $\Gamma$ is a finitely generated group and has a virtual isomorphism $\varphi$ such that*

$$\bigcap_{n \geq 1} \varphi^n(\Gamma) = \{1\}.$$

*(This is the case, e.g., when $\varphi$ is expanding.) Does this imply that $\Gamma$ has polynomial growth?*

A condition weaker than in the last exercise is the following: a group $\Gamma$ is called **scale-invariant** if it has a chain of subgroups $\Gamma = \Gamma_0 \geq \Gamma_1 \geq \Gamma_2 \geq \cdots$ such that each $\Gamma_n$ is isomorphic to $\Gamma$, $[\Gamma : \Gamma_n] < \infty$, and $\bigcap \Gamma_n = \{1\}$. This notion was introduced by Itai Benjamini, and he had conjectured that it implies polynomial growth of $\Gamma$. However, this was disproved in [NekP09], by the following examples. The proofs use the self-similar actions of these groups on rooted trees, see Section 15.1.

{t.scaleinv}

**Theorem 4.17.** *The following groups of exponential growth are scale-invariant:*

- *the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}$, described below in Section 5.1;*
- *the affine group $\mathbb{Z}^d \rtimes \mathsf{GL}(d, \mathbb{Z})$;*
- *the solvable Baumslag-Solitar group $\mathsf{BS}(1, m)$, defined in Section 4.3.*

On the other hand, torsion-free Gromov-hyperbolic groups are not scale-invariant, a result that can be a little bit motivated by the well-known fact that hyperbolic spaces do not have homotheties.

Benjamini's question was motivated by the "renormalization method" of percolation theory. In [NekP09], we formulated a more geometric version of this question, on "scale-invariant tilings" in transitive graphs, which is still relevant to percolation renormalization and is still open; see Question 12.33.

Going back to volume growth, here are some big theorems:

**Theorem 4.18** (Tits' alternative [Tit72])**.** *A linear group (subgroup of a matrix group) $\Gamma$ either has $F_2 \leq \Gamma$ or it is solvable. In particular, $\Gamma$ has either exponential growth or polynomial growth.*

**Theorem 4.19** (Grigorchuk [Gri83])**.** *There exist finitely generated groups with intermediate growth.*

**Conjecture 4.20.** *There are no groups with superpolynomial growth but with growth of order $\exp(o(\sqrt{n}))$.*

For an introduction to groups of intermediate growth, see [GriP08], and for more details, see [dlHar00, BartGN03]. The proof of Grigorchuk's group being of intermediate growth relies on the following observation:

{l.highfranks}

**Lemma 4.21** (Higher order Franks' lemma)**.** *If $\Gamma$ is a group with growth function $v_\Gamma(n)$ and there exists an expanding virtual isomorphism*

$$\underbrace{\Gamma \times \Gamma \times \cdots \times \Gamma}_{m \geq 2} \longrightarrow \Gamma,$$

*then $\exp(n^{\alpha_1}) \preceq v_\Gamma(n) \preceq \exp(n^{\alpha_2})$ for some $0 < \alpha_1 \leq \alpha_2 < 1$.*

$\triangleright$ **Exercise 4.16.** *Prove the higher order Franks' lemma. (Hint: $\Gamma^m \hookrightarrow \Gamma$ implies the existence of $\alpha_1$, since $v(n)^m \leq C\, v(kn)$ for all $n$ implies that $v(n)$ has some stretched exponential growth. The expanding virtual isomorphism gives the existence of $\alpha_2$.)*

# 5 Isoperimetric inequalities

{s.isop}

## 5.1 Basic definitions and examples

{ss.isopbasic}

We start with a coarse geometric definition that is even more important than volume growth.

{d.isoperimetr

**Definition 5.1.** *Let $\psi$ be an increasing, positive function and let $G$ be a graph of bounded degree. Then we say that $G$ satisfies the* $\psi$**-isoperimetric inequality** $IP_\psi$ *if $\exists\ \kappa > 0$ such that $|\partial_E S| \geq \kappa\, \psi(|S|)$ for any finite subset of vertices $S \subseteq V(G)$, where the boundary $\partial_E S$ is defined to be the set of edges which are adjacent to a vertex in $S$ and a vertex outside of $S$. The supremum of all $\kappa$'s with this property is usually called the $\psi$-isoperimetric constant, denoted by $\iota_{\psi,E}$.*

Besides the edge boundary $\partial_E$ defined above, we can also consider the outer vertex boundary $\partial_V^{\mathrm{out}} S$, the set of vertices outside of $S$ with at least one neighbour in $S$, and the inner vertex boundary $\partial_V^{\mathrm{in}} S$, the set of vertices inside $S$ with at least one neighbour outside $S$. Since $G$ has bounded degrees, any of these could have been used in the definition. If the degrees are unbounded, or we are interested in the optimal constant factor $\kappa$, then we need to distinguish between edge- and vertex-isoperimetric inequalities, and in the edge-version it is natural to replace the size of $S$ by $\sum_{x\in S} \deg(x)$.

▷ **Exercise 5.1.** *Show that the satisfaction of an isoperimetric inequality is a quasi-isometry invariant among bounded degree graphs, but not among graphs with unbounded degrees.*

$\mathbb{Z}^d$ satisfies $IP_{n^{1-1/d}}$, often denoted $IP_d$. (This is so well-known that one might forget that it needs a proof. See Sections 5.3 and 5.4.) If a group satisfies $IP_\infty$, i.e., a linear isoperimetric inequality, then it is called nonamenable, as described in the next definition. The isoperimetric constant $\iota_{\infty,E}$ in this case is usually called the Cheeger constant.

{d.amenable}

**Definition 5.2.** *A bounded degree graph $G$ is* **amenable** *if there exists a sequence $\{S_n\}$ of connected subsets of vertices, $S_n \subseteq V(G)$, such that*

$$\frac{|\partial S_n|}{|S_n|} \to 0.$$

*Such an $\{S_n\}$ is called a* **Følner sequence***. We say that a group is Følner amenable if any of its finitely generated Cayley graphs is. A finite set $S$ with $|\partial S|/|S| < \delta |S|$ will sometimes be called a $\delta$-Følner set.*

If, in addition to connectedness, the $S_n$'s also satisfy $S_n \nearrow V(G)$, i.e., $S_n \subseteq S_{n+1}\ \forall\, n$ and $\bigcup_n S_n = V(G)$, then the sequence $\{S_n\}$ is called a **Følner exhaustion**. Not every amenable graph has a Følner exhaustion, as the example after the next exercise shows.

▷ **Exercise 5.2.**

**(a)** *Find the edge Cheeger constant $\iota_{\infty,E}$ of the infinite binary tree.*

**(b)** *Show that a bounded degree tree without leaves is amenable iff there is no bound on the length of "hanging chains", i.e., chains of vertices with degree 2. (Consequently, for trees, $IP_{1+\epsilon}$ implies $IP_\infty$.)*

Consider now the following tree. Take a bi-infinite path $\mathbb{Z}$, and, for each even integer $2k$, root an infinite binary tree at every vertex of $\mathbb{Z}$ whose distance from the origin is between $2^{2k}$ and $2^{2k+1}$. This tree has unbounded hanging chains and is thus amenable; however, it is clear that you cannot both have $S_n$ connected and $S_n \subseteq S_{n+1}$ in a Følner sequence, because then the $S_n$'s would contain larger and larger portions of binary trees, causing the $S_n$'s to have a large boundary.

**Lemma 5.3.** *In Definition 5.2, you can achieve $S_n \nearrow V(G)$ in the case of amenable Cayley graphs.*

*Proof.* Given a Følner sequence $S_n$, set $S_n^r := \bigcup_{g \in B_r} gS_n$, where $B_r$ is the ball in $G$ of radius $r$ centered at the origin. Without loss of generality we can assume $e \in S_n$, since $\Gamma$ acts on $G$ by graph automorphisms, and thus choosing any $g_n \in S_n$ we can consider $g_n^{-1}S_n$. Now, since $e \in B_r$, $S_n \subseteq S_n^r$, hence $|S_n| \le |S_n^r|$. Also,

$$\left|\partial S_n^r\right| \le \sum_{g \in B_r} |\partial(gS_n)| \le |B(r)|\, |\partial S_n|.$$

Thus we have

$$\frac{|\partial S_n^r|}{|S_n^r|} \le |B_r| \frac{|\partial S_n|}{|S_n|}.$$

Now, for each $r \in \mathbb{N}$, choose $n_r$ such that $|\partial S_{n_r}|/|S_{n_r}| \le \frac{1}{r\,|B_r|}$, and set $S_r^* := S_{n_r}^r$. Then $\{S_r^*\}$ is a Følner sequence, and $e \in S_n$ implies that $B_r \subseteq S_r^*$. If we take now a rapidly growing sequence $\{r(i)\}_i$ such that $S_{r(i)}^* \subseteq B_{r(i+1)}$, then $\{S_{r(i)}^*\}_i \nearrow G$ is a Følner exhaustion. □

The archetypical examples for the difference between amenable and non-amenable graphs are the Euclidean versus hyperbolic lattices, e.g., tilings in the Euclidean versus hyperbolic plane. The notions "non-amenable", "hyperbolic", "negative curvature" are very much related to each other, but there are also important differences. Here is a down-to-earth exercise to practice these notions; it might not be obvious at first sight, but part (a) is a special case of part (b).

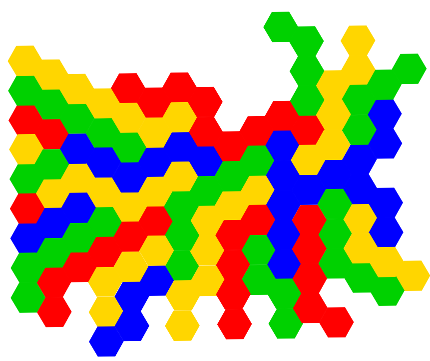


Figure 5.1: Trying to create at least 7 neighbours for each country. It works fine for a while, but then we seem to run out of space. {f.hexsept}

▷ **Exercise 5.3.**

**(a)** *Consider the standard hexagonal lattice. Show that if you are given a bound $B < \infty$, and can group the hexagons into countries, each being a connected set of at most $B$ hexagons, then it is not possible to have at least 7 neighbours for each country.*

**(b)** *In a locally finite planar graph $G$, define the* **combinatorial curvature** *at a vertex $x$ by*

$$\mathsf{curv}_G(x) := 2\pi - \sum_i \frac{(L_i - 2)\,\pi}{L_i}\,,$$

*where the sum runs over the faces adjacent to $x$, and $L_i$ is the number of sides of the $i^{th}$ face. Show that if there exists some $\delta > 0$ such that curvature is less than $-\delta\pi$ at each vertex, then it is not possible that both $G$ and its planar dual $G^*$ are edge-amenable.*

{ex.endsnoname

▷ **Exercise 5.4.** *Show that a group with a continuum number of ends (see Exercises 3.3 and 3.4) must be non-amenable.*

We now look at an important example of a Følner amenable group with exponential growth.

The **lamplighter group** is defined to be the wreath product $\mathbb{Z}_2 \wr \mathbb{Z}$, which is defined to be

$$\left(\bigoplus_{\mathbb{Z}} \mathbb{Z}_2\right) \rtimes_\sigma \mathbb{Z}\,,$$

where the left group consists of all bi-infinite binary sequences with only finitely many nonzero terms, and $\sigma$ is the left shift automorphism on this group. Thus a general element of the lamplighter group looks like $(f,m)$, where $f : \mathbb{Z} \longrightarrow \mathbb{Z}_2$ has $|\mathrm{supp}(f)| < \infty$, interpreted as a configuration of $\mathbb{Z}_2$-lamps on a $\mathbb{Z}$-street, and $m \in \mathbb{Z}$ is the lamplighter or marker. Such pairs multiply according to the semidirect product rules, see Section 4.2. The most transparent description is just to describe the Cayley graph w.r.t. some simple generating set:

Let $e_k \in \oplus_{\mathbb{Z}}\mathbb{Z}_2$ denote the function that has a 1 in the $k^{\text{th}}$ place and zeroes everwhere else. Then the group is generated by the following three elements:

$$\mathsf{s} := (e_0, 0); \quad \mathsf{R} := (0,1); \quad \text{and} \quad \mathsf{L} := (0,-1).$$

Multiplication by these generators gives $\mathsf{s}\cdot(f,m) = (e_0,0)\cdot(f,m) = (e_m+f,m)$, while $\mathsf{R}\cdot(f,m) = (f,m+1)$ and $\mathsf{L}\cdot(f,m) = (f,m-1)$, and so they can be interpreted as "switch", "Right" and "Left". (Unfortunately, because of the way we defined semidirect multiplication, we need to multiply from the left with these nice generators to get the above nice interpretation, even if we generally prefer taking right Cayley graphs in these notes.)

The volume of a ball in this generating set has the bound $|B_n(\mathrm{id})| \geq 2^{n/2}$, therefore the Cayley graph has exponential growth. This bound is clear because at each step you can either "switch" (apply $\mathsf{s}$), or not. On the other hand, it is not hard to see that this left Cayley graph is amenable: set

$$S_n = \{(f,m) : -n \leq m \leq n \text{ and } \mathrm{supp}(f) \subseteq [-n,n]\}$$

and observe that $|S_n| = 2^{2n+1}(2n+1)$ and $|\partial_V^{\mathrm{in}} S_n| = 2^{2n+1}\cdot 2$, since the points on the boundary correspond to $m=-n$ or $m=n$.

One may consider generalizations $\mathsf{F} \wr \Gamma = (\bigoplus_\Gamma \mathsf{F}) \rtimes \Gamma$, with $\Gamma$ being any base group as a "city", and $\mathsf{F}$ being any "lamp group". Moreover, one can even define the **lamplighter graph** $F \wr G$ built from any two graphs $F$ and $G$: the lamplighter can adjust the status of the current lamp according to $F$, or can move on the base graph $G$.

The lamplighter groups $\mathbb{Z}_p \wr \mathbb{Z}$ have some surprising Cayley graphs. The **Diestel-Leader graph** $\mathsf{DL}(k,\ell)$ is the so-called **horocyclic product** of $\mathbb{T}_{k+1}$ and $\mathbb{T}_{\ell+1}$: pick an end of each tree, organize all the vertices into layers labeled by $\mathbb{Z}$ with labels tending to $+\infty$ towards that end (with the location of the zero level being arbitrary), then let $V(G)$ consist of all the pairs $(v,w) \in \mathbb{T}_{k+1} \times \mathbb{T}_{\ell+1}$ with labels $(n,-n)$ for some $n \in \mathbb{Z}$, with an edge from $(v,w)$ to $(v',w')$ if $(v,v') \in E(\mathbb{T}_{k+1})$ and $(w,w') \in E(\mathbb{T}_{\ell+1})$. See Figure 5.2.

{ex.DLamen}

▷ **Exercise 5.5.** *Show that* $\mathsf{DL}(k,\ell)$ *is amenable iff* $k=\ell$.

{ex.LLDL}

▷ **Exercise 5.6.** *Show that the Cayley graph of the lamplighter group* $\Gamma = \mathbb{Z}_2 \wr \mathbb{Z}$ *with generating set* $S = \{\mathsf{R}, \mathsf{Rs}, \mathsf{L}, \mathsf{sL}\}$ *is the Diestel-Leader graph* $\mathsf{DL}(2,2)$. *How can we obtain* $\mathsf{DL}(p,p)$ *from* $\mathbb{Z}_p \wr \mathbb{Z}$*?*

## 5.2 Amenability, invariant means, wobbling paradoxical decompositions

{ss.amen}

The algebraic counterpart of amenability is the following.

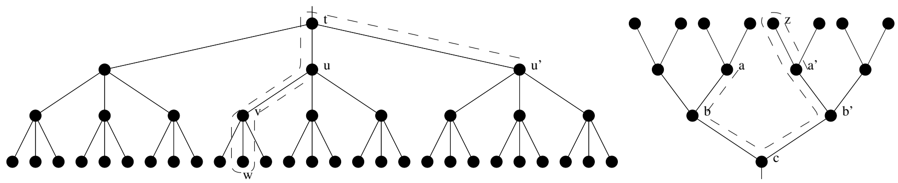



Figure 5.2: The Diestel-Leader graph $\mathsf{DL}(3,2)$. A sample path: $(u,a)$, $(v,b)$, $(w,c)$, $(v,b')$, $(u,a')$, $(t,z)$, $(u',a')$. {f.43}

{d.vNamenable}

**Definition 5.4** (von Neumann 1929)**.** *A finitely generated group* $\Gamma$ *is* **von Neumann amenable** *if it has an invariant mean on* $L^\infty(\Gamma)$*, i.e., there exists a linear map* $m : L^\infty(\Gamma) \longrightarrow \mathbb{R}$ *with the following two properties for every bounded* $f : \Gamma \longrightarrow \mathbb{R}$*:*

1. $m(f) \in [\inf(f), \sup(f)]$
2. *For all* $\gamma \in \Gamma$, $m(f_\gamma) = m(f)$*, where* $f_\gamma(x) := f(\gamma x)$*.*

{p.measurevN}

**Proposition 5.5.** *A finitely generated group* $\Gamma$ *is von Neumann amenable if and only if there exists a finitely additive invariant (with respect to translation by group elements) probability measure on all subsets of* $\Gamma$*.*

To prove the proposition, identify $\mu(A)$ and $m(\mathbf{1}_A)$ and approximate general bounded functions by step functions.

**Example:** $\mathbb{Z}$ is amenable, but the obvious candidate

$$\mu(A) = \limsup_{n\to\infty} \frac{|[-n,n] \cap A|}{2n+1}$$

is not good enough, as it fails finite additivity. But there is some sort of limit argument to find an appropriate measure, which requires the Axiom of Choice.

▷ **Exercise 5.7.**

**(a)** *Prove that subgroups of amenable groups are amenable.*

**(b)** *Given a short exact sequence* $1 \longrightarrow A_1 \longrightarrow \Gamma \longrightarrow A_2 \longrightarrow 1$*, show that if* $A_1$ *and* $A_2$ *are amenable, then* $\Gamma$ *is as well.*

From the amenability of $\mathbb{Z}$, this exercise gives that $\mathbb{Z}^d$ is amenable. Moreover, it can be proved that an infinite direct sum of amenable groups is also amenable, hence the lamplighter group $\Gamma$ is also amenable: it is two-step solvable, with $[\Gamma,\Gamma] = \oplus_{\mathbb{Z}}\mathbb{Z}_2$. More generally, a group is amenable if and only if all finitely generated subgroups of it are amenable, and hence the exercise implies that any solvable group is amenable.

{t.folner}

**Theorem 5.6** (Følner [Føl55])**.** *If* $\Gamma$ *is a finitely generated group, it is von Neumann amenable if and only if any of its Cayley graphs is Følner amenable.*

{t.kestenorig}

**Theorem 5.7** (Kesten [Kes59])**.** *A finitely generated group* $\Gamma$ *is amenable if and only if the spectral radius* $\rho$ *of any of its Cayley graphs, as defined in (1.8), equals 1.*

We will sketch a proof of Følner's theorem below, and prove Kesten's theorem in a future lecture, Section 7.2.

{pr.F2vN}

**Proposition 5.8.** $F_2$ *is nonamenable in the von Neumann sense.*

*Proof.* Denote $F_2$ as $\langle a, b\rangle$, and let $A^+$ denote the set of words in $F_2$ beginning with $a$. Let $A^-$ denote the set of words beginning with $a^{-1}$, and let $A = A^+ \cup A^-$. Define $B^+$, $B^-$, and $B$ similarly. Notice that $F_2 = A \cup B \cup \{e\}$, and that $F_2$ also equals $A^+ \sqcup aA^-$, as well as $B^+ \sqcup bB^-$.

Now, suppose that we have $\mu$ as in Proposition 5.5. Certainly $\mu(\{e\}) = 0$, and so we have

$$\begin{aligned}\mu(F_2) &= \mu(A) + \mu(B) + \mu(\{e\}) \\ &= \mu(A^+) + \mu(A^-) + \mu(B^+) + \mu(B^-) \\ &= \mu(A^+) + \mu(aA^-) + \mu(B^+) + \mu(bB^-) \\ &= \mu(A^+ \sqcup aA^-) + \mu(B^+ \sqcup bB^-) \\ &= 2\mu(F_2).\end{aligned}$$

This is a contradiction, and thus no such measure exists. □

▷ **Exercise 5.8.*** $\mathsf{SO}(3) \geq F_2$ *(Use the Ping Pong Lemma, Lemma 2.19).*

This exercise and Proposition 5.8 form the basis of the **Banach-Tarski paradox**: the 3-dimensional solid ball can be decomposed into finitely many pieces that can be rearranged (using rigid motions of $\mathbb{R}^3$) to give two copies of the original ball (same size!). See [Lub94] or [Wag93] for more on this.

The following theorem was proved by Olshanski in 1980, Adian in 1982, Gromov in 1987, and again by Olshanski and Sapir in 2002, this time for finitely presented groups.

{t.olshanski}

**Theorem 5.9** (Olshanski 1980)**.** *There exist nonamenable groups without* $F_2$ *as a subgroup.*

An example of this is the Burnside group $B(m, n) = \langle g_1, ..., g_m \,|\, g^n = 1 \forall g\rangle$ for $m \geq 2$ and $n \geq 665$ and odd.

Now we sketch the proof of the Følner theorem, but first we define some notions and state an exercise we will use. I learned about this approach, using wobbling paradoxicity, from Gábor Elek; see [ElTS05] and the references there, although I am not sure that this proof has appeared anywhere before.

{d.wobbling}

**Definition 5.10.** *Let $X$ be a metric space. As in Exercise 3.1, a map $\varphi : X \longrightarrow X$ is at a bounded distance from the identity, or* **wobbling**, *if* $\sup_x d(x, \varphi(x)) < \infty$. *Further, the maps $\alpha$ and $\beta$ form a* **wobbling paradoxical decomposition** *of $X$ if they are wobbling injections such that $\alpha(X) \sqcup \beta(X) = X$.*

{ex.wobbling}

▷ **Exercise 5.9.*** *A bounded degree graph is nonamenable if and only if it has a wobbling paradoxical decomposition. (Hint: State, prove and use the locally finite infinite bipartite graph version of the Hall marriage theorem [Die00, Theorem 2.1.2], called the Hall-Rado theorem.)*

*Sketch of proof of Theorem 5.6.* For the reverse direction, if there exists a Følner sequence $\{S_n\}$, define $\mu_n(A) := \frac{|A \cap S_n|}{|S_n|}$, and show that some sort of limit exists. Now, because $\{S_n\}$ is a Følner sequence, $|S_n g^{-1} \triangle S_n| < \epsilon |S_n|$ for $g$ a generator. So $\mu_n(Ag) = \frac{|Ag \cap S_n|}{|S_n|} = \frac{|(A \cap S_n g^{-1})g|}{|S_n|}$ will have the same limit as $\mu_n(A)$, giving invariance of $\mu$.

For the forward direction, we prove the contrapositive, so assume that $G$ is nonamenable. Then by Exercise 5.9, it has a paradoxical decomposition $\alpha$ and $\beta$. Suppose that both of these maps move a vertex a distance of at most $r$. Then we can decompose

$$V = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_k = B_1 \sqcup B_2 \sqcup \cdots \sqcup B_\ell,$$

where $\alpha|_{A_i}$ is translation by some $g_i \in B_r$, $\beta|_{B_i}$ is a translation by some $h_i \in B_r$, and $k, \ell \leq |B_r|$. If we let $C_{i,j} := A_i \cap B_j$, and if we assume for contradiction that there exists some invariant probability measure $\mu$ on $\Gamma$ as in Definition 5.5, then since $\mu(\alpha(C_{i,j})) = \mu(C_{i,j}) = \mu(\beta(C_{i,j}))$, we have

$$\mu(V) = \mu(\alpha(V) \sqcup \beta(V)) = \sum_{i,j} \mu(\alpha(C_{i,j})) + \mu(\beta(C_{i,j})) = 2\mu(V)\,.$$

Hence $\Gamma$ is von Neumann nonamenable. □

## 5.3 From growth to isoperimetry in groups

{ss.growthisop

The following theorem was proved by Gromov for groups, and generalized by Coulhon and Saloff-Coste in 1993 for any transitive graph.

{t.CSC}

**Theorem 5.11** ([CouSC93])**.** *Let $\Gamma$ be a finitely generated group, with a right Cayley graph $G(\Gamma, S)$. Define the inverse growth rate by*

$$\rho(n) := \min\{r : |B_r(o)| \geq n\}\,.$$

*Let $K$ be any finite subset of $\Gamma$. Recall the definition $\partial_V^{\mathrm{in}} K = \{v \in K : \exists \gamma \in S\ v\gamma \notin K\}$. Then we have*

$$\frac{\left|\partial_V^{\mathrm{in}} K\right|}{|K|} \geq \frac{1}{2\rho(2|K|)}.$$

*Proof.* Take any $s \in S$. It is clear that $x \in K \setminus Ks^{-1}$, i.e., $x \mapsto xs$ moves $x$ out of $K$, only if $x \in \partial_V^{\mathrm{in}} K$. Thus $|K \setminus Ks^{-1}| \leq \left|\partial_V^{\mathrm{in}} K\right|$. More generally, if $g = s_1 \cdots s_r \in \Gamma$ is a product of $r$ generators, then, by iterating the above argument, we see that $|K \setminus Kg^{-1}| \leq r\left|\partial_V^{\mathrm{in}} K\right|$. In more detail: writing $K \setminus Kg^{-1} = \bigcup_{j=1}^r H_j$, where $H_j := \{x \in K : xs_1 \cdots s_i \in K \text{ for } i = 1, \ldots, j-1, \text{ but } xs_1 \cdots s_j \notin K\}$, the map $x \mapsto xs_1 \cdots s_{j-1}$ from $H_j$ to $\partial_V^{\mathrm{in}} K$ is injective.

On the other hand, let $\rho = \rho(2|K|)$, and observe that for any $x \in K$,

$$\left|\left\{xg : g \in B_\rho(o)\right\} \setminus K\right| \ \geq\ |B_\rho(o)|/2\,,$$

since the size of $\{xg : g \in B_\rho(o)\}$ is at least $2|K|$. Therefore, if we pick $g \in B_\rho(o)$ uniformly at random, then

$$\mathbf{P}\big[\, g \text{ moves } x \text{ out of } K \,\big] \ \geq\ 1/2,$$

which implies $\mathbf{E}\big[\text{number of } x\text{'s moved out of } K\big] \geq |K|/2$. Hence, there is a $g$ that moves at least $|K|/2$ elements out of $K$. Combining our upper and lower bounds on the number of elements of $K$ moved out of $K$ by $g$, we have

$$\rho\left|\partial_V^{\mathrm{in}} K\right| \geq \|g\|_S \left|\partial_V^{\mathrm{in}} K\right| \geq |K|/2\,,$$

and we are done. □

**Examples:** For $\Gamma = \mathbb{Z}^d$, $\rho(n) = n^{1/d}$, hence we get that it satisfies $IP_d$. (We will see another proof strategy in the next section.) It obviously does not satisfy anything stronger, hence $\mathbb{Z}^d$ shows that the above inequality is sharp (up to constant factors), at least in the regime of polynomial growth. The lamplighter group shows that the inequality is also sharp for groups of exponential growth.

{ex.Timar}

▷ **Exercise 5.10** (Ádám Timár)**.** *** *For any f.g. group* $\Gamma$*, does* $\exists ? C_1$ *s.t.* $\forall A \subset \Gamma$ *finite* $\exists g \in \Gamma$ *with*

$$0 < d(A, gA) \leq C_1 \,,$$

*and does* $\exists ? C_2$ *s.t.* $\forall A, B \subset \Gamma$ *finite,* $\exists g \in \Gamma$ *with*

$$0 < d(A, gB) \leq C_2 ?$$

▷ **Exercise 5.11** (Iva Kozáková-Špakulová)**.** *Give an example of a group* $\Gamma$ *where* $C_2 > 1$ *is needed.*

One reason for these exercises to appear here is the important role translations played also in Theorem 5.11. A simple application of an affirmative answer to Exercise 5.10 would be that the boundary-to-volume ratio would get worse for larger sets, since we could glue translated copies of any small set to get larger sets with worse boundary-to-volume ratio. In other words, the isoperimetric profile $\phi(r) := \{|\partial S|/|S| \,:\, |S| \leq r\}$ would be roughly decreasing. An actual result using the same idea is [LyPer16, Theorem 6.2]: for any finite set, the boundary-to-volume ratio is strictly larger than the Cheeger constant.

Related isoperimetric inequalities were proved in [BabSz92] and [Żuk00].

## 5.4 Isoperimetry in $\mathbb{Z}^d$ and $\{0,1\}^n$

{ss.Zdisop}

On $\mathbb{Z}^d$, the following sharp result is known:

{t.integerIP}

**Theorem 5.12.** *For any* $S$ *in* $\mathbb{Z}^d$*,* $|\partial_E S| \geq 2d|S|^{1-\frac{1}{d}}$*, where* $\partial_E S$ *is the set of edges with one vertex in* $S$ *and one in* $S^c$*.*

One proof of Theorem 5.12, using some very natural compression methods, was done by Bollobás and Leader [BolL91]. The basic idea is that one can apply "gravity" (acting in any coordinate direction) to any subset $S$ of vertices, which compresses $S$, reducing its boundary without changing its volume. Subsets with minimal boundary are more-or-less stable under this compression, and one can describe such stable sets quite well. A beautiful alternative method can be seen in [LyPer16, Section 6.7]: it goes through the following theorem, which is proved using conditional entropy inequalities. See [BaliBo12] for a concise treatment and a mixture of both the entropy and compression methods, with applications to combinatorial number theory.

{t.LoomisWhitn

**Theorem 5.13** (Discrete Loomis-Whitney Inequality)**.**

$$\forall S \subseteq \mathbb{Z}^d, \quad |S|^{d-1} \leq \prod_{i=1}^{d} |P_i(S)|,$$

*where* $P_i(S)$ *is the projection of* $S$ *in the direction of the* $i^{th}$ *coordinate (that is, we simply delete the* $i$*th coordinate).*

The Loomis-Whitney inequality gives:

$$|S|^{\frac{d-1}{d}} \;\leq\; \left(\prod_{i=1}^{d} |P_i(S)|\right)^{\frac{1}{d}} \;\leq\; \frac{1}{d}\sum_{i=1}^{d} |P_i(S)| \;\leq\; \frac{|\partial_E S|}{2d},$$

where the last inequality holds because each line in the $i$th coordinate direction must leave the set $S$ through at least two edges. And this is exactly Theorem 5.12.

{ex.HyperIsopE

▷ **Exercise 5.12** ([Hart76]). *The classical edge-isoperimetric inequality for the hypercube says that for sets of size $|S| = 2^k$, $0 \le k \le n$, the edge-boundary is minimized by the $k$-dimensional subcubes, hence $|\partial_E S| \ge 2^k(n-k)$. Also more generally, $|\partial_E S| \ge |S| \log_2 \big(2^n/|S|\big)$. Prove this.*

Let me give my personal endorsement for the conditional entropy proof of Theorem 5.13 in [LyPer16, Section 6.7]. In [Pet08], I needed to prove an isoperimetric inequality in the wedge $\mathcal{W}_h \subset \mathbb{Z}^3$, the subgraph of the lattice induced by the vertices $V(\mathcal{W}_h) = \{(x, y, z) : x \ge 0 \text{ and } |z| \le h(x)\}$, where $h(x)$ is some increasing function. Namely, using the flow criterion of transience, Theorem 6.9 below, it was shown in [Lyo83] that $\mathcal{W}_h$ is transient iff

$$\sum_{j=1}^{\infty} \frac{1}{j h(j)} < \infty \,. \tag{5.1}$$ {e.tlyons}

For example, $h(j) = \log^r j$ gives transience iff $r > 1$. I wanted to show that this implies Thomassen's condition for transience [Ths92], which is basically an isoperimetric inequality $IP_\psi$ with

$$\sum_{k=1}^{\infty} \psi(k)^{-2} < \infty \,. \tag{5.2}$$ {e.thomassen}

(The reason for this goal will be clear in Section 12.5.) In such a subgraph $\mathcal{W}_h$, the Bollobás-Leader compression methods seem completely useless, but I managed to prove the result using conditional entropy inequalities for projections. What I proved was that $\mathcal{W}_h$ satisfies $IP_\psi$ with

$$\psi(v) := \sqrt{v h\Big(\sqrt{v/h(\sqrt{v})}\Big)}.$$

As can be guessed from its peculiar form, this is not likely to be sharp, but is good enough to deduce (5.2) from (5.1). E.g., for $h(v) = v^\alpha$, we get $\psi(v) = v^{\frac{1}{2}+\frac{\alpha}{4}-\frac{\alpha^2}{8}}$, which is not close to the easily conjectured isoperimetric function $v^{\frac{1+\alpha}{2+\alpha}}$, except for $\alpha$ close to 0, but that is exactly the interesting regime here, hence this strange $\psi(v)$ suffices.

▷ **Exercise 5.13.**

**(a)** *Show that the wedge $\mathcal{W}_h$ with $h(v) = v^\alpha$, $\alpha > 0$, does not satisfy $IP_\psi$ whenever $\psi(v)/v^{\frac{1+\alpha}{2+\alpha}} \to \infty$.*

**(b)** *** *Is it true that $\mathcal{W}_h$ with $h(v) = v^\alpha$ does satisfy $IP_\psi$ with $\psi(v) = v^{\frac{1+\alpha}{2+\alpha}}$ ?*

# 6 Random walks, discrete potential theory, martingales

{s.potential}

Probability theory began with the study of sums of i.i.d. random variables: LLN, CLT, large deviations. One can look at this as the theory of 1-dimensional random walks. One obvious generalization is to consider random walks on graphs with more interesting geometries, or on arbitrary graphs in general. The first two sections here will introduce a basic and very useful technic for this: electric networks and discrete potential theory. Another obvious direction of generalization is to study stochastic processes that still take values in $\mathbb{R}$, resemble random walks in some sense, but whose increments are not i.i.d. any more: these will be the so-called martingales, the subject of the third section. Discrete harmonic functions will connect martingales to the first two sections.

## 6.1 Markov chains, electric networks and the discrete Laplacian

{ss.networks}

Simple random walks on groups (for which we saw examples in Section 1.1), are special cases of reversible Markov chains, which we now define.

A **Markov chain** is a sequence of random variables $X_1, X_2, X_3, \ldots \in V$ such that

$$\mathbf{P}\big[\, X_{n+1} = y \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,\big] = \mathbf{P}\big[\, X_{n+1} = y \mid X_n = x_n \,\big] = p(x_n, y).$$

That is, the behaviour of the future states is governed only by the current state. The values $p(x, y)$ are called the **transition probabilities**, where it is usually assumed that $\sum_{y \in V} p(x, y) = 1$ for all $x \in V$. We will also use the notation $p_n(x, y) = \mathbf{P}\big[\, X_n = y \mid X_0 = x \,\big]$, just as in Section 1.1.

Given any measure $\pi$ (finite or infinite) on the state-space $V$, the Markov chain tells us how it evolves in time: after one step, we get the measure $\pi P(y) = \sum_{x \in V} \pi(x) p(x, y)$. A measure $\pi$ is called **stationary** if the chain leaves it invariant: $\pi P = \pi$. An important basic theorem is that any finite Markov chain has a stationary distribution, which is unique if the chain is irreducible (which means that for any $x, y \in V$ there is some $n$ with $p_n(x, y) > 0$). These facts follow from the Perron-Frobenius theorem for the matrix of transition probabilities, which we will not state here. There is also a probabilistic proof, constructing a stationary measure using the expected number of returns to any given vertex. See, for instance, [Dur10, Section 6.5]. For an infinite state space, there could be no stationary measure, or several. We will not discuss these matters in this generality, but see the example below.

We say that a Markov chain is **reversible** if there exists a reversible measure, i.e., non-negative function $\pi(x)$, not identically zero, that satisfies

$$\pi(x) p(x, y) = \pi(y) p(y, x) \quad \forall x, y.$$

Note that this reversible measure is unique up to a global factor, since $\pi(x)/\pi(y)$ is given by the Markov chain. Note also that reversible measures are also stationary. But not every stationary measure is reversible: it is good to keep in mind the following simple examples.

**Example.** Consider the $n$-cycle $\mathbb{Z} \pmod{n}$ with transition probabilities $p(i, i+1) = 1 - p(i+1, i) = p > 1/2$ for all $i$. It is easy to see that this is a non-reversible chain; intuitively, a movie of the evolving chain looks different from the reversed movie: it moves more in the $+$ direction. But the uniform distribution is a stationary measure.

On the other hand, the chain with the same formula for the transition probabilities on $\mathbb{Z}$ is already reversible. Although the uniform measure $\pi(i) = 1 \; \forall\, i \in \mathbb{Z}$ is still stationary but non-reversible, the measure $\pi'(i) = \big(p/(1-p)\big)^i$ is reversible. The above intuition with reversing the movie goes wrong now because $\pi'$ is not a finite measure, hence looking at a typical realization of the movie is simply meaningless. A simple real-life example of an infinite chain having some unexpected stationary measures is the following joking complaint of my former advisor Yuval Peres: "Each day is of average busyness: busier than yesterday but less busy than tomorrow."

▷ **Exercise 6.1.** *Show that a Markov chain $(V, P)$ has a reversible measure if and only if for all oriented cycles $x_0, x_1, \ldots, x_n = x_0$, we have $\prod_{i=0}^{n-1} p(x_i, x_{i+1}) = \prod_{i=0}^{n-1} p(x_{i+1}, x_i)$.*

Now, all Markov chains on a countable state space are in fact random walks on weighted graphs:

{d.network}

**Definition 6.1.** *Consider a directed graph with weights on the directed edges: for any two vertices $x$ and $y$, let $c(x, y)$ be any non-negative number such that $c(x, y) > 0$ only if $x$ and $y$ are neighbours.*

*Let $C_x = \sum_y c(x,y)$, and define the* **weighted random walk** *to be the Markov chain with transition probabilities*

$$p(x,y) = \frac{c(x,y)}{C_x}.$$

*An important special case is when $c(x,y) = c(y,x)$ for every $x$ and $y$, i.e., we have weights on the undirected edges. Such weighted graphs are usually called* **electric networks***, and the edge weights $c(e)$ are called* **conductances***. The inverses $r(e) = 1/c(e) \in (0,\infty]$ are called* **resistances***. The associated weighted random walk is sometimes called the network walk.*

The walk associated to an electric network is always reversible: $C_x$ is a reversible measure. On the other hand, any reversible Markov chain on a countable state space comes from an electric network, since we can define $c(x,y) := \pi(x)p(x,y) = \pi(y)p(y,x) = c(y,x)$.

▷ **Exercise 6.2.** *Show by examples that, in directed weighted graphs, the measure $(C_x)_{x\in V}$ might be non-stationary, and might be stationary but non-reversible. Can the walk associated to a finite directed weighted graph have a reversible measure?*

An electric network is like a discrete geometric space: we clearly have some notion of closeness coming from the neighbouring relation, where a large resistance should mean a larger distance. Indeed, we will define discrete versions of some of the notions of multidimensional calculus (or more generally, differential and Riemannian geometry), like gradient, divergence, and harmonicity, which will turn out to be very relevant for studying the random walk associated to the network.

Let us start with the following problem. We are given a network $G(V,E,c)$, with two subsets $A,Z \subset V$, we start the random walk at some $x \in V$, and want to compute

$$f(x) := \mathbf{P}_x[\,\tau_A < \tau_Z\,], \tag{6.1}$$ {e.hitP}

where $\tau_A$ is the first time we hit $A$. In order for (6.1) to be well-defined, we assume that $\tau_A \wedge \tau_Z$ is almost surely finite. A simple example to keep in mind is SRW on the interval $V = \{0,1,\dots,n\}$, with $A = \{0\}$ and $Z = \{n\}$. We can solve this problem by writing a recursion, or a discrete differential equation, or a dynamic program, call it according to your taste. Namely, for any inner vertex $x \notin A \cup Z$, condition on what the first step $X_1$ of the walk could be:

$$f(x) = \sum_{y\in V} \mathbf{P}_x[\,X_1 = y\,]\,\mathbf{P}_x[\,\tau_A < \tau_Z \mid X_1 = y\,] = \sum_{y\in V} \mathbf{P}_x[\,X_1 = y\,]\,\mathbf{P}_y[\,\tau_A < \tau_Z\,] = \sum_{y\in V} p(x,y)f(y)\,,$$

or more concisely,

$$f(x) = \mathbf{E}_x\big[\,\mathbf{P}_x[\,\tau_A < \tau_Z \mid X_1\,]\,\big] = \mathbf{E}_x\big[\,\mathbf{P}_{X_1}[\,\tau_A < \tau_Z\,]\,\big] = \mathbf{E}_x[\,f(X_1)\,]\,.$$

The boundary values are of course $f|_A = 1$ and $f|_Z = 0$. In the example of the interval, $f(x) = (f(x-1) + f(x+1))/2$ for all $1 \le x \le n-1$, and one can easily see that the unique solution is the linear extension of the boundary values: $f(x) = (n-x)/n$.

More generally, let $W$ be any subset of $V$, and let $f : W \longrightarrow \mathbb{R}$ be any real-valued function. Then, for $x \in V \setminus W$, define

$$f(x) := \mathbf{E}\big[\,f(X_\tau) \mid X_0 = x\,\big]\,, \tag{6.2}$$ {e.hitE}

where $\tau$ is the first time we visit $W$, and we assume that $\tau < \infty$ almost surely. We get back the previous example by $W = A \cup Z$, $f|_A = 0$, $f|_Z = 1$. And, by the exact same argument as above, we have $f(x) = \mathbf{E}_x[\,f(X_1)\,]$ again. It is now time to make the following fundamental definitions:

{d.markovop}

**Definition 6.2.** *The* **Markov operator** $P$ *of a Markov chain on* $V$ *is defined by*

$$(Pf)(x) = \sum_{y \in V} p(x,y) f(y) = \mathbf{E}_x[\, f(X_1)\,]\,,$$

*acting on functions* $f : V \longrightarrow \mathbb{R}$*: taking the one-step average by the Markov chain. The* **Laplacian** *operator is* $\Delta := I - P$*. A function* $f : V \longrightarrow \mathbb{R}$ *is* **harmonic** *at some* $x \in V$ *if* $\Delta f(x) = 0$*; that is, if the* **mean value property** *holds:*

$$\mathbf{E}_x[\, f(X_1)\,] = f(x)\,,$$

*the average of the function values after one step of the chain equals the value at the starting point* $x$*. A function is harmonic if it is harmonic at every vertex.*

Examples (6.1) and (6.2) are harmonic in $V \setminus (A \cup Z)$ and $V \setminus W$, respectively. One justification for calling this property harmonicity is the mean value property, which is a direct analogue of the continuous property

$$f(x) = \frac{1}{\mathrm{Vol}_{d-1}(\partial B_r(x))} \int_{\partial B_r(x)} f(y)\,dy = \frac{1}{\mathrm{Vol}_d(B_r(x))} \int_{B_r(x)} f(y)\,dy\,.$$

And, just like in the case of continuous harmonic functions, an immediate consequence of the mean value property is the **maximum principle**: if $f : V \longrightarrow \mathbb{R}$ is harmonic for all $x \in V \setminus W$, then there are no strict local maxima in $V \setminus W$, and if there is a global maximum over $V$, then it must be achieved also at some point of $W$.

{l.harmext}

**Lemma 6.3.** *Given a finite network* $G(V,E,c)$*, a subset* $W \subset V$*, and any real-valued function* $f$ *on* $W$*, there exists a unique extension of* $f$ *to* $V$ *that is harmonic on* $V \setminus W$*.*

*Proof.* Existence: use the extension (6.2) above.

Uniqueness: Suppose that $f_1$ and $f_2$ are two extensions of $f$. Let $g = f_1 - f_2$, again harmonic on $V \setminus W$. Since $V$ is finite, the global maximum and minimum of $g$ is attained, and by the maximum principle, it must also be attained on $W$, where $g \equiv 0$. Hence $g \equiv 0$ on $V$. $\square$

In the continuous world, the Laplacian is the second order differential operator $\sum_{i=1}^d \partial_{x_i x_i}$. Then we should also have first order operators in the discrete world! Here they are.

Consider an electric network on the graph $G = (V,E)$. If $e$ is an edge with vertices $e^+$ and $e^-$, decompose it into two directed edges $e$ and $\widehat{e}$ such that $e$ runs from $e^-$ to $e^+$, while $\widehat{e}$ runs from $e^+$ to $e^-$. Denote by $\overleftrightarrow{E}$ the set of directed edges formed.

Take $f, g : V \longrightarrow \mathbb{R}$. Define

$$\nabla f(e) = [f(e^+) - f(e^-)]\, c(e)$$

to be the **gradient** of $f$. The weight $c(e) = 1/r(e)$ is natural since the resistance $r(e)$ of a physical link is proportional to the length of $e$. In a cohomological language, this is a **coboundary operator**, since, from functions on zero-dimensional objects (the vertices), it produces functions on one-dimensional objects (the edges). Also define the inner product

$$(f,g) = (f,g)_C = \sum_{x \in V} f(x)\, g(x)\, C_x.$$

This is again natural from the point of view that imposing a large potential at a vertex that has larger total conductance has a larger effect on the network. Also mathematically, it is natural to consider the

Hilbert space $L^2(V,\pi)$ based on any stationary measure $\pi$, and $(C_x)_{x\in V}$ is a reversible stationary measure for the network walk; see, e.g., Exercise 6.3 below.

Take $\theta,\eta : \overleftrightarrow{E} \longrightarrow \mathbb{R}$ such that $\theta(\widehat{e}) = -\theta(e)$. Define the **boundary operator**

$$\nabla^*\theta(x) = \sum_{e:e^+=x} \theta(e)\frac{1}{C_x}.$$

Also define another inner product,

$$(\theta,\eta) = (\theta,\eta)_r = \frac{1}{2}\sum_{e\in\overleftrightarrow{E}} \theta(e)\,\eta(e)\,r(e).$$

Some works, e.g. [LyPer16], omit the "vertex conductances" $C_x$ from the inner product on $V$ and the boundary operator $\nabla^*$, while our definition agrees with [Woe00]. This is an inessential difference, but one needs to watch out for it.

{p.adjoint}

**Proposition 6.4.** $(\nabla f,\theta)_r = (f,\nabla^*\theta)_C$, *i.e.,* $\nabla$ *and* $\nabla^*$ *are the adjoints of each other (hence the notation).*

*Proof.* The right hand side is

$$\sum_{x\in V}\Big(\sum_{e^+=x}\theta(e)\frac{1}{C_x}\Big)f(x)C_x = \sum_{x\in V}\sum_{e^+=x}\theta(e)f(x).$$

The left hand side is

$$\frac{1}{2}\sum_{e\in\overleftrightarrow{E}}\left[\big(f(e^+)-f(e^-)\big)\,c(e)\right]\theta(e)\frac{1}{c(e)} = \frac{1}{2}\sum_{e\in\overleftrightarrow{E}}\left[f(e^+)-f(e^-)\right]\theta(e).$$

In this sum, for each $x$ and $e$ such that $e^+=x$, the term $f(x)\theta(e)$ is counted twice: once when $y=e^+$ and once when $y=\widehat{e}^+=e^-$. Since $\theta(\widehat{e})=-\theta(e)$, the two sums are equal. □

A very important exercise is the following:

{ex.selfadj}

▷ **Exercise 6.3.** *Show that the Markov operator $P$ is self-adjoint with respect to the inner product $(\cdot,\cdot)_\pi$, with some $\pi : V\longrightarrow\mathbb{R}_{\geq 0}$, if and only if $\pi$ is a reversible measure for the Markov chain.*

Now, how is the Laplacian related to the boundary and coboundary operators? For any $f : V\longrightarrow\mathbb{R}$, we have

$$\begin{aligned}
\nabla^*\nabla f(x) &= \sum_{e^+=x}\nabla f(e)\frac{1}{C_x}\\
&= \sum_{e^+=x}\left[f(e^+)-f(e^-)\right]\frac{c(e)}{C_x}\\
&= f(x)\sum_{e^+=x}\frac{c(e)}{C_x} - \sum_y f(y)\frac{c(x,y)}{C_x}\\
&= f(x) - \sum_y f(y)p(x,y)\\
&= f(x) - (Pf)(x) = \Delta f(x)\,.
\end{aligned}$$

{d.flow}

**Definition 6.5.** *Let $A$ and $Z$ be disjoint subsets of $V(G)$.*

- *A* **voltage** *between $A$ and $Z$ is a function $f : V \longrightarrow \mathbb{R}$ that is harmonic at every $x \in V \setminus (A \cup Z)$, while $\Delta f|_A \geq 0$ and $\Delta f|_Z \leq 0$.*

- *A* **flow** *from $A$ to $Z$ is a function $\theta : \overleftrightarrow{E} \longrightarrow \mathbb{R}$ with $\theta(\widehat{e}) = -\theta(e)$ such that $\nabla^*\theta(x) = 0$ for every $x \in V \setminus (A \cup Z)$, while $\nabla^*\theta|_A \geq 0$ and $\nabla^*\theta|_Z \leq 0$. In words, it satisfies* **Kirchhoff's node law***: the inflow equals the outflow at every inner vertex.*

- *The* **strength** *of a flow is*

$$\|\theta\| := \sum_{e:e^+ \in A} \theta(e) = \sum_{e:e^- \in Z} \theta(e) = \sum_{a \in A} \nabla^*\theta(a)\, C_a = -\sum_{z \in Z} \nabla^*\theta(z)\, C_z\,,$$

  *the total net amount flowing from $A$ to $Z$. (Or maybe from $Z$ to $A$ — it is unclear what the best terminology is. Electricity flows from higher to lower voltage, but the gradient vector field is pointing towards higher values.)*

- *A main example of a flow is the* **current flow** *$\theta := \nabla f$ associated to a voltage function $f$ between $A$ and $Z$. (It is a flow precisely because of the harmonicity of $f$.)*

**Example 1.** The hitting probability function $f(x) = \mathbf{P}_x[\,\tau_A < \tau_Z\,]$ defined in (6.1) is harmonic on $V \setminus (A \cup Z)$, satisfies $f|_A = 1$ and $f|_Z = 0$, while $f(x) \in [0,1]$ for all $x \in V$, hence $\Delta f|_A \geq 0$ and $\Delta f|_Z \leq 0$. Therefore, $f$ is a voltage function from $A$ to $Z$.

**Example 2.** More generally, consider the function $f(x) = \mathbf{E}_x\big[\,f(X_\tau)\,\big]$ defined in (6.2). If we now define $W_0 = \{u \in W \mid \Delta f(u) < 0\}$ and $W_1 = \{u \in W \mid \Delta f(u) > 0\}$, then $f$ is a voltage from $W_1$ to $W_0$, and $\theta = \nabla f$ is the associated current flow from $W_1$ to $W_0$.

**Example 3.** Flows are the discrete analogue of divergence-free vector fields. Current flows are the gradient vector fields of harmonic functions. Of course, not every flow is a current flow. E.g., on the cycle $C_4 = \{0,1,2,3\}$ with unit conductances, define $\theta(0,1) = \theta(1,2) = -1$, $\theta(0,3) = \theta(3,2) = 0$. Then $\nabla^*\theta(0) = 1 = -\nabla^*\theta(2)$ and $\nabla^*\theta(1) = \nabla^*\theta(3) = 0$, hence $\theta$ is a flow from 0 to 2 of strength 1. However, any current flow from 0 to 2 assigns equal amounts to the four edges $(0,1),(1,2),(0,3),(3,2)$.

**Example 4.** Let $Z \subset V$ be such that $V \setminus Z$ is finite, and let $x \notin Z$. Then

$$G^Z(x,y) := \mathbf{E}_x\big[\,\text{number of times the walk goes through } y \text{ before reaching } Z\,\big]\,,$$

is called Green's function killed at $Z$. It is *almost* the standard Green's function corresponding to the Markov chain killed at $Z$, i.e., with transition probabilities $p^Z(x,y) = p(x,y)\mathbf{1}_{x \notin Z}$. (This is a sub-Markovian chain, i.e., the sum of transition probabilities from certain vertices is less than 1: once in $Z$, the particle is killed instead of moving.) The only difference between $G^Z(x,y)$ and $\sum_{n \geq 0} p_n^Z(x,y)$ is at $x \notin Z, y \in Z$, where the former is zero, the latter is not.

Quite similarly to the previous examples, $G^Z$ is harmonic in its first coordinate outside of $Z \cup \{y\}$:

$$\begin{aligned} G^Z(x,y) &= \sum_{n\geq 1} p_n^Z(x,y) \qquad \text{(since } x \notin Z\cup\{y\}) \\ &= \sum_{n\geq 1}\sum_{x'} p^Z(x,x') p_{n-1}^Z(x',y) \\ &= \sum_{x'} p^Z(x,x') \sum_{n\geq 1} p_{n-1}^Z(x',y) \\ &= \sum_{x'} p^Z(x,x') \sum_{n\geq 0} p_n^Z(x',y) \\ &= \sum_{x'} p^Z(x,x') G^Z(x',y)\,. \end{aligned}$$

To get harmonicity in the second coordinate, we need to change $G^Z$ a little bit. Notice that $C_x\, G^Z(x,y) = C_y\, G^Z(y,x)$, by reversibility when $x,y \notin Z$, and by definition when $x$ or $y$ is in $Z$. Therefore, $f(x) := G^Z(o,x)/C_x = G^Z(x,o)/C_o$ is now harmonic in $x \notin Z \cup \{o\}$. It is clear that $f|_Z = 0$ and $f(o) > 0$, so, by the maximum principle, $\Delta f(o) > 0$ and $\Delta f|_Z \leq 0$. So, $f$ is a voltage function from $o$ to $Z$.

Using Lemma 6.3, we can define a certain electrical distance between disjoint nonempty subsets $A$ and $Z$ of $V(G)$. Namely, consider the unique voltage function $v$ from $A$ to $Z$ with $v|_A \equiv \alpha$ and $v|_Z \equiv \beta$, where $\alpha > \beta$. The associated current flow $i = \nabla v$ has strength $|||i||| > 0$. It is easy to see that

$$\mathcal{R}(A \leftrightarrow Z) := \frac{\alpha - \beta}{|||i|||} \tag{6.3}$$ {e.Reff}

is independent of the values $\alpha > \beta$. It is called the **effective resistance** between $A$ and $Z$. Its inverse $\mathcal{C}(A \leftrightarrow Z) := 1/\mathcal{R}(A \leftrightarrow Z)$ is the **effective conductance**.

▷ **Exercise 6.4.** *Show that effective resistances add up when combining networks in series, while effective conductances add up when combining networks in parallel.*

{ex.Ceff}

▷ **Exercise 6.5.**

**(a)** *Show that for the voltage function $f(x) = G^Z(o,x)/C_x$ of Example 4 above, the associated current flow has unit strength, hence $\mathcal{R}(o \leftrightarrow Z) = G^Z(o,o)/C_o$. (Hint: show that $\nabla f(y,z)$ for $z \in Z$ and $y \notin Z$ is exactly the probability that the random walk hits $Z$ through the edge $(y,z)$.)*

**(b)** *Using (a), show that $\mathcal{C}(o \leftrightarrow Z) = C_o\, \mathbf{P}_o[\,\tau_Z < \tau_o^+\,]$, where $\tau_o^+$ is the first positive hitting time on $o$. (Hint: we have seen the connection between the probability of return and the expected number of visits in Section 1.1.)*

The last exercise provides a great electric interpretation for recurrence of infinite networks $G(V,E,c)$. Consider a sequence of finite cutsets $Z_n \subset V$ between the starting vertex $o$ and infinity, such that each $Z_n$ is also a cutset between $o$ and $Z_{n+1}$. Then the sequence of events $\{\tau_{Z_n} < \tau_o^+\}$ is decreasing in $n$, hence $\lim_{n\to\infty} \mathbf{P}_o[\,\tau_{Z_n} < \tau_o^+\,]$ exists. Moreover, this limit is independent of the sequence $\{Z_n\}$. Indeed, given two such sequences $\{Z_n\}$ and $\{Z'_n\}$, for any $Z_n$ there exists a $Z'_m$ such that $Z_n$ is a cutset between $o$ and $Z'_m$, and vice versa, giving two interlacing subsequences of probabilities, which then must have the same limit. Therefore, Exercise 6.5 (b) says that $\lim_{n\to\infty} \mathcal{C}(o \leftrightarrow Z_n)$ also exists and is independent of the sequence $\{Z_n\}$, and hence may be called the effective conductance between $o$ and infinity, denoted by $\mathcal{C}(o \leftrightarrow \infty)$. It is equal to $C_o\, \mathbf{P}_o[\,\tau_{Z_n} < \tau_o^+ \text{ for all } n\,]$. We claim that this can be written more simply as

$$\mathcal{C}(o \leftrightarrow \infty) = C_o\, \mathbf{P}_o[\,\tau_o^+ = \infty\,]\,. \tag{6.4}$$ {e.CeffEsc}

In particular, the effective conductance to infinity is positive iff the network is transient. To verify (6.4), if the right hand side is 0, then $\mathbf{P}_o[\,\tau_{Z_n} < \tau_o^+ \text{ for all } n\,] = 0$, as well, since the latter is a smaller event. On the other hand, if the right hand side is positive, then the network is transient, and Exercise 1.1 says that the walk spends in any finite set only finite time. Thus every $Z_n$ will be eventually reached almost surely, and hence $\{\tau_o^+ = \infty\}$ actually implies $\{\tau_{Z_n} < \tau_o^+ \text{ for all } n\}$.

{e.hitsym}

▷ **Exercise 6.6.** *Let $G(V, E, c)$ be a transitive network (i.e., the group of graph automorphisms preserving the edge weights have a single orbit on $V$). Show that, for any $u, v \in V$,*

$$\mathbf{P}_u[\,\tau_v < \infty\,] = \mathbf{P}_v[\,\tau_u < \infty\,]\,.$$

▷ **Exercise 6.7** ("Green's function is the inverse of the Laplacian")**.** *Let $(V, P)$ be a transient Markov chain with a stationary measure $\pi$ and associated Laplacian $\Delta = I - P$. Assume that the function $y \mapsto G(x, y)/\pi_y$ is in $L^2(V, \pi)$. Let $f : V \longrightarrow \mathbb{R}$ be an arbitrary function in $L^2(V, \pi)$. Solve the equation $\Delta u = f$.*

## 6.2 Dirichlet energy and transience

{ss.Dirichlet}

In PDE theory, the Laplace equation arises as the Euler-Lagrange variational PDE for minimizing the $L^2$ norm of the gradient. The same phenomenon holds in the discrete setting.

{d.dirichlet}

**Definition 6.6.** *For any $f : V(G) \longrightarrow \mathbb{R}$, define the* **Dirichlet energy** *by*

$$\mathcal{E}(f) := (\nabla f, \nabla f)_r = \frac{1}{2} \sum_{e \in \overleftrightarrow{E}(G)} |f(e^+) - f(e^-)|^2 c(e)\,.$$

*With a slight abuse of notation, for any antisymmetric $\theta : \overleftrightarrow{E} \longrightarrow \mathbb{R}$ we can define $\mathcal{E}(\theta) := (\theta, \theta)_r$.*

{l.harmmin}

**Lemma 6.7.** *The unique harmonic extension in Lemma 6.3 is the unique minimizer of Dirichlet energy.*

{ex.harmmin}

▷ **Exercise 6.8.** *Prove this lemma. (Hint: for a quadratic function $f(x) = \sum_i (x - a_i)^2$, what is the solution of $f'(x) = 0$?)*

Note that an antisymmetric function $\theta : \overleftrightarrow{E}(G) \longrightarrow \mathbb{R}$ is the gradient $\theta = \nabla f$ of some $f : V(G) \longrightarrow \mathbb{R}$ if and only it satisfies **Kirchhoff's cycle law**, i.e., is circulation-free:

$$\sum_{e \in C} \theta(e)\, r(e) = 0 \qquad \text{for all directed cycles } C \subset \overleftrightarrow{E}\,.$$

Thus, Lemma 6.7 can be reformulated as follows: among all antisymmetric $\theta$ satisfying Kirchhoff's cycle law, with a given flux along the boundary (the values $\nabla^*\theta(u)$, $u \in W$), the one that minimizes the Dirichlet energy $\mathcal{E}(\theta)$ also satisfies Kirchhoff's node law in $V \setminus W$ (i.e., it is a flow). There is a dual statement:

{l.Thomson}

**Lemma 6.8** (Thomson's principle)**.** *If $A$ and $Z$ are disjoint subsets of $V(G)$, then among all the flows $\theta$ with given values $\nabla^*\theta|_{A\cup Z}$, the current flow has the smallest energy (i.e., the one that satisfies Kirchhoff's cycle law).*

*Proof.* The strategy is the same as in Exercise 6.8, just need to perturb flows instead of functions: if $\theta$ is a flow with minimal energy $\mathcal{E}(\theta)$, and $C$ is an oriented cycle, then consider the flow $\gamma$ that is constant 1 on $C$, and a simple quadratic computation shows that $\mathcal{E}(\theta + \epsilon\gamma) \geq \mathcal{E}(\theta)$ can hold for all $\epsilon \in \mathbb{R}$ only if $\theta$ satisfies the cycle law along $C$. □

Let us compute this minimum Dirichlet energy for the unique harmonic extension (the voltage $v$) of $v|_A = \alpha$ and $v|_Z = \beta$, where $\alpha > \beta$. Here $\mathcal{E}(v) = (v, \nabla^* \nabla v) = \alpha \sum_{a \in A} C_a \nabla^* \nabla v(a) + \beta \sum_{z \in Z} C_z \nabla^* \nabla v(z) = \alpha |||\nabla v||| - \beta |||\nabla v|||$. So, if the voltage difference $\alpha - \beta$ is adjusted so that we have a unit flow from $A$ to $Z$, i.e., $|||\nabla v||| = 1$, then, by the definition (6.3),

$$\mathcal{E}(v) = \alpha - \beta = \mathcal{R}(A \leftrightarrow Z)\,. \tag{6.5}$$ {e.DReff}

{t.transientfl

**Theorem 6.9** (Terry Lyons [Lyo83])**.** *A graph $G$ is transient if and only if there exists a non-zero flow of finite energy from some vertex $o \in V(G)$ to infinity (i.e., a flow with $A = \{o\}$ and $Z = \emptyset$).*

*Proof.* If $G$ is transient, then Green's function $G(o, x)$ is finite for all $o, x$, hence we can consider $f(x) := G(o, x)/C_x$, as in Example 4 after Definition 6.5. Then $\nabla f$ is a non-zero flow to infinity from $o \in G$, and, by Exercise 6.5 (a), it has unit strength. Its energy, by the above calculation, is $\mathcal{E}(f) = (f, \nabla^* \nabla f) = f(o)|||\nabla f||| = f(o) = G(o, o)/C_o < \infty$, which is indeed finite.

For the other direction, assuming recurrence, take an exhaustion of $G$ by subgraphs $F_n$. As discussed around (6.4), recurrence of $G$ implies that $\mathcal{C}(o \leftrightarrow G \setminus F_n) \to 0$ as $n \to \infty$. So, the effective resistance blows up. By Thomson's principle (Lemma 6.8) and (6.5), this means that there are no unit strength flows from $o$ to $G \setminus F_n$ whose energies remain bounded as $n \to \infty$. On the other hand, if there was a finite energy flow on $G$ from $o$ to infinity, then there would also be one with unit strength, and from that we could produce a unit strength flow of smaller energy from $o$ to each $G \setminus F_n$, contradicting the previous sentence. $\square$

$\triangleright$ **Exercise 6.9.** *Without consulting Lyons (Terry or Russ), find an explicit flow of finite energy on $\mathbb{Z}^3$. (Hint: what would a lot of water or electrons do in $\mathbb{R}^3$ when pumped into the origin? Mimic this in $\mathbb{Z}^3$, and hope that it will have finite energy.)*

Finding a flow of finite energy is a good possible way to prove transience. Unsurprisingly (especially in view of the Max Flow Min Cut theorem), in order to prove recurrence, it is useful to find small cutsets:

{pr.NW}

**Proposition 6.10** (Nash-Williams criterion [NasW61])**.** *If $\{\Pi_n\}_{n \geq 1}$ is a sequence of pairwise disjoint finite cutsets in a locally finite graph $G(V, E)$, each separating $o$ from $\infty$, then*

$$\mathcal{R}(o \leftrightarrow \infty) \geq \sum_{n=1}^{\infty} \frac{1}{\sum_{e \in \Pi_n} c(e)}\,.$$

*Sketch of proof.* Instead of proving this in detail (which can be found, e.g., in [LyPer16, Section 2.5]), let us just discuss a natural special case, in which the intuitive meaning of the formula becomes transparent. Assume that each $\Pi_n$ separates $\Pi_{n-1}$ from $\Pi_{n+1}$. Then we can collapse all vertices between $\Pi_n$ and $\Pi_{n+1}$ into a single vertex $v_n$ (including the collapse of all vertices inside $\Pi_1$ into a $v_0$). By this, we can only decrease the effective resistance, and we obtain a chain of vertices, where $v_{n-1}$ and $v_n$ are joined by parallel edges of total conductance $\sum_{e \in \Pi_n} c(e)$. To get the effective resistance to infinity, we sum up the reciprocals of these conductances, and get the result. $\square$

$\triangleright$ **Exercise 6.10.** *Use the Nash-Williams criterion to prove the recurrence of $\mathbb{Z}^2$.*

In Section 1.1, we computed return probabilities and concluded about recurrence and transience by rather explicit calculations, relying heavily on the structure of $\mathbb{Z}^d$ and $\mathbb{T}_d$. Flows of finite energy provide us with a much more robust tool, as shown by the following result.

Any network $G(V,E,c)$ is naturally a metric space, with

$$d(x,y) := \inf \left\{ \sum_{e\in\gamma} r(e) : \gamma \text{ is a path connecting } x \text{ and } y \right\}.$$

We say that $G$ is uniformly locally finite if $\sup_{x\in V} |B_R(x)| < \infty$ for any radius $R > 0$.

{t.Kanai}

**Theorem 6.11** (Kanai, 1986)**.** *If $\phi : (G_1, d_1) \longrightarrow (G_2, d_2)$ is a quasi-isometric embedding between uniformly locally finite networks, then $\exists\, C < \infty$ such that*

$$\mathcal{E}_1(f \circ \phi) \le C\, \mathcal{E}_2(f).$$

*Furthermore, if $G_1$ is transient, then so is $G_2$.*

*Thus, if $G_1 \simeq_q G_2$, then the $G_i$'s are transient at the same time and $\mathcal{E}_1 \asymp \mathcal{E}_2$ (in the obvious sense that can be read off from the above inequality).*

*Proof.* We will prove only the transience statement, the other being very similar. Suppose $\theta_1$ is a flow of finite energy on $G_1$, say, from $a$ to infinity. For any $e$ in $G_1$, we choose one of the shortest paths in $G_2$ going from the vertex $\phi(e_-)$ to $\phi(e_+)$, in an arbitrary way, and we call this the image of $e$ under $\phi$. Define $\theta_2$ as follows:

$$\theta_2(e_2) = \sum_{e_1 : e_2 \in \phi(e_1)} \pm\theta_1(e_1),$$

where the sign of $\theta_1(e_1)$ depends on the orientation of the path $\phi(e_1)$ with respect to the orientation of $e_2$. Now $\theta_2$ is a flow from $\phi(a)$ to $\infty$, since the contribution of each path $\phi(e)$ to each $\nabla^*\theta_2(x)$ is zero unless $x \in \phi(e_\pm)$. Define

$$\alpha := \sup_{e\in G_1} d_2(\phi(e^+), \phi(e^-))/d_1(e^+, e^-)$$

and

$$\beta := \sup_{e_2 \in G_2} \#\{e_1 \in G_1 : e_2 \in \phi(e_1)\}.$$

Note that $\alpha$ is finite, since $\phi$ cannot increase distances by a lot. Then $\beta$ is finite, as well: since $\alpha < \infty$ and $\phi$ cannot decrease distances by a lot, all edges $e_1$ counted here must come from within some distance $\alpha'$ of the set $\phi^{-1}(e_2)$ of uniformly bounded diameter, which is altogether a finite set of uniformly bounded size, because of the uniform local finiteness of $G_1$. Finally, the energy of $\theta_2$ is finite, since:

$$|\theta_2(e_2)|^2 \le \beta \sum_{e_1 : e_2 \in \phi(e_1)} \theta_1(e_1)^2$$

by the convexity of $x \mapsto x^2$, and then

$$\mathcal{E}(\theta_2) = \sum_{e_2 \in E(G_2)} |\theta_2(e_2)|^2 r(e_2) \le \beta \sum_{e_2} \sum_{e_1 : e_2 \in \phi(e_1)} \theta_1(e_1)^2 r(e_2) \le \alpha\,\beta \sum_{e_1 \in E(G_1)} \theta_1(e_1)^2 r(e_1) < \infty,$$

and we are done. □

## 6.3 Martingales

{ss.MG}

We now define one of the most fundamental notions of probability theory: martingales. We will use them in the evolving sets method of Chapter 8, in the study of bounded harmonic functions in Chapter 9, and in several other results related to random walks. We will also use them in static models where there is no a priori time: to prove Kolmogorov's 0-1 law Theorem 9.20, to prove concentration results for the Erdős-Rényi random graph later in this section, and in the study of the appearance of the giant cluster in the evolution of the Erdős-Rényi random graph in Section 12.3.

**Definition 6.12.** *A sequence $M_1, M_2, \dots$ of $\mathbb{R}$-valued random variables (or in any value set where averaging makes sense) is called a* **martingale** *if $\mathbf{E}[\, M_{n+1} \mid M_1, \dots, M_n \,] = M_n$. More generally, given an increasing sequence of $\sigma$-algebras, $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \dots$ (called a filtration), and $M_n$ is measurable w.r.t. $\mathscr{F}_n$ for each $n$, we want that $\mathbf{E}[\, M_{n+1} \mid \mathscr{F}_n \,] = M_n$.*

**Example 1**: We start with a trivial example: if $M_n = \sum_{k=1}^n X_k$ with $\mathbf{E}[\, X_k \,] = 0$ for all $k$, with the $X_k$'s being independent, then $\mathbf{E}[\, M_{n+1} \mid M_1, \dots, M_n \,] = \mathbf{E}[\, X_{n+1} \,] + M_n = M_n$.

**Example 2**: A classical source of martingales is gambling. In a completely fair casino, if a gambler chooses based on her history of winnings what game to play next and in what value, then the increments in her fortune will not be i.i.d., but her fortune will be a martingale: she cannot make and cannot lose money in expectation, whatever she does. In particular, for any martingale,

$$\begin{aligned} \mathbf{E}\big[\, M_k \bigm| M_0 \,\big] &= \mathbf{E}\Big[\, \mathbf{E}\big[\, M_k \bigm| M_{k-1}, M_{k-2}, \dots, M_0 \,\big] \Bigm| M_0 \,\Big], \text{ by Fubini} \\ &= \mathbf{E}\big[\, M_{k-1} \bigm| M_0 \,\big], \text{ by being a martingale} \\ &= \dots = M_0, \text{ by iterating}, \end{aligned} \tag{6.6}$$ {e.MGE1}

and

$$\mathbf{E}[\, M_k \,] = \mathbf{E}[\, M_{k-1} \,] = \dots = \mathbf{E}[\, M_0 \,]. \tag{6.7}$$ {e.MGE2}

A famous gambling example is the "**double the stake until you win**" strategy: **0.** start with fortune $M_0 = 0$; **1.** borrow one dollar, double or lose it with probability $1/2$ each, arriving at a fortune $M_1 = 2 - 1 = 1$ or $M_1 = 0 - 1 = -1$; **2.** if the first round is lost, then borrow two more dollars, double or lose it with probability $1/2$ each; **3.** if the second round is also lost, then borrow four more dollars, and so on, until first winning a round, and then pay back all the debts. This will eventually happen almost surely, in the random $\tau$th round, and then the net payoff is $-1 - 2 - \dots - 2^{\tau-1} + 2^\tau = 1$, so this is an almost certain way to make money in a fair game — assuming one has a friend with infinite fortune to borrow from! Now, it is easy to see that the fortune $M_k$ (set to be constant from time $\tau$ on) is a martingale: on the event $\{M_k = 1\}$, we have $M_{k+1} = 1 = M_k$ a.s., while $\mathbf{E}\big[\, M_{k+1} \bigm| M_k = -1 - 2 - \dots - 2^{k-1} \,\big] = M_k + 1/2 \cdot 2^k - 1/2 \cdot 2^k = M_k$. But then, how does $\mathbf{E}M_k = \mathbf{E}M_0 = 0$ square with $\mathbf{E}M_\tau = 1$? Well, $\tau$ is a random time, so why would (6.7) apply? We will further discuss this issue a few paragraphs below, under Optional Stopping Theorems.

**Example 3**: $G = (V, E, c)$ is a network, $W \subseteq V$ is given (thought of as some sort of boundary), and $f : V \longrightarrow \mathbb{R}$ is harmonic on $V \setminus W$; see Definition 6.2. Define $M_n = f(X_n)$, where $X_n$ is the random walk on the network, and let $\tau \in \mathbb{N}$ be the time of hitting $W$. It may be that $\tau = \infty$ with positive probability. On the event $\tau < \infty$, we set $X_{\tau+i} = X_\tau$ for all $i \in \mathbb{N}$. Now, by the Markov property of $X_n$ and the harmonicity of $f$, we have

$$\mathbf{E}[\, M_{n+1} \mid X_1, \dots, X_n,\ n < \tau \,] = \mathbf{E}\big[\, f(X_{n+1}) \bigm| X_n \,\big] = f(X_n) = M_n\,,$$

and the same holds trivially also on the event $\{n \geq \tau\}$ instead of $\{n < \tau\}$, thus $M_n$ is a martingale w.r.t. $\mathscr{F}_n = \sigma(X_0, \dots, X_n)$. It is also a martingale in the more restricted sense: given the sequence

$(M_n)_{n=0}^\infty$ only,

$$\begin{aligned}\mathbf{E}[\, M_{n+1} \mid M_n, \dots, M_0 \,] &= \sum_{\substack{x_0,\dots,x_n:\\ f(x_i)=M_i\ \forall i}} \mathbf{E}[\, f(X_{n+1}) \mid X_i = x_i \ \forall i \le n \,]\, \mathbf{P}[\, X_i = x_i \ \forall i \le n \mid M_n, \dots, M_0 \,] \\ &= \sum_{\substack{x_0,\dots,x_n:\\ f(x_i)=M_i\ \forall i}} f(x)\, \mathbf{P}[\, X_i = x_i \ \forall i \le n \mid M_n, \dots, M_0 \,] \\ &= f(x_n) = M_n \,.\end{aligned}$$

This averaging argument can be used in general to show that it is easier to be a martingale w.r.t. to a filtration of smaller (coarser) $\sigma$-algebras.

▷ **Exercise 6.11.** *Give an example of a random sequence $(M_n)_{n=0}^\infty$ such that $\mathbf{E}[\, M_{n+1} \mid M_n \,] = M_n$ for all $n \ge 0$, but which is not a martingale w.r.t. the filtration $\mathscr{F}_n = \sigma(M_0, \dots, M_n)$.*

**Example 4**: If we have a filtration $\mathscr{F}_n \uparrow \mathscr{F}_\infty$, possibly $\mathscr{F}_\infty = \mathscr{F}_N$ for some finite $N$, and $Y$ is an integrable variable in $\mathscr{F}_\infty$, then $M_n := \mathbf{E}[\, Y \mid \mathscr{F}_n \,]$ is a martingale:

$$\begin{aligned}\mathbf{E}[\, M_{n+1} \mid \mathscr{F}_n \,] &= \mathbf{E}\big[\, \mathbf{E}[\, Y \mid \mathscr{F}_{n+1} \,] \bigm| \mathscr{F}_n \,\big] \\ &= \mathbf{E}[\, Y \mid \mathscr{F}_n \,] = M_n \,.\end{aligned}$$

In words: our best guesses about a random variable $Y$, as we learn more and more information about it, form a martingale. As we will see, this is a far-reaching idea.

It might not be obvious at first sight, but Example 3 is a special case of Example 4, at least in the case when $\tau < \infty$ almost surely. As we saw in Section 6.1, given $f : W \longrightarrow \mathbb{R}$, one harmonic extension to $V \setminus W$ is given by $f(x) = \mathbf{E}_x[\, f(X_\tau) \,]$. So, taking $\mathscr{F}_n = \sigma(X_0, \dots, X_n)$ and $\mathscr{F}_\infty = \sigma(X_0, X_1, \dots)$, we have that $Y = f(X_\tau)$ is $\mathscr{F}_\infty$-measurable, and $M_n = \mathbf{E}[\, Y \mid \mathscr{F}_n \,]$ equals $f(X_n)$. A generalization of this correspondence between harmonic functions and limiting values of random walk martingales, for the case when $\tau = \infty$ is a possibility, will be Theorem 9.25.

Another instance of Example 4, very different from random walks, but typical in probabilistic combinatorics, is the edge- and vertex-exposure martingales for the **Erdős-Rényi random graph** model $G(n,p)$. An instance $\mathcal{G}$ of this random graph is generated by having each edge of the complete graph $K_n$ on $n$ vertices be present with probability $p$ and missing with probability $1-p$, independently from each other. In other words, the probability space is all subgraphs of $K_n$, with measure $\mathbf{P}[\, \mathcal{G} = H \,] = p^{|E(H)|}(1-p)^{\binom{n}{2} - |E(H)|}$ for any $H$. Now, given a variable $Y$ on this probability space, e.g., the chromatic number $\chi(\mathcal{G})$, we can associate to it two martingales, as follows. Fix an ordering $e_1, \dots, e_m$ of the edges of $K_n$, where $m = \binom{n}{2}$. Let $\mathcal{G}[e_1, \dots, e_i] \in \{0,1\}^{\{e_1,\dots,e_i\}}$ be the states of the edges $e_1, \dots, e_i$ in $\mathcal{G}$, and let $M_i^E(\mathcal{G}) := \mathbf{E}\big[\, Y \bigm| \mathcal{G}[e_1, \dots, e_i] \,\big]$ for $i = 0, 1, \dots, m$. This is the **edge exposure martingale** associated to $Y$, a random sequence determined by $\mathcal{G}$, with $M_0^E = \mathbf{E}[\, Y \,]$ and $M_m^E = Y$. Similarly, one can fix an ordering $v_1, \dots, v_n$ of the vertices, let $\mathcal{G}[v_1, \dots, v_i]$ be the states of the $\binom{i}{2}$ edges spanned by $v_1, \dots, v_i$, and let $M_i^V(\mathcal{G}) := \mathbf{E}\big[\, Y \bigm| \mathcal{G}[v_1, \dots, v_i] \,\big]$ for $i = 0, 1, \dots, n$. This is the **vertex exposure martingale**. One reason for the interest in these martingales is that the Azuma-Hoeffding inequalities, Propositions 1.8 and 1.9, can be applied to prove the concentration of certain variables $Y$ around their mean (even if the value of the mean is unknown). For instance, when $Y = \chi(\mathcal{G})$, then we have the Lipschitz property $|M_{i+1}^V - M_i^V| \le 1$ almost surely, hence, for $M_n^V = \chi(\mathcal{G})$, Proposition 1.8 and Exercise 1.8 give

$$\mathbf{P}\big[\, |\chi(\mathcal{G}) - \mathbf{E}\chi(\mathcal{G})| > \lambda\sqrt{n} \,\big] \le 2\exp(-\lambda^2/2) \,. \tag{6.8}$$ {e.chromcont}

The reason for the Lipschitz property is that we clearly have $|\chi(H) - \chi(H')| \le 1$ if $H$ and $H'$ are two graphs on the same vertex set whose symmetric difference is a set of edges incident to a single vertex. Now, when we reveal the states of the edges between $v_{i+1}$ and $\{v_1, \dots, v_i\}$, we can write $M_i^V = \mathbf{E}\big[\, M_{i+1}^V \bigm| \mathcal{G}[v_1, \dots, v_i] \,\big]$, and think of the right hand side as a double averaging: first over the edges not spanned by $\{v_1, \dots, v_i, v_{i+1}\}$, then over the edges between $v_{i+1}$ and $\{v_1, \dots, v_i\}$. Fixing the states of the edges not spanned by $\{v_1, \dots, v_i, v_{i+1}\}$ and varying the states of the edges between $v_{i+1}$ and $\{v_1, \dots, v_i\}$, we get a set of possible outcomes whose $Y$-values differ by at most 1. Hence the $M_{i+1}^V$-values, given $M_i^V$, all differ by at most 1. But their average is $M_i^V$, hence they can all be at distance at most 1 from $M_i^V$, as claimed.

We give here two more examples of this kind. Yet another one can be found in Exercise 14.4 part (b).

{ex.chromcont2

▷ **Exercise 6.12.** *Let $G = (V, E)$ be an arbitrary finite graph, $1 \le k \le |V|$ an integer, and $\mathcal{K}$ a uniform random $k$-element subset of $V(G)$. Then the chromatic number $\chi(G[\mathcal{K}])$ of the subgraph of $G$ spanned by $\mathcal{K}$ is concentrated: for the number $c(G,k) := \mathbf{E}[\,\chi(G[\mathcal{K}])\,]$, we have*

$$\mathbf{P}\big[\, \big|\chi(G[\mathcal{K}]) - c(G,k)\big| > \epsilon k \,\big] \le \exp(-\epsilon^2 k/2)\,.$$

The **concentration of measure** phenomenon shown by the next exercise is strongly related to isoperimetric inequalities in high-dimensional spaces. (See Exercise 5.12 and the discussion around Exercises 12.49 and 12.50.) For a subset $A$ of the hypercube $\{0,1\}^n$, let $B(A,t) := \big\{x \in \{0,1\}^n : \mathrm{dist}(x, A) \le t\big\}$.

{ex.hypcubeinc

▷ **Exercise 6.13.** *Let $\epsilon, \lambda > 0$ be constants satisfying $\exp(-\lambda^2/2) = \epsilon$. Then, for $A \subset \{0,1\}^n$,*

$$|A| \ge \epsilon\, 2^n \quad \Longrightarrow \quad \big|B(A, 2\lambda\sqrt{n})\big| \ge (1-\epsilon)\, 2^n\,.$$

*That is, even small sets become huge if we enlarge them a little.*

There are two basic and very useful groups of results regarding martingales. One is known as **Martingale Convergence Theorems**: e.g., any bounded martingale $M_n$ converges to some limiting variable $M_\infty$, almost surely and in $L^1$. An example of this was Example 3, in the case when $f$ is bounded and $\tau < \infty$ almost surely. More generally, in Example 4, we have a natural candidate for the limit: $M_n \to Y$. This convergence follows from $\mathscr{F}_n \uparrow \mathscr{F}_\infty$, but in a non-trivial way, known as Lévy's 0-1 law (see Theorem 9.19 in Section 9.4). We will state but not prove a general version of the Martingale Convergence Theorem as Theorem 9.8 in Section 9.2; a thorough source is [Dur10, Chapter 5].

One version of the Martingale Convergence Theorem implies that Example 4 is not at all that special: the class of martingales arising there coincides with the **uniformly integrable** ones:

$$\lim_{K \to \infty} \sup_{n \ge 0} \mathbf{E}\big[\, M_n \mathbf{1}_{\{|M_n| > K\}} \,\big] = 0\,, \tag{6.9}$$ {e.UI}

where the corresponding $Y$ is the a.s. and $L^1$-limit of $M_n$, as $n \to \infty$; see [Dur10, Theorem 5.5.6].

The second important group of results about martingales is the **Optional Stopping Theorems** (OST): given a stopping time $\tau$ for a martingale (i.e., an integer-valued random variable $\tau$ such that the event $\{\tau > k\}$ is $\mathscr{F}_k$-measurable for all $k \in \mathbb{N}$), when do we have $\mathbf{E}[\, M_\tau \,] = \mathbf{E}[\, M_0 \,]$? In Example 3, assuming that $\tau < \infty$ a.s., we had $\mathbf{E}_x[\, M_\tau \,] = \mathbf{E}_x[\, Y \,] = f(x) = M_0$ almost by definition. More generally, if $M_n$ is a uniformly integrable martingale, as defined in (6.9), and $\tau < \infty$ a.s., then $\mathbf{E}[\, M_\tau \,] = \mathbf{E}[\, M_0 \,]$ does hold, see [Dur10, Section 5.7]. On the other hand, in Example 2, we had $\mathbf{E}[\, M_\tau \,] = 1 \ne 0 = M_0$. An even simpler counterexample is SRW on $\mathbb{Z}$, started from 1, viewed as a martingale $\{M_n\}_{n=0}^\infty$, with $\tau$ being the first hitting time on 0. By recurrence, $\tau < \infty$ a.s., but $\mathbf{E}[\, M_\tau \,] = 0 \ne 1 = M_0$.

Let us sketch the proof of the Optional Stopping Theorem for bounded martingales. $\tilde{M}_n := M_{n\wedge\tau}$ is a martingale again, hence $\mathbf{E}[\,\tilde{M}_n\,] = \mathbf{E}[\,\tilde{M}_0\,] = \mathbf{E}[\,M_0\,]$ for any $n \in \mathbb{N}$. On the other hand, $\tau < \infty$ (a.s.) implies that $\lim_{n\to\infty} \tilde{M}_n = M_\tau$ almost surely. Hence the Dominated Convergence Theorem says that $\mathbf{E}[\,M_\tau\,] = \lim_{n\to\infty} \mathbf{E}[\,\tilde{M}_n\,] = \lim_{n\to\infty} \mathbf{E}[\,M_0\,] = \mathbf{E}[\,M_0\,]$, as desired.

We have already seen applications of martingales to concentration results and to harmonic functions defined on general graphs, and we will see more later, but let us demonstrate now that martingale techniques are useful even in the simplest example, random walk on $\mathbb{Z}$.

Start a symmetric simple random walk $X_0, X_1, \dots$ at $k \in \{0, 1, \dots, n\}$, and stop it when first hitting 0 or $n$, at time $\tau = \tau_0 \wedge \tau_n$. What is $h(k) := \mathbf{P}_k[\,\tau_0 > \tau_n\,]$? Since $(X_i)$ is a bounded martingale, we have $\mathbf{E}_k[\,X_\tau\,] = k$ by the Optional Stopping Theorem. On the other hand, $\mathbf{E}_k[\,X_\tau\,] = h(k)\cdot n + (1-h(k))\cdot 0$, thus $h(k) = k/n$. This is of course the harmonic extension of $h(0) = 0$ and $h(n) = 1$, which one could also get by solving a system of linear equations, which can probably be considered less elegant than using the OST. Finer information on $\tau$ can also be obtained from the Optional Stopping Theorem:

{ex.HittingTim

▷ **Exercise 6.14.** *Let $(X_i)_{i\geq 0}$ be a random walk on $\mathbb{Z}$, with i.i.d. increments $\xi_i$ that have zero mean and a bounded support. Let the moments of $\xi$ be denoted by $\mathbf{E}[\,\xi^i\,] = \mu_i$. Starting from $X_0 = \ell \in \{1, 2, \dots, k-1\}$, let $\tau_0$ be the first time the walk is at most 0, and let $\tau_k$ be the first time the walk is at least $k$.*

- **(a)** *For any $0 < \ell < k$, prove $\mathbf{P}_\ell[\,\tau_k < \tau_0\,] \asymp \ell/k$, with constants in $\asymp$ that are close to 1 if both $k$ and $\ell$ are large compared to the support of the jump distribution.*
- **(b)** *Show that $X_n^2 - \mu_2 n$ is a martingale. Proving an exponential tail for $\tau_k \wedge \tau_0$ (which depends on $k$), prove that it is uniformly integrable (6.9). Using the OST, deduce that $\mathbf{E}_\ell[\,\tau_k \wedge \tau_0\,] \asymp \ell(k-\ell)/\mu_2$.*
- **(c)** *Deduce from part (b) that $\mathbf{E}_\ell[\,\tau_k \wedge \tau_0 \mid \tau_k < \tau_0\,] \leq Ck(k-\ell)$, with a constant $C$ that depends only on the jump distribution.*
- **(d)** *Find a martingale of the form $X_n^4 + AnX_n^2 + BnX_n + Cn^2 + Dn$ with some constants $A, B, C, D \in \mathbb{R}$. (Solution: $A = -6{\mu_2}^{1/2}$, $B = -4\mu_3$, $C = 3\mu_2^{3/2}$, $D = 3\mu_2^{3/2} - \mu_4$.) Use this and the previous parts to prove the second moment bound $\mathbf{E}_\ell[\,(\tau_k \wedge \tau_0)^2\,] \leq O(\ell k^3)$.*
- **(e)** *Using the expectation result of part (b), improve the tail bound on $\tau_k \wedge \tau_0$: show for some $c = c(\mu_2) > 0$, uniformly for all $\ell$, that $\mathbf{P}_\ell[\,\tau_k \wedge \tau_0 > n\,] < \exp(-c\,n/k^2)$.*
- **(f)** *Using part (d) and the Second Moment Method (see (12.15) in case you do not know what that is), prove that $\mathbf{P}_{ck}[\,\tau_0 \wedge \tau_k > k^2\,] > \delta(c) > 0$ for any $c \in (0,1)$, uniformly for all large enough $k$.*

*In summary, starting from $X_0 = ck$ with $c \in (0,1)$ fixed, the sequence of random variables $(\tau_0 \wedge \tau_k)/k^2$ is tight and non-vanishing. In fact, Donsker's functional limit theorem [Dur10, Theorem 8.6.5] implies that there is a limit distribution, given by the hitting time of a Brownian motion. Note here that the hitting time is not a continuous functional on the space of continuous functions, hence Donsker's theorem cannot be applied automatically. Nevertheless, Brownian trajectories are almost surely continuity points of the functional (which needs a proof), hence the natural extension [Dur10, Theorem 8.6.6] applies.*

One can use similar ideas also in the asymmetric case:

{ex.asymMG}

▷ **Exercise 6.15.** *Consider asymmetric simple random walk $(X_i)$ on $\mathbb{Z}$, with probability $p > 1/2$ for a right step and $1-p$ for a left step. Find a martingale of the form $r^{X_i}$ for some $0 < r \neq 1$, and calculate $\mathbf{P}_k[\,\tau_0 > \tau_n\,]$. Then find a martingale of the form $X_i - \mu\, i$ for some $\mu > 0$, and calculate $\mathbf{E}_k[\,\tau_0 \wedge \tau_n\,]$. (Hint: to prove that the second martingale is uniformly integrable, first show that $\tau_0 \wedge \tau_n$ has an exponential tail.)*

Now, condition the symmetric simple random walk $(X_i)$ to reach $n$ before 0. This conditioning concerns the entire trajectory, hence it might happen, a priori, that we get a complicated non-Markovian process. A simple but beautiful result is that, in fact, we get a nice Markov process. This construction is a version of **Doob's $h$-transform** [Doo59], with $h$ being the harmonic function $h(k) := \mathbf{P}_k[\mathcal{A}]$ below:

{l.condMC}

**Lemma 6.13.** *Let $(X_i)_{i=0}^{\infty}$ be any time-homogeneous Markov chain on the state space $\mathbb{N}$, and $\mathcal{A} := \{\tau_A < \tau_Z\}$ for some $A, Z \subset \mathbb{N}$ (more generally, it could be any event in the invariant $\sigma$-field of the chain, see Definition 9.21 in Section 9.4). Then $(X_i)$ conditioned on $\mathcal{A}$ is again a Markov chain, with transition probabilities*

$$\mathbf{P}[\, X_{i+1} = \ell \mid X_i = k, \mathcal{A} \,] = \frac{\mathbf{P}_\ell[\,\mathcal{A}\,]}{\mathbf{P}_k[\,\mathcal{A}\,]}\, \mathbf{P}[\, X_{i+1} = \ell \mid X_i = k \,]\,,$$

*where $\mathbf{P}_k[\,\mathcal{A}\,] = \mathbf{P}[\,\mathcal{A} \mid X_0 = k\,]$ is supposed to be positive.*

*Proof.* Note that $\mathbf{P}[\,\mathcal{A} \mid X_{i+1} = \ell, X_i = k\,] = \mathbf{P}[\,\mathcal{A} \mid X_{i+1} = \ell\,] = \mathbf{P}_\ell[\,\mathcal{A}\,]$. Then,

$$\begin{aligned}
\mathbf{P}[\, X_{i+1} = \ell \mid X_i = k, \mathcal{A} \,] &= \frac{\mathbf{P}[\, X_{i+1} = \ell, X_i = k, \mathcal{A} \,]}{\mathbf{P}[\, X_i = k, \mathcal{A} \,]} \\
&= \frac{\mathbf{P}[\,\mathcal{A} \mid X_{i+1} = \ell, X_i = k\,]\, \mathbf{P}[\, X_{i+1} = \ell, X_i = k \,]}{\mathbf{P}[\,\mathcal{A} \mid X_i = k\,]\, \mathbf{P}[\, X_i = k \,]} \\
&= \frac{\mathbf{P}_\ell[\,\mathcal{A}\,]}{\mathbf{P}_k[\,\mathcal{A}\,]}\, \mathbf{P}[\, X_{i+1} = \ell \mid X_i = k \,]\,,
\end{aligned}$$

as claimed. □

Back to our example, if $(X_i)$ is simple random walk with $X_0 = k$, stopped at 0 and $n$, and $\mathcal{A} = \{\tau_n < \tau_0\}$, then $\mathbf{P}_k[\,\mathcal{A}\,] = k/n$. Therefore, the new transition probabilities are

$$p(k, k-1) = \frac{k-1}{2k}\,, \qquad p(k, k+1) = \frac{k+1}{2k}\,, \tag{6.10}$$ {e.Bessel3}

for $k = 1, \ldots, n-1$. Note the consistency property that these values do not depend on $n$. In particular, the conditional measures have a weak limit as $n \to \infty$: the Markov chain with transition probabilities given in (6.10) for all $k = 1, 2, \ldots$. This chain can naturally be called **SRW on $\mathbb{Z}$ conditioned not to ever hit zero**. Since the drift at level $k$ is $+1/k$, this is the discrete analogue of the Bessel(3) process $dX_t = \frac{1}{X_t} dt + dB_t$, for those who have or will see stochastic differential equations. The reason for the index 3 is that the Bessel($n$) process, given by $dX_t = \frac{n-1}{2X_t} dt + dB_t$, is the Euclidean distance of an $n$-dimensional Brownian motion from the origin. (But I do not think that anyone knows a direct combinatorial link between Brownian motion in $\mathbb{R}^3$ and the conditioned one in $\mathbb{R}$.)

▷ **Exercise 6.16.** *Show that the conditioned random walk (6.10) is transient. (Hint: construct an electric network on $\mathbb{N}$ that gives rise to this random walk, then use Sections 6.1 and/or 6.2. Or is the statement obvious from the construction?)*

▷ **Exercise 6.17.** *Show that the Doob $h$-transform of any reversible Markov chain w.r.t. any discrete harmonic function $h$ is again reversible.*

The following result on random walks on $\mathbb{Z}$ will be used in our proof of the Erdős-Rényi giant cluster phase transition, Theorem 12.23.

{ex.returntail

▷ **Exercise 6.18.** *Let $(X_i)_{i\geq 0}$ be a random walk on $\mathbb{Z}$, with i.i.d. increments $\xi_i$ that have zero mean and an exponential tail: there exist $K \in \mathbb{N}$ and $0 < q < 1$ such that $\mathbf{P}[\,\xi \geq k+1\,] \leq q\,\mathbf{P}[\,\xi \geq k\,]$ for all $k \geq K$.*

*Starting from $X_0 = \ell \in \{1, 2, \dots, n-1\}$, let $\tau_0$ be the first time the walk is at most 0, and let $\tau_n$ be the first time the walk is at least $n$. Show that, for $X_0 = 1$,*

$$\mathbf{P}_1[\,\tau_0 > n\,] \asymp \mathbf{P}_1[\,\tau_{\sqrt{n}} < \tau_0\,] \asymp 1/\sqrt{n}.$$

*In more detail:*

**(a)** *For any $0 < \ell < n$, prove $\mathbf{P}_\ell[\,\tau_n < \tau_0\,] \asymp \ell/n$. For this, first prove that $X_{\tau_n} - n$, conditioned on $\tau_n < \tau_0$, has an exponential tail, independently of $n$. This proves the second equality.*

*For the first equality, we need some basic understanding of $M_n = \max_{0\leq i\leq n} X_i$, with $X_0 = 0$.*

**(b)** *From Donsker's theorem on $\{X_{tn}/(\sigma\sqrt{n}) : t \geq 0\}$ converging to Brownian motion, show that $M_n/\sqrt{n}$ has a continuous limit distribution.*

**(c)** *Now notice that $\mathbf{P}_1[\,\tau_0 > n \mid \tau_{\sqrt{n}} < \tau_0\,] > \mathbf{P}_0[\,M_n < \sqrt{n}\,] > c_1 > 0$.*

**(d)** *Show that $X_i \mid \{\tau_0 > n\}$ stochastically dominates the unconditioned $X_i$, hence $\mathbf{P}_1[\,\tau_{\sqrt{n}} < \tau_0 \mid \tau_0 > n\,] > \mathbf{P}_1[\,M_n > \sqrt{n}\,] > c_2 > 0$.*

*The above ingredients together give the result.*

*One can give a proof using only martingale arguments, avoiding the use of the Brownian limit, as follows:*

**(d)** *Extend Exercise 6.14 from bounded jump distributions to the exponential tail considered here.*

**(e)** *Consider the following decomposition, according to how big $\max_{0\leq i\leq \tau_0} X_i$ roughly is:*

$$\begin{aligned}\mathbf{P}_1[\,\tau_0 > n\,] &= \mathbf{P}_1\big[\,\tau_0 > n \;\big|\; \tau_{\sqrt{n}} < \tau_0\,\big]\,\mathbf{P}_1[\,\tau_{\sqrt{n}} < \tau_0\,] \\ &\quad + \sum_{k=0}^{\lfloor \log_2 \sqrt{n}\rfloor} \mathbf{P}_1\Big[\,\tau_0 > n \;\Big|\; \tau_{\sqrt{n}/2^{k+1}} < \tau_0 < \tau_{\sqrt{n}/2^k}\,\Big]\,\mathbf{P}_1\big[\,\tau_{\sqrt{n}/2^{k+1}} < \tau_0 < \tau_{\sqrt{n}/2^k}\,\big]\,.\end{aligned}$$

*Using parts (a), (e) from the extended Exercise 6.14, show that the first line on the RHS is $\asymp 1/\sqrt{n}$. Using a Doob transform argument and parts (a), (d), show that the second line is at most $O(1/\sqrt{n})$.*

# 7 Cheeger constant and spectral gap

{s.Cheeger}

The previous chapter introduced a certain geometric view on reversible Markov chains: many natural dynamically defined objects (hitting probabilities, Green's functions) turned out to have good encodings as harmonic functions over the associated electric network, and a basic probabilistic property (recurrence versus transience) turned out to have a useful reformulation via the Dirichlet energy of flows (the usefulness having been demonstrated by Kanai's quasi-isometry invariance Theorem 6.11).

We will now make these connections even richer: isoperimetric inequalities satisfied by the underlying graph (the electric network) will be expressed as linear algebraic or functional analytic properties of the Markov operator acting on functions over the state space, which can then be translated into probabilistic behaviour of the Markov chain itself.

## 7.1 Spectral radius and the Markov operator norm

{ss.specrad}

Consider some Markov chain $P$ on the graph $(V, E)$. We will assume in this section that $P$ is reversible. Given any reversible measure $\pi$, we can consider the associated electric network $c(x, y) = \pi(x)p(x, y) = c(y, x)$, $C_x = \sum_y c(x, y)$, and the usual inner products $(\cdot, \cdot)_C$ on $\ell_0(V)$ and $(\cdot, \cdot)_r$ on $\ell_0(\overleftrightarrow{E})$, where $\ell_0$ stands for functions with finite support. Note that $C_x = \pi(x)$ now.

The Markov operator $P : \ell_0(V) \longrightarrow \ell_0(V)$, introduced in Definition 6.2, satisfies

$$\|P\| = \sup_{f \in \ell_0(V)} \frac{\|Pf\|}{\|f\|} \leq 1 \,,$$

with $\|\cdot\|$ denoting $\ell^2$-norm w.r.t. $\pi$. The reason for this bound is that $P$ is an averaging operator, while $f \mapsto \|f\|^2$ is a convex function of the variables $f(x), x \in V$, hence we can use Jensen's inequality, plus the stationarity of $\pi$:

$$\begin{aligned} \|Pf\|^2 = \sum_x \pi(x) \left| \sum_y p(x,y) f(y) \right|^2 &\leq \sum_x \pi(x) \sum_y p(x,y) |f(y)|^2 \\ &= \sum_y |f(y)|^2 \sum_x \pi(x) p(x,y) = \sum_y |f(y)|^2 \pi(y) = \|f\|^2 \,. \end{aligned}$$

Now, since $\ell_0(V)$ is dense in $\ell^2(V, \pi)$, we can extend $P$ to an operator $\ell^2(V, \pi) \longrightarrow \ell^2(V, \pi)$ without increasing the operator norm, and Exercise 6.3 implies that this extension is self-adjoint:

$$(Pf, g)_\pi = (f, Pg)_\pi \,.$$

Observe furthermore that the Dirichlet energy can be written as

$$\mathcal{E}(f) = \mathcal{E}_{P,\pi}(f) = (\nabla f, \nabla f) = (f, \nabla^* \nabla f) = (f, (I - P)f) = \|f\|^2 - (f, Pf) \,. \tag{7.1}$$ {e.DIP}

Recall the definition of the spectral radius from (1.8),

$$\rho(P) = \limsup_{n \to \infty} p_n(o, o)^{1/n} \,.$$

Now here is the reason for calling $\rho(P)$ the spectral radius:

{p.specrad}

**Proposition 7.1.** $\|P\| = \rho(P)$.

*Proof.* Using self-adjointness,

$$\begin{aligned} \big(\pi(o) p_{2n}(o, o)\big)^{\frac{1}{2n}} = (P^{2n}\delta_o, \delta_o)^{\frac{1}{2n}} = (P^n \delta_o, P^n \delta_o)^{\frac{1}{2n}} &= \|P^n \delta_o\|^{1/n} \\ &\leq \big(\|P\|^n \, \|\delta_o\|\big)^{1/n} = \|P\| \, \pi(o)^{1/n} \,. \end{aligned}$$

Letting $n \to \infty$, we almost find that $\rho(P) \leq \|P\|$; the missing step is the following exercise:

▷ **Exercise 7.1.** *Show that in any reversible chain,* $\limsup_{n\to\infty} p_n(o, o)^{1/n} = \limsup_{n\to\infty} p_{2n}(o, o)^{\frac{1}{2n}}$. *(Hint: you may want to use Lemma 9.1 from the beginning of Section 9.1.)*

To prove the direction $\|P\| \leq \rho(P)$, we want to show that the behavior of $\|P\delta_o\|$ detects the behavior of $\|Pf\|$ for any function $f \in \ell_0(V)$. To this end, start with

$$\begin{aligned} \|P^{n+1} f\|^2 &= (P^{n+1} f, P^{n+1} f) \\ &= (P^n f, P^{n+2} f) \\ &\leq \|P^n f\| \cdot \|P^{n+2} f\| \,. \end{aligned}$$

The second equality holds since $P$ is self-adjoint, and the final step is by the Cauchy-Schwarz inequality. So, we get

$$\frac{\|P^{n+1}f\|}{\|P^n f\|} \leq \frac{\|P^{n+2}f\|}{\|P^{n+1}f\|}\,. \tag{7.2}$$ {e.ratioineq}

For any non-negative sequence $(a_n)_{n=1}^{\infty}$ we know that:

$$\liminf_{n\to\infty} \frac{a_{n+1}}{a_n} \leq \liminf_{n\to\infty} a_n^{1/n} \leq \limsup_{n\to\infty} a_n^{1/n} \leq \limsup_{n\to\infty} \frac{a_{n+1}}{a_n}\,.$$

By (7.2), the sequence $\frac{\|P^{n+1}f\|}{\|P^n f\|}$ has a limit, and hence

$$\frac{\|Pf\|}{\|f\|} \leq \lim_{n\to\infty} \frac{\|P^{n+1}f\|}{\|P^n f\|} = \lim_{n\to\infty} \|P^n f\|^{1/n}\,.$$

So, it is enough to show that $\limsup_{n\to\infty} \|P^n f\|^{1/n} \leq \rho(P)$. We have

$$\begin{aligned}
\|P^n f\|^2 &= (f, P^{2n} f) \\
&= \sum_x f(x) P^{2n} f(x) \pi(x) \\
&= \sum_{x,y} f(x) f(y) p_{2n}(x,y) \pi(x) \\
&\leq \sum_{x,y} f(x) f(y) \mathbf{1}_{\{f(x)f(y)>0\}}\, p_{2n}(x,y)\pi(x)\,.
\end{aligned} \tag{7.3}$$ {e.xyfxfy}

This is a finite sum since $f$ has finite support. By Exercise 1.3, the radius of convergence of $G(x,y|z)$ does not depend on $x$ and $y$. So, for every $\epsilon > 0$ there is some $N > 0$ such that for all $n > N$ and for every pair $x, y$ appearing in this finite sum, $p_{2n}(x,y)^{\frac{1}{2n}} < \rho + \epsilon$. Thus, by (7.3),

$$\begin{aligned}
\|P^n f\|^2 &< \sum_{x,y} f(x) f(y) \mathbf{1}_{\{f(x)f(y)>0\}}\, (\rho+\epsilon)^{2n} \pi(x) \\
&< C(\rho+\epsilon)^{2n}
\end{aligned}$$

for some constant $C < \infty$, since $f$ is finitely supported. Taking $2n$th roots yields $\lim_{n\to\infty} \|P^n f\|^{1/n} \leq \rho$, and we are done. □

{l.hilbertnorm

**Lemma 7.2.** *For a self-adjoint operator $P$ on a Hilbert space $H = \ell^2(V)$, $\|P\| := \sup_{f\in H} \frac{\|Pf\|}{\|f\|} = \sup_{f\in\ell_0(V)} \frac{(Pf,f)}{\|f\|^2}$.*

*Proof.* We give two proofs. The first uses "theory", the second uses a "trick". (Grothendieck had the program of doing all of math without tricks: the right abstract definitions should lead to solutions automatically. I think the problem with this is that, without understanding the tricks first, people would not find or even recognize the right definitions.)

In the finite-dimensional case (i.e., when $V$ is finite, hence $\ell^2(V) = \ell_0(V)$), since $P$ is self-adjoint, its eigenvalues are real, and then both $\sup_{f\in H} \frac{\|Pf\|}{\|f\|}$ and $\sup_{f\in\ell_0(V)} \frac{(Pf,f)}{\|f\|^2}$ are expressions for the largest eigenvalue of $P$. For the infinite dimensional case, one has to use the spectral theorem, and note that it is enough to consider the supremum over the dense subset $\ell_0(V)$. For the details, see, e.g., [Rud73, Theorem 12.25].

For the tricky proof, which I learnt from [LyPer16, Exercise 6.6], first notice that

$$\sup_{f\in\ell_0(V)} \frac{\|Pf\|}{\|f\|} \geq \sup_{f\in\ell_0(V)} \frac{|(Pf,f)|}{\|f\|^2} \geq \sup_{f\in\ell_0(V)} \frac{(Pf,f)}{\|f\|^2}\,,$$

where the first inequality follows from $(Pf, Pf)(f, f) \geq |(Pf, f)|^2$, which is just Cauchy-Schwarz.

For the other direction, if the rightmost supremum is $C$, then, using that $P$ is self-adjoint,

$$\begin{aligned}(Pf, g) &= \frac{(P(f+g), f+g) - (P(f-g), f-g)}{4} \\ &\leq C\frac{(f+g, f+g) + (f-g, f-g)}{4} = C\frac{(f,f) + (g,g)}{2}.\end{aligned}$$

Taking $g := Pf\|f\|/\|Pf\|$ yields $\|Pf\| \leq C\|f\|$. Again by the denseness of $\ell_0(V)$ in $H$, this finishes the proof.

WAIT. THIS WAS NOT A COMPLETE PROOF AT ALL. □

## 7.2 The infinite case: the Kesten-Cheeger-Dodziuk-Mohar theorem

{ss.kesten}

In this section, we are going to prove Kesten's characterization of the amenability of a group through the spectral radius of the simple random walk, Theorem 5.7. However, we want to do this in larger generality. The isoperimetric inequalities of Chapter 5 were defined not only for groups but also for locally finite graphs, moreover, we can naturally define them for electric networks, as follows. Satisfying the (edge) isoperimetric inequality $IP_\infty$ will mean that there exists a $\kappa > 0$ such that $C(\partial_E S) \geq \kappa\pi(S)$ for any finite connected subset $S$, where $\pi(S) = \sum_{x\in S} C_x$ is the natural stationary measure for the associated random walk, and $C(\partial_E S) = \sum_{\substack{x\in S \\ y\notin S}} c(x,y)$. The largest possible $\kappa$, i.e., $\iota_{\infty,E,c} := \inf_S C(\partial_E S)/\pi(S)$, is the Cheeger constant of the network. Note that if we are given a reversible Markov chain $(V, P)$, then the Cheeger contant of the associated electric network does not depend on which reversible measure $\pi$ we take, so we can talk about $\iota_{\infty,E}(P) = \kappa(P)$, the **Cheeger constant of the chain**.

The following theorem was proved for groups by Kesten in his PhD thesis [Kes59]. A differential geometric version was proved in [Che70]. Then [Dod84] and [Moh88] proved generalizations for graphs and reversible Markov chains, respectively. We will be proving Mohar's generalization, following [Woe00, Sections 4.A and 10.A].

{t.kesten}

**Theorem 7.3** (Kesten, Cheeger, Dodziuk, Mohar)**.** *Let $(V, P)$ be an infinite reversible Markov chain on the state space $V$, with Markov operator $P$. The following are equivalent:*

- **(1)** *$(V, P)$ satisfies $IP_\infty$ with $\kappa > 0$;*
- **(2)** *For all $f \in \ell_0(V)$ the* **Dirichlet inequality** *$\bar\kappa\|f\|_2^2 \leq \mathcal{E}_{P,\pi}(f)$ is satisfied for some $\bar\kappa > 0$, where $\mathcal{E}_{P,\pi}(f) = (\nabla f, \nabla f) = \frac{1}{2}\sum_{x,y} |f(x) - f(y)|^2 c(x,y)$ and $(f, f) = \sum_x f^2(x)C_x$;*
- **(3)** *$\rho(P) \leq 1 - \bar\kappa$, where the spectral radius satisfies $\rho(P) = \|P\|$, shown in the previous section.*

*In fact, $\kappa$ and $\bar\kappa$ are related by $\kappa(P)^2/2 \leq 1 - \rho(P) \leq \kappa(P)$.*

*Proof of (2) ⇔ (3).* Recall from (7.1) that $\mathcal{E}_{P,\pi}(f) = \|f\|_2^2 - (f, Pf)$. Now rearrange the Dirichlet inequality $\bar\kappa\|f\|_2^2 \leq \|f\|_2^2 - (f, Pf)$ into $(f, Pf) \leq (1-\bar\kappa)\|f\|_2^2$ for all $f \in \ell_0(V)$. By Lemma 7.2, this is true precisely when $\|P\| \leq 1 - \bar\kappa$. □

*Proof of (2) ⇒ (1).* Take any finite connected set $S$. Then $\pi(S) = \sum_{x\in S} C_x = \|\mathbf{1}_S\|_2^2$, and

$$\mathcal{E}_{P,\pi}(\mathbf{1}_S) = \frac{1}{2}\sum_{x,y} |\mathbf{1}_S(x) - \mathbf{1}_S(y)|^2 c(x,y) = C(\partial S).$$

Now apply the Dirichlet inequality to $\mathbf{1}_S$ to get $C(\partial S) = \mathcal{E}_{P,\pi}(\mathbf{1}_S) \geq \bar\kappa\|\mathbf{1}_S\|_2^2 = \bar\kappa\pi(S)$, and so $(V, P)$ satisfies $IP_\infty$ with $\kappa = \bar\kappa$. □

To show that the first statement implies the second, we want to show that the existence of "almost invariant" functions, or functions $f$ such that $(Pf, f)$ is close to $\|f\|^2$, gives the existence of "almost invariant" sets, i.e., sets $S$ with $P\mathbf{1}_S$ close to $\mathbf{1}_S$, which are exactly the Følner sets.

{d.sobolev}

**Definition 7.4.** *The* **Sobolev norm** *of $f$ is $S_{P,\pi}(f) := \frac{1}{2}\sum_{x,y} |f(x) - f(y)|c(x, y) = \|\nabla f\|_{\text{count},1}$, where, by $\|\cdot\|_{\text{count},1}$, we mean the $L^1$-norm w.r.t. counting measure on the edge set, as opposed to weighting by the resistances.*

{p.sobolevIP}

**Proposition 7.5.** *For any $d \in [1, \infty]$, a reversible chain $(V, P)$ satisfies $IP_d(\kappa)$ if and only if the* **Sobolev inequality** *$\kappa\|f\|_{\frac{d}{d-1}} \leq S_{P,\pi}(f)$ holds for all $f \in \ell_0(V)$, where $IP_d(\kappa)$ means $C(\partial_E S) \geq \kappa\pi(S)^{\frac{d-1}{d}}$ for any finite connected subset $S$, and $\|f\|_p = (\sum_x |f(x)|^p C_x)^{\frac{1}{p}}$.*

*Proof.* To prove that the Sobolev inequality implies good isoperimetry, note that $\|\mathbf{1}_S\|_{\frac{d}{d-1}} = \pi(S)^{\frac{d-1}{d}}$ and $S_{P,\pi}(\mathbf{1}_S) = C(\partial_E S)$.

To prove the other direction, first note that $S_{P,\pi}(f) \geq S_{P,\pi}(|f|)$ by the triangle inequality, so we may assume $f \geq 0$. For $t > 0$, we are going to look at the super-level sets $S_t = \{f > t\}$, which will be finite since $f \in \ell_0$. Now,

$$\begin{aligned}
S_{P,\pi}(f) &= \sum_x \sum_{y: f(y) > f(x)} (f(y) - f(x))c(x, y) \\
&= \sum_x \sum_{y: f(y) > f(x)} c(x, y) \int_0^\infty \mathbf{1}_{[f(x), f(y))}(t)dt \\
&= \int_0^\infty \sum_x \sum_{y: f(y) > f(x)} c(x, y)\mathbf{1}_{[f(x), f(y))}(t)dt \\
&= \int_0^\infty \sum_{\substack{x,y \text{ s.t.} \\ f(x) \leq t < f(y)}} c(x, y)dt \\
&= \int_0^\infty C(\partial_E\{f > t\})dt.
\end{aligned}$$

For $d = \infty$ ($\frac{d}{d-1} = 1$), we get

$$\begin{aligned}
S_{P,\pi}(f) &= \int_0^\infty C(\partial_E\{f > t\})dt \\
&\geq \kappa \int_0^\infty \pi(\{f > t\})dt \\
&= \kappa\, \mathbb{E}[f] = \kappa\, \|f\|_1.
\end{aligned}$$

For $d = 1$, since $\partial(\{f > t\}) \neq 0 \Leftrightarrow 0 < t \leq \|f\|_\infty$, we get that $S_{P,\pi}(f) \geq \int_0^{\|f\|_\infty} \kappa\, dt = \kappa\, \|f\|_\infty$.

The case $1 < d < \infty$ is just slightly more complicated than the $d = \infty$ case, and is left as an exercise. □

▷ **Exercise 7.2.** *Let $d \in (1, \infty)$, and let $p = \frac{d}{d-1}$.*

**(a)** *Show that $\int_0^\infty pt^{p-1}F(t)^p dt \leq (\int_0^\infty F(t)dt)^p$ for $F \geq 0$ decreasing.*

**(b)** *Using part (a), finish the proof of the proposition, i.e., the $1 < d < \infty$ case.*

We now use the proposition to complete the proof of the Kesten theorem.

*Proof of (1) ⇒ (2) of Theorem 7.3.*

$$\begin{aligned}
\|f\|_2^2 &= \|f^2\|_1 \\
&\le \frac{1}{\kappa} S_{P,\pi}(f^2) \quad \text{by the proposition} \\
&= \frac{1}{\kappa}\frac{1}{2}\sum_{x,y} |f^2(x) - f^2(y)|\, c(x,y) \\
&\le \frac{1}{\kappa}\frac{1}{2}\sum_{x,y} \big|f(x) - f(y)\big| \big(|f(x)| + |f(y)|\big)\, c(x,y) \\
&\le \frac{1}{\kappa}\bigg(\frac{1}{2}\sum_{x,y} |f(x) - f(y)|^2 c(x,y)\bigg)^{1/2} \bigg(\frac{1}{2}\sum_{x,y} \big(|f(x)| + |f(y)|\big)^2 c(x,y)\bigg)^{1/2},
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz.

The first sum above is precisely $\mathcal{E}_{P,\pi}(f)^{1/2}$. The second sum can be upper bounded by $\sqrt{2}\|f\|_2$, using the inequality $\frac{1}{2}(|x|+|y|)^2 \le x^2 + y^2$. So, after squaring the entire inequality, we have $\|f\|_2^4 \le \frac{2}{\kappa^2}\mathcal{E}_{P,\pi}(f)\|f\|_2^2$, which gives

$$\|f\|_2^2 \le \frac{2}{\kappa^2}\mathcal{E}_{P,\pi}(f).$$

Therefore, the first statement in the Kesten theorem implies the second, with $\bar{\kappa} = \frac{\kappa^2}{2}$. □

## 7.3 The finite case: expanders and mixing

{ss.mixing}

The decay of the return probabilities has a very important analogue for finite state Markov chains: convergence to the stationary distribution. In most chains, the random walk will gradually forget its starting distribution, i.e., it gets mixed, and the speed of this mixing is a central topic in probability theory and theoretical computer science, with applications to almost all branches of sciences and technology. (As I heard once from László Lovász, when I was still in high school: "With the right glasses, *everything* is a Markov chain.") A great introductory textbook to Markov chain mixing is [LevPW17]. The main topic in this section will be to understand mixing times from a spectral point of view; in the next chapter, isoperimetry will be the main tool.

Let $(V, P)$ be a finite state Markov chain, with its Markov operator $Pf(x) = \sum_y p(x,y)f(y) = \mathbf{E}\big[\, f(X_1) \,\big|\, X_0 = x \,\big]$ acting on $\ell_0(V)$. The constant 1 function is obviously an eigenfunction with eigenvalue 1, i.e., $P\mathbf{1} = \mathbf{1}$. On the other hand, recall that a probability measure $\pi$ on $V$ is called stationary if $\pi P = \pi$, i.e., if $\pi P(y) = \sum_x p(x,y)\pi(x) = \pi(y)$, i.e., if it is a left eigenfunction with eigenvalue 1. The existence of such a $\pi$ is much less obvious than the case of the above constant eigenfunction, but we are dealing in this section only with reversible chains, for which the reversible distribution is also stationary.

{ex.genspec}

▷ **Exercise 7.3.** *Let $(V, P)$ be a reversible, finite Markov chain, with stationary distribution $\pi(x)$. Recall that $P$ is self-adjoint with respect to $(f, g) = \sum_{x\in V} f(x)g(x)\pi(x)$. Show:*

- **(a)** *If $f : V \longrightarrow \mathbb{R}$ is a right eigenfunction of $P$, then $x \mapsto g(x) = f(x)\pi(x)$ is a left eigenfunction, with the same eigenvalue.*
- **(b)** *All eigenvalues $\lambda_i$ satisfy $-1 \le \lambda_i \le 1$;*
- **(c)** *If we write $-1 \le \lambda_n \le \cdots \le \lambda_1 = 1$, then $\lambda_2 < 1$ if and only if $(V, P)$ is connected (the chain is irreducible);*
- **(d)** *$\lambda_n > -1$ if and only if $(V, P)$ is not bipartite. (Recall here the easy lemma that a graph is bipartite if and only if all cycles are even.)*

Recall from the Kesten Theorem 7.3 that $\frac{\kappa^2}{2} \le 1-\rho \le \kappa$ for infinite reversible Markov chains. For finite chains, we have the following analogue, where $1-\lambda_2$ is usually called the **spectral gap**.

{t.AlonMilman}

**Theorem 7.6** (Alon, Milman, Dodziuk [Alo86, AloM85, Dod84])**.** *For a finite reversible Markov chain $(V,P)$ with stationary distribution $\pi$ and conductances $c(x,y)=\pi(x)p(x,y)$, set*

$$h(V,P) = \inf_{S \subseteq V} \frac{C(\partial_E S)}{\pi(S) \wedge \pi(S^c)}\,,$$

*the Cheeger constant of the finite chain. Then $\frac{h^2}{2} \le 1-\lambda_2 \le 2h$.*

*Sketch of proof.* The proof is almost identical to the infinite case, with two differences. One, the definition of the Cheeger constant is slightly different now, with the complement of the set also appearing in the denominator. Two, we have to deal with the fact that 1 is an eigenvalue, so we are interested in the spectral gap, not in the spectral radius. For this, recall that $\lambda_2 = \sup\{\frac{\|Pf\|}{\|f\|} : \sum_x f(x)\pi(x) = 0\}$ (Raleigh, Ritz, Courant, Fisher), since this is the subspace orthogonal to the constant functions, the eigenspace corresponding to the eigenvalue 1. By an obvious modification of Lemma 7.2 and by (7.1), this says that

$$1-\lambda_2 = \inf\left\{ \frac{\mathcal{E}(f)}{\|f\|^2} : \sum_x f(x)\pi(x) = 0 \right\}. \tag{7.4}$$ {e.gapDir}

Now, what are the effects of these modifications on the proofs?

When we have some isoperimetric constant $h(V,P)$ and want to bound the spectral gap from below, start with a function $f$ attaining the infimum in (7.4), and consider its super-level sets $\{x : f(x) > t\}$. From the argument of Proposition 7.5, for $f_+ := f \vee 0$, we get

$$S_{P,\pi}(f_+) \ge h \int_0^\infty \pi(\{f > t\}) \wedge \pi(\{f \le t\})\, dt\,.$$

By symmetry, we may assume that $\pi(\{f > 0\}) \le 1/2$, hence $\pi(\{f > t\}) \le \pi(\{f \le t\})$ for all $t \ge 0$, and we get $S(f_+) \ge h\,\|f_+\|_1$. Similarly, we get $S(f_+^2) \ge h\,\|f_+^2\|_1 = h\,\|f_+\|_2^2$. Now, just as in the proof of (1) $\Rightarrow$ (2) of Theorem 7.3, we have $S(f_+^2) \le \sqrt{2}\mathcal{E}(f_+)^{1/2}\|f_+\|_2$. These two inequalities combined,

$$\frac{h^2}{2} \le \frac{\mathcal{E}(f_+)}{\|f_+\|^2} \le \frac{\mathcal{E}(f)}{\|f\|^2} = 1-\lambda_2\,,$$

where the second inequality is left as a simple exercise.

For the other direction, if we have a subset $A$ with small boundary, then take $f = \mathbf{1}_A - \beta\mathbf{1}_{A^c}$ instead of $f = \mathbf{1}_A$, with a $\beta$ chosen to make the average 0, and compute the Dirichlet norm. □

▷ **Exercise 7.4.** *Fill in the missing details of the above proof.*

The main reason for the interest in the spectral gap of a chain is the following result, saying that if the gap is large, then the chain has a small **mixing time**: starting from any vertex or any initial distribution, after not very many steps, the distribution will be close to the stationary measure. The speed of convergence to stationarity is basically the finite analogue of the heat kernel decay in the infinite case, and the theorem is the finite analogue of the result that a spectral radius smaller than 1 implies exponential heat kernel decay (i.e., the quite obvious inequality $\rho(P) \le \|P\|$ in Proposition 7.1). The converse direction (does fast mixing require a uniformly positive spectral gap?) has some subtleties (pointed out to me by Ádám Timár), which will be discussed after the proof of the theorem, together with proper definitions of the different notions of mixing time.

{t.gapmix}

**Theorem 7.7.** *Let $(V, P, \pi)$ be a reversible, finite Markov chain. Set $g_{\mathrm{abs}} := 1 - \max_{i \geq 2} |\lambda_i|$, and $\pi_* := \min_{x \in V} \pi(x)$. Then for all $x$ and $y$, $|\frac{p_t(x,y) - \pi(y)}{\pi(y)}| \leq \frac{(1-g_{\mathrm{abs}})^t}{\pi_*}$, where $p_t(x,y)$ is the probability of being at $y$ after $t$ steps starting from $x$.*

The result is empty for a chain that is not irreducible: there $g_{\mathrm{abs}} = 0$, and the upper bound $1/\pi_*$ is trivial. Note also that we use $g_{\mathrm{abs}}$, called the **absolute spectral gap**, instead of $1 - \lambda_2$. The reason is that $\lambda_n$ being close to $-1$ is an obvious obstacle for mixing: say, if $\lambda_n = -1$, then the graph is bipartite (Exercise 7.3 (c)), and the distribution at time $t$ depends strongly on the parity of $t$.

Before proving the theorem, let us see four examples.

**Example 1.** For simple random walk on a complete $n$-vertex graph with loops, whose transition matrix

$$\begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

has eigenvalues $1, 0, \ldots, 0$, the bound in Theorem 7.7 is 0, that is, $p_t(x,y)$ is $1/n$ for every $t \geq 1$.

**Example 2.** For simple random walk on the $n$-cycle $C_n$, the exercise below tells us that the eigenvalues are $\cos(2\pi j/n)$, $j = 0, 1, \ldots, n-1$. If $n$ is even, $C_n$ is bipartite, and indeed $-1$ is an eigenvalue. If $n$ is odd, then the absolute spectral gap is $1 - \cos(2\pi/n) \asymp n^{-2}$. So, the upper bound in the theorem gets small at $Cn^2 \log n$. The true mixing time is actually $Cn^2$, as we will see in Exercises 7.12 (c), 7.15 (c), and in Section 8.2.

{ex.specCn}

▷ **Exercise 7.5.** *Verifying the above example, compute the spectrum of simple random walk on the $n$-cycle $C_n$. (Hint: think of the Markov operator as an average of two operators: rotating clockwise or counter-clockwise. The eigenvalues for the rotations are easy to calculate, with a basis of eigenfunctions that are actually the same for the two rotations, hence they are also eigenfunctions for the average.)*

**Example 3.** A random walk can be made "lazy" by flipping a fair coin before each step, and only moving if the coin turns up heads (otherwise, the walk stays in the same place for that step.) If $\{0,1\}^k$ is the $k$-dimensional hypercube, consider the simple random walk on this hypercube and let $P$ be its Markov operator. Then $\bar{P} = \frac{I+P}{2}$ is the Markov operator for the lazy random walk on the hypercube, and it has eigenvalues $\frac{1+\lambda_i}{2}$.

It turns out that $g_{\mathrm{abs}} = \frac{1}{k}$ in this case, and $\pi_* = \frac{1}{2^k}$. So, the upper bound in the theorem gets small for $t = Ck^2$, but the true mixing time is actually $C\, k \log k$; see Exercises 7.12 (c), 7.14, and Exercise 8.6.

{ex.hyperspec}

▷ **Exercise 7.6.** *Compute the spectrum of $\bar{P}$ on $\{0,1\}^k$ in the above example. (Hint: think of this as a product of Markov chains $P_i$, $i = 1, \ldots, k$: with probability $1/k$, move in the $i^{th}$ coordinate according to $P_i$. Then, if $f_{i,j}$ is an eigenfunction of $P_i$ with eigenvalue $\lambda_{i,j}$, what happens to the function $F(x_1, \ldots, x_k) := \prod_{i=1}^k f_{i,j}(x_i)$ when $\bar{P}$ is applied?)*

Our last example concerns graphs that are easy to define, but whose existence is far from clear. See the next section for more on this.

{ndcExpander}

**Definition 7.8.** *An $(n,d,c)$-**expander** is a $d$-regular graph on $n$ vertices, with $h(V) \geq c$. A $d$-regular **expander sequence** is a sequence of $(n_k, d, c)$-expanders with $n_k \to \infty$ and $c > 0$. By Theorem 7.6, the existence of this $c > 0$ is equivalent to having a uniformly positive spectral gap.*

**Example 4.** For simple random walk on an $(n, d, c)$ expander, $\pi$ is uniform, and $\pi_* = \frac{1}{n}$. For $t = C(d, c) \log n$, the bound in the theorem is small. This is sharp, up to a constant factor. In fact, note that any $d$-regular graph with $n$ vertices has diameter at least $\log_{d-1} n$, which implies that $\max_{x,y} |\frac{p_t(x,y)-\pi(y)}{\pi(y)}| = 1$ for $t \leq \log_{d-1} n$. Therefore, an expander mixes basically as fast as possible for a constant degree graph. (The converse is false, see Exercise 7.18.)

*Proof of Theorem 7.7.* Define $\varphi_x(z) := \frac{\mathbf{1}_{\{x=z\}}}{\pi(x)}$. Then $(P^t\varphi_y)(x) = \frac{\mathbf{P}[\,X_t=y|X_0=x\,]}{\pi(y)} = \frac{p_t(x,y)}{\pi(y)}$. So,

$$\begin{aligned}\frac{|p_t(x,y)-\pi(y)|}{\pi(y)} &= |(P^t\varphi_y - \mathbf{1}, \varphi_x)| = |(P^t(\varphi_y - \mathbf{1}), \varphi_x)| \\ &\leq \|\varphi_x\| \cdot \|P^t(\varphi_y - \mathbf{1})\| = \frac{1}{\sqrt{\pi(x)}}\|P^t(\varphi_y - \mathbf{1})\|\,,\end{aligned}$$

where the inequality is from Cauchy-Schwarz. Now, since $(\varphi_y - \mathbf{1}, \mathbf{1}) = 0$, we can write $\varphi_y - \mathbf{1} = \sum_{i=2}^n a_i f_i$, where $f_1, \dots, f_n$ is an orthonormal basis of eigenvectors. Thus,

$$\begin{aligned}\|P^t(\varphi_y - \mathbf{1})\|^2 &= \|\sum_{i=2}^n a_i\lambda_i^t f_i\|^2 = \sum_{i=2}^n |\lambda_i^t|^2 \|a_i f_i\|^2 \\ &\leq \lambda_*^{2t} \sum_{i=2}^n \|a_i f_i\|^2 = \lambda_*^{2t}\|\varphi_y - \mathbf{1}\|^2 \leq \lambda_*^{2t}\|\varphi_y\|^2,\end{aligned}$$

where $\lambda_* := \max_{i\geq 2} |\lambda_i|$, and the last inequality is by the Pythagorean theorem. So $|\frac{p_t(x,y)-\pi(y)}{\pi(y)}| \leq \lambda_*^t \frac{1}{\sqrt{\pi(x)\pi(y)}}$, and we are done. □

This theorem measures closeness to stationarity in a very strong sense, in the $L^\infty$-norm between the functions $p_t(x, \cdot)/\pi(\cdot)$ and $\mathbf{1}(\cdot)$, called the **uniform distance**. The proof itself used, as an intermediate step, the $L^2$-distance

$$\|P^t\varphi_y - \mathbf{1}\|_2 = \left\|\frac{p_t(\cdot, y)}{\pi(y)} - \mathbf{1}(\cdot)\right\|_2 = \left\|\frac{p_t(y, \cdot)}{\pi(\cdot)} - \mathbf{1}(\cdot)\right\|_2 = \chi^2\big(p_t(y,\cdot), \pi(\cdot)\big)\,,$$

where the middle equality uses that $p_t(x, y)/\pi(y) = p_t(y, x)/\pi(x)$ for reversible chains, and $\chi^2$ is the **chi-square distance**, the following asymmetric distance between two measures:

$$\chi^2(\mu, \nu) := \sum_{x\in V} \left|\frac{\mu(x)}{\nu(x)} - 1\right|^2 \nu(x)\,, \tag{7.5}$$ {e.chi2dist}

which we will again use in Section 8.3. The most popular notion of distance uses the $L^1$-norm, or more precisely the **total variation distance**, defined as follows: for any two probability measures $\mu$ and $\nu$ on $V$,

$$\begin{aligned}d_{\mathrm{TV}}(\mu, \nu) :&= \max\big\{|\mu(A) - \nu(A)| : A \subseteq V\big\} \\ &= \frac{1}{2}\sum_{x\in V} |\mu(x) - \nu(x)| = \frac{1}{2}\sum_{x\in V}\left|\frac{\mu(x)}{\nu(x)} - 1\right|\nu(x)\,.\end{aligned} \tag{7.6}$$ {e.TV}

{ex.TVminA}

▷ **Exercise 7.7.** *Prove the equality between the two lines of (7.6). This also shows $d_{\mathrm{TV}}$ to be a metric. (Hint: show first that $A := \{x \in V : \mu(x) \geq \nu(x)\}$ maximizes $|\mu(A) - \nu(A)|$.)*

{ex.TVcoupling

▷ **Exercise 7.8.** *Show that $d_{\mathrm{TV}}(\mu, \nu) = \min\big\{\mathbf{P}[\,X \neq Y\,] : (X, Y)$ is a coupling of $\mu$ and $\nu\big\}$.*

Of course, one can use any $L^p$ norm, or quantities related to the entropy, and so on, see [SaC97, BobT06]. In any case, given a notion of distance, the **mixing time** of a chain is usually defined to be the smallest time $t$ when the distance between $p_t(x,\cdot)$ and $\pi(\cdot)$ becomes less than some small constant, say 1/4. This time will be denoted by $t_{\mathrm{mix}}(1/4)$. Why is this is a good definition? Let us discuss the case of the total variation distance. Let

$$d(t) := \sup_{x\in V} d_{\mathrm{TV}}\big(p_t(x,\cdot),\pi(\cdot)\big) \quad \text{and} \quad \bar{d}(t) := \sup_{x,y\in V} d_{\mathrm{TV}}\big(p_t(x,\cdot),p_t(y,\cdot)\big)\,.$$

The following two exercises explain why we introduced $\bar{d}(t)$.

▷ **Exercise 7.9.** *Show that $d(t) \le \bar{d}(t) \le 2d(t)$.*

▷ **Exercise 7.10.** *Using Exercise 7.8, show that $\bar{d}(t+s) \le \bar{d}(t)\,\bar{d}(s)$.*

Therefore, for $t_{\mathrm{mix}}(1/4) = t^{\mathrm{TV}}_{\mathrm{mix}}(1/4)$,

$$d(\ell\, t_{\mathrm{mix}}(1/4)) \le \bar{d}(\ell\, t_{\mathrm{mix}}(1/4)) \le \bar{d}(t_{\mathrm{mix}}(1/4))^{\ell} \le (2d(t_{\mathrm{mix}}(1/4))^{\ell} \le 2^{-\ell}\,.$$

For instance, $t_{\mathrm{mix}}(2^{-100}) \le 100\, t_{\mathrm{mix}}(1/4)$, so $t_{\mathrm{mix}}(1/4)$ captures well the magnitude of the time needed for the chain to get close to stationarity.

Define the **separation distance** at time $t$ by $s(t) := \sup_{x\in V}\Big(1-\frac{p_t(x,y)}{\pi(y)}\Big)$. Note that this is a one-sided version of the uniform distance: $s(t) < \epsilon$ means that all states have collected at least a $(1-\epsilon)$-factor of their share at stationarity, but there still could be some states where the walk is very likely to be, compared to the stationary measure of those states. Since the total measure of this singularly favoured set is at most $\epsilon$, the separation distance is a natural and useful measure of mixing.

{ex.separation

▷ **Exercise 7.11.** *Show that, in any finite reversible chain, $d(t) \le s(t) \le 4d(t/2)$. (Hint: since $t$ and $t/2$ appear together, it seems like a good idea to write a trajectory of length $t$ as $t/2 + t/2$ and reverse time for the second leg. Then, using $\sqrt{p_{t/2}(x,z)p_{t/2}(y,z)} \ge p_{t/2}(x,z) \wedge p_{t/2}(y,z)$, we can get to a formula that resembles $d(t/2)$ by Exercise 7.7.)*

The **relaxation time** of a reversible chain is defined to be $t_{\mathrm{relax}} := 1/g_{\mathrm{abs}}$, where $g_{\mathrm{abs}}$ is the absolute spectral gap. This is certainly related to mixing: Theorem 7.7 implies that

$$t^{\infty}_{\mathrm{mix}}(1/e) \le \big(1+\ln\frac{1}{\pi_*}\big)\, t_{\mathrm{relax}}\,. \tag{7.7}$$ {e.inftyrelax}

It also has an interpretation as a genuine temporal quantity, shown by the following exercise that is just little more than a reformulation of the proof of Theorem 7.7:

{ex.L2mixing}

▷ **Exercise 7.12.** *Let $(V,P,\pi)$ be a reversible chain with eigenvalues $1=\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_n \ge -1$ and corresponding eigenfunctions $f_1,\dots,f_n$.*

**(a)** *For $f: V \longrightarrow \mathbb{R}$, let $\mathrm{Var}_\pi[f] := \mathbf{E}_\pi[f^2] - (\mathbf{E}_\pi f)^2 = \sum_x f(x)^2\pi(x) - \big(\sum_x f(x)\pi(x)\big)^2$. Show that $g_{\mathrm{abs}} > 0$ implies that $\lim_{t\to\infty} P^t f(x) = \mathbf{E}_\pi f$ for all $x \in V$. Moreover,*

$$\mathrm{Var}_\pi[P^t f] \le (1-g_{\mathrm{abs}})^{2t}\,\mathrm{Var}_\pi[f]\,,$$

*with equality at the eigenfunction corresponding to the $\lambda_i$ giving $g_{\mathrm{abs}} = 1-|\lambda_i|$. Hence $t_{\mathrm{relax}}$ is the time needed to reduce the standard deviation of any function to $1/e$ of its original standard deviation.*

**(b)** *Show that*

$$4\, d_{\mathrm{TV}}\big(p_t(x,\cdot),\pi(\cdot)\big)^2 \le \left\| \frac{p_t(x,\cdot)}{\pi(\cdot)} - \mathbf{1}(\cdot) \right\|_2^2 = \sum_{i=2}^n f_i(x)^2 \lambda_i^{2t}\,.$$

**(c)** *When the chain $(V,P)$ is transitive, then, for any $x \in V$,*

$$4\, d_{\mathrm{TV}}\big(p_t(x,\cdot),\pi(\cdot)\big)^2 \le \left\| \frac{p_t(x,\cdot)}{\pi(\cdot)} - \mathbf{1}(\cdot) \right\|_2^2 = \sum_{i=2}^n \lambda_i^{2t}\,.$$

*For instance, assuming the answer to Exercise 7.6, one can easily get the bound $d\big(1/2\,k\ln k + c\,k\big) \le e^{-2c}/2$ for $c > 1$ on the TV distance for the lazy walk on the hypercube $\{0,1\}^k$. (This is sharp even regarding the constant 1/2 in front of $k \ln k$ — see Exercise 7.14 below.) Also, assuming the answer to Exercise 7.5, one can get that $t_{\mathrm{mix}}^{\mathrm{TV}}(C_n) = O(n^2)$ for the $n$-cycle.*

Furthermore, the finite chain analogue of the proof of the $\|P\| \le \rho(P)$ inequality in Proposition 7.1 gives the following:

{pr.convspeed}

**Proposition 7.9.** *For the distance in total variation,* $\lim_{t\to\infty} d(t)^{1/t} = 1 - g_{\mathrm{abs}}$.

*Proof.* The direction $\le$ follows almost immediately from Theorem 7.7. For the other direction, let $f$ be an eigenfunction corresponding to the $\lambda_i$ giving $g = 1 - |\lambda_i|$. This is orthogonal to the constant functions (the eigenfunctions for $\lambda_1 = 1$), hence $\sum_x f(x)\pi(x) = 0$. Therefore,

$$|\lambda_i^t f(x)| = |P^t f(x)| = \Big| \sum_y \big(p_t(x,y) - \pi(y)\big) f(y) \Big| \le \|f\|_\infty\, 2\, d_{\mathrm{TV}}\big(p_t(x,\cdot),\pi(\cdot)\big) \le 2\,\|f\|_\infty\, d(t)\,.$$

Taking $x \in V$ with $f(x) = \|f\|_\infty$, we get $(1-g_{\mathrm{abs}})^t \le 2d(t)$, then taking $t^{\mathrm{th}}$ roots gives the result. □

The above inequality $(1-g_{\mathrm{abs}})^t \le 2d(t)$ easily gives that $g_{\mathrm{abs}} \ge c/t_{\mathrm{mix}}^{\mathrm{TV}}(1/4)$, or in other words,

$$t_{\mathrm{relax}} \le C\, t_{\mathrm{mix}}^{\mathrm{TV}}(1/4)\,, \tag{7.8}$$ {e.relaxTV}

for some absolute constants $0 < c, C < \infty$.

Besides (7.8), it is clear that $t_{\mathrm{mix}}^{\mathrm{TV}}(\epsilon) \le t_{\mathrm{mix}}^{\infty}(\epsilon)$ for any $\epsilon$. It is therefore more than natural to ask: going from the relaxation time to the uniform mixing time, how bad is the $\ln 1/\pi_*$ factor in (7.7), and when and how could we eliminate that? In some cases, this factor is not very important: e.g., in the example of simple random walk on the cycle $C_n$, it gives a bound $O(n^2 \log n)$, instead of the true mixing time $\asymp n^2$. However, for rapidly mixing chains, where $t_{\mathrm{relax}} = O(\log |V|)$, this is a serious factor. For the case of SRW on an expander, it is certainly essential, and (7.7) is basically sharp. In some other cases, this factor will turn out to be a big loss: e.g., for SRW on the hypercube $\{0,1\}^k$, or on a lamplighter group, say, on $\mathbb{Z}_2 \wr \mathbb{Z}_n^d$, or for chains on spin systems, such as the Glauber dynamics of the Ising model in a box $\mathbb{Z}_n^d$ at critical or high temperature. In these cases, $|V_n|$ is exponential in the parameter $n$ that is the dimension of the hypercube or the linear size of the base box $\mathbb{Z}_n^d$, while we still would like to have polynomial mixing in $n$, and the factor $\ln 1/\pi_*$ might ruin the exponent of the polynomial. Looking at the proof of Theorem 7.7, this factor comes from two sources. The first one is an application of Cauchy-Schwarz that looks indeed awfully wasteful: it seems very unlikely that $P^t\varphi_y - \mathbf{1}$ is almost collinear with $\varphi_x$, it seems even more unlikely that this happens for a lot of $y$'s and a fixed $x$, and it is certainly impossible for a fixed $y$ and several $x$'s, since different $\varphi_x$'s are orthogonal. This suggests that if we do not want mixing in $L^\infty$, only in some average, then we should be able to avoid this wasteful Cauchy-Schwarz. The second source is the possibility that all the non-unit eigenvalues are close to $\lambda_2$. To exclude this, one needs to understand the spectrum very well. The above Exercise 7.12 (c) gave two examples where both sources of waste can be eliminated.

In the case of the hypercube $\{0,1\}^k$, mixing times can be calculated very precisely, without using the spectrum, by the following direct approach. The first idea is that the steps can be thought of as rerandomizing a uniformly chosen coordinate, and hence, at the random stopping time $\tau_{\text{all}}$ when we have rerandomized all the coordinates, we are exactly at stationarity. This is exactly the classical coupon-collector problem: if there are $k$ different types of coupons, and in each step we obtain a uniform random one, how much time does it take to collect one of each type? In expectation, this takes time $1 + k/(k-1) + \cdots + k/1 \sim k \ln k$, and, because of the independence of stages in collecting the coupons, one clearly expects sharp concentration around this expectation. Therefore, at time $(1+o(1))\, k \log k$, we must be very close to stationarity in a strong sense:

{ex.coupconc}

▷ **Exercise 7.13.** *By proving concentration for the coupon collecting problem, show that the* $1/2$*-lazy random walk on* $\{0,1\}^k$ *has separation distance* $s(k \ln k + tk) \le Ce^{-ct}$ *for some* $0 < c, C < \infty$ *and all* $t > 0$*.*

This chain turns out to have a **separation distance cutoff** at the mixing time $\tau_{\text{mix}}^{\text{sep}} = (1+o(1))k \log k$:

$$\forall \epsilon > 0 : \qquad \frac{\min\{t : s(t) < \epsilon\}}{\min\{t : s(t) < 1-\epsilon\}} \to 1, \text{ as } k \to \infty\,. \tag{7.9}$$ {e.cutoff}

The reason is that $\tau_{\text{all}}$ is a so-called **strong stationary stopping time** (i.e., even the conditional distribution $X_{\tau_{\text{all}}} \,\big|\, \{\tau < t\}$ is stationary, for any $t > 0$) that has a **halting state**: there is a vertex such that being there ensures that the stopping time has already occurred (starting from all-0, the all-1 vertex is such). For such stopping times, the separation distance starting from $x$ satisfies $s_x(t) = \mathbf{P}_x[\,\tau > t\,]$; see [LevPW17, Remark 6.12]. Thus, the concentration of $\tau_{\text{all}}$ from below gives the lower bound for the separation cutoff on the hypercube.

However, still for the 1/2-lazy walk on the hypercube, to be close to stationarity in total variation distance, we do not need to rerandomize every single coordinate! Starting, say, from the all-0 vertex, TV-mixing should occur when enough coordinates have been rerandomized so that the few remaining 0's disappear in the fluctuations of the rerandomized coordinates. The stationary fluctuations in the number of 0 coordinates is of order $\sqrt{k}$, hence the coupon collecting should be okay to stop when fewer than $\sqrt{k}$ coordinates are missing. The following exercise makes this intuition precise.

{ex.hyperTV}

▷ **Exercise 7.14.** *The goal here is to show that the total variation mixing time of the* $1/2$*-lazy random walk* $X_0, X_1, \dots$ *on the hypercube* $\{0,1\}^k$ *is* $(1/2 + o(1))\, k \log k$*.*

**(a)** *Let* $Y_t$ *be the number of missing coupons at time* $t$ *in the coupon collector's problem with* $k$ *coupons. For* $\alpha < 1$*, show that*

$$\mathbf{E}\, Y_{\alpha\, k \log k} \sim k^{1-\alpha} \quad \textit{and} \quad \mathbb{D}\, Y_{\alpha\, k \log k} = o(k^{1-\alpha}).$$

*Using Markov's and Chebyshev's inequalities, deduce that* $Y_{\alpha\, k \log k}/\sqrt{k} \to 0$ *or* $\infty$ *in probability, for* $\alpha > 1/2$ *and* $< 1/2$*, respectively.*

**(b)** *Let* $\mathsf{N}(\mu, \sigma^2)$ *denote the normal distribution. Show that, for any sequence* $\sigma_k \to \sigma \in (0,\infty)$*, we have that* $d_{\text{TV}}\big(\mathsf{N}(0,\sigma^2),\, \mathsf{N}(\mu_k, \sigma_k^2)\big) \to 0$ *or* $1$*, for* $\mu_k \to 0$ *and* $\mu_k \to \infty$*, respectively. Using this and the local version of the de Moivre–Laplace theorem, prove that*

$$d_{\text{TV}}\big(\mathsf{Binom}(k,1/2),\, \mathsf{Binom}(k-k^\beta, 1/2) + k^\beta\big) \to \begin{cases} 0 & \text{if } \beta < 1/2\,, \\ 1 & \text{if } \beta > 1/2\,. \end{cases}$$

**(c)** *For $X_0 = (0, 0, \dots, 0) \in \{0,1\}^k$, let the distribution of $X_t$ be $\pi_t$. What is it, conditioned on $\|X_t\|_1 = \ell$?*

**(d)** *Let $Y_t$ be the number of coordinates that have not been rerandomized by time $t$. Compare the distribution of $k - \|X_t\|_1$, conditioned on $Y_t \geq y$, to $\mathsf{Binom}(k - y, 1/2) + y$. And what is the distribution of $\|Z\|_1$, where $Z$ has distribution $\pi$, uniform on $\{0,1\}^k$?*

**(e)** *Deduce from the previous parts that $d_{\mathrm{TV}}\big(\pi_{\alpha\, k \log k}, \pi\big) \to 0$ or $1$, for $\alpha > 1/2$ and $< 1/2$, respectively.*

In particular, we have a **cutoff phenomenon** again, as in (7.9), but now in total variation, at half the time. (So, the factor $1/2$ between the times in Exercise 7.11 was sharp). See [LevPW17, Chapter 18] for more on cutoff.

A general method to bound total variation mixing times from above, very similar to stationary stopping times, is by Exercise 7.8: for any coupling of the two chains $x = X_0, X_1, \dots$ and $y = Y_0, Y_1, \dots$ on $V(G)$,

$$d_{\mathrm{TV}}\big(p_t(x,\cdot), p_t(y,\cdot)\big) \leq \mathbf{P}[\, X_t \neq Y_t \mid X_0 = x,\, Y_0 = y\,]\,.$$

If we choose $y = Y_0$ according to the stationary measure, then $Y_t$ is also stationary, hence

$$d(t) \leq \max_x \mathbf{P}[\, X_t \neq Y_t \mid X_0 = x,\, Y_0 \sim \pi\,]\,.$$

As an example, a simple natural coupling for the lazy walk on the hypercube $\{0,1\}^k$ comes from our arguments above: we first choose a coordinate $i \in \{1, 2, \dots, k\}$ uniformly, then, if $X_n(i) \neq Y_n(i)$, then either $X_n$ or $Y_n$ moves in the $i$th coordinate, each with probability half. If $X_n(i) = Y_n(i)$, then the two walks move or stay put together. In this coupling, we reach $X_n = Y_n$ when all the $h$ coordinates in which $X_0$ and $Y_0$ differed have already been sampled, so we get the coupon-collector problem again, with expected coupling time $k/h + k/(h-1) + \dots + k/1 \sim k \ln h$ if both $k$ and $h$ are large, with a sharp concentration as in Exercise 7.13. In the worst case $h = k$, this gives a coupling time $(1 + o(1))\, k \ln k$, but even when $y = Y_0$ is chosen uniformly, which makes $h$ concentrated around $k/2$, we get the same asymptotics. That is, for total variation mixing, this method is off here by a factor of 2, as shown by Exercise 7.14.

To get the sharp TV-mixing on the hypercube, coupling can also be used, but in a somewhat indirect way. First, part (c) of Exercise 7.14 shows that the walk $X_t$ on $\{0,1\}^k$ is close to stationarity iff the Hamming weight $\|X_t\|_1$ is close to its stationary distribution $\mathsf{Binom}(k, 1/2)$. But the projected walk $\|X_t\|_1$ is clearly a Markov chain again, called the **lazy Ehrenfest urn**, whose TV-mixing time can be found precisely, using the simplest possible coupling: with probability $1/2$ one chain moves, with probability $1/2$ the other. See [LevPW17, Theorem 18.3] for an analysis of the coupling time. However, it is not hard to see that this coupling cannot be pulled back to a coupling of the walks on the hypercube.

In order to get good lower bounds on mixing times, it would of course be nice to know all the eigenvalues and eigenfunctions. A less algebraic hence more widely applicable idea is to find a natural observable that shows efficiently if we are close to stationarity. In our above discussion for the hypercube, this was the Hamming weight. The following exercise gives two ways in which such observables can be useful.

{ex.mixlower}

▷ **Exercise 7.15** (Lower bounds from eigenfunctions and similar observables).

**(a)** *For the lazy random walk on $\{0,1\}^k$, consider the total Hamming weight $W(\underline{x}) := \sum_{i=1}^k x_i$. Estimate its Dirichlet energy $\mathcal{E}(W)$ and variance $\mathrm{Var}_\pi[W] = \|W - \mathbf{E}_\pi W\|^2$, and deduce by the variational formula (7.4) that the spectral gap of the walk is at most of order $1/k$ (which we already knew from the exact computation in Exercise 7.6), i.e., the relaxation time is at least of order $k$. Then find the $L^2$-decomposition of $W$ into the eigenfunctions of the Markov operator, and deduce the decay of $\mathrm{Var}_\pi[P^t W]$ as a function of $t$.*

**(b)** *Take SRW on the $n$-cycle $C_n$. Find a function that shows by (7.4) that the spectral gap of the chain is a most of order $1/n^2$. (Again, we already knew this from computing the spectrum exactly in Exercise 7.5, but we want to point out here that finding one good observable is typically much easier than determining the entire spectrum.)*

**(c)** *For SRW on the $n$-cycle $C_n$, consider the function $F(i) := \mathbf{1}\{n/4 \le i < 3n/4\}$ for $i \in C_n$. Observe that although this $F$ gives only an $\Omega(n)$ lower bound on the relaxation time if the variational formula (7.4) is used, the variance $\mathrm{Var}_\pi[P^t F]$ does not start decaying before $t$ reaches order $n^2$, hence Exercise 7.12 (a) does imply a quadratic lower bound on the relaxation time. (That is, the same observable might fail or succeed, depending on the method.) Explain how this is possible in terms of the $L^2$-decomposition of $F$ into eigenfunctions of the Markov operator.*

The following theorem, often called "Wilson's method for lower bounds on mixing" and stated here without a proof, is a very useful generalization of the idea how the Hamming weight gave us a sharp lower bound for the TV mixing time in Exercise 7.14. See [LevPW17, Theorem 13.28] for a proof.

{t.WilsonLower

**Theorem 7.10** (Wilson's method [Wil04])**.** *Consider an irreducible aperiodic Markov chain $(X_t)_{t\ge 0}$ on a finite state space $V$. Let $\Phi$ be an eigenfunction with real eigenvalue $1/2 < \lambda < 1$, satisfying*

$$\mathbf{E}_x\big[\,|\Phi(X_1) - \Phi(x)|^2\,\big] \le R$$

*for some $R > 0$, for all starting points $X_0 = x$. Then, for any $\epsilon \in (0,1)$ and any $x \in V$,*

$$t_{\mathrm{mix}}^{\mathrm{TV}}(\epsilon) \ge \frac{1}{2\log(1/\lambda)}\left(\log\frac{(1-\lambda)\,\Phi(x)^2}{2R} + \log\frac{1-\epsilon}{\epsilon}\right).$$

For instance, for the 1/2-lazy random walk on $\{0,1\}^k$, for $\Phi(x)$ being the Hamming weight $W(x)$, we have $\lambda = 1-1/k$, $R = 1$, and $\max_x \Phi(x) = k$, hence indeed get $t_{\mathrm{mix}}^{\mathrm{TV}}(\epsilon) \ge (1/2+o(1))\,k\ln k$, for any $\epsilon \in (0,1)$ fixed. A more advanced example is the **Interchange Process** on the symmetric group $S_n$ given by the path $[1,n]$, i.e., the random walk on $S_n$ generated by picking uniformly at random one of the adjacent transpositions $(i, i+1)$ in each step, $1 \le i \le n-1$. Thinking of this process as permuting $n$ particles, each particle is performing a very lazy random walk on $[1,n]$, with one step in each $\asymp n$ steps of the chain. Hence each particle needs time $\asymp n^3$ to get mixed on $[1,n]$, and to get most of them mixed, similarly to the hypercube example, we will need time $\asymp n^3 \log n$. An exact lower bound, via Theorem 7.10, is again given by $\Phi$ being a sum of eigenfunctions, each with an eigenvalue close to 1. Namely, the movement of the $i$th particle is a projection of the full Markov chain, again Markovian, with second eigenfunction $\phi(k) = \cos\frac{(2k-1)\pi}{2n}$, and then $\Phi(\sigma) := \sum_{i=1}^n \phi(i)\,\phi(\sigma(i))$ is an eigenfunction of the full chain, with eigenvalue $\lambda = 1 - \frac{\pi^2}{n^3} + O(n^{-4})$. The coefficients $\phi(i)$ are not constant now (as opposed to the hypercube example) to make the resulting eigenfunction non-trivial, and they are chosen so that $\max_\sigma \Phi(\sigma)$ and $R$ are easy to compute and are useful: the first is quite large while the second is quite small. Indeed, with a little bit of work it turns out that $\Phi(\mathrm{id}) \asymp n$ and $R \asymp 1/n^2$, and hence we get a lower bound $t_{\mathrm{mix}}^{\mathrm{TV}}(\epsilon) \ge \frac{1+o(1)}{\pi^2} n^3 \ln n$. This was proved to be sharp in [Lac16].

▷ **Exercise 7.16.** *Fill in the details of the above calculation for the Interchange Process on $[1,n]$.*

We have seen that the TV mixing time on the hypercube is significantly larger than the relaxation time, and showed the cutoff phenomenon in Exercise 7.14. Exercise 7.15 showed for the cycle $C_n$ that $n^2 \preccurlyeq t_{\mathrm{relax}} \preccurlyeq t_{\mathrm{mix}}^{\mathrm{TV}}$, which is the right order of magnitude also from above, by Exercise 7.12 (c). But we did not talk about cutoff, which was not an accident:

▷ **Exercise 7.17.** *Show that if $t^{\mathrm{TV}}_{\mathrm{mix}}(V_n) \asymp t_{\mathrm{relax}}(V_n)$ for a sequence of finite reversible Markov chains on $n$ vertices, then the sequence cannot exhibit cutoff for the total variation mixing time.*

There are natural chains where the three mixing times, $t_{\mathrm{relax}} \preccurlyeq t^{\mathrm{TV}}_{\mathrm{mix}} \preccurlyeq t^{\infty}_{\mathrm{mix}}$, are in fact of different orders: e.g., simple random walk on the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}_n^d$, with a natural set of generators. For $d = 2$,

$$t_{\mathrm{relax}} \asymp n^2 \log n\,, \qquad t^{\mathrm{TV}}_{\mathrm{mix}}(\epsilon) \asymp n^2 \log^2 n\,, \qquad t^{\infty}_{\mathrm{mix}}(\epsilon) \asymp n^4\,. \tag{7.10}$$ {e.LLmix}

The reason is, very briefly, that the relaxation time is given by the maximal hitting time for SRW on the base graph torus $\mathbb{Z}_n^2$, the TV mixing time is given by the expected cover time of the base graph, while the uniform mixing time is given by the time when the set $S_t$ of unvisited sites on the base graph satisfies $\mathbf{E}[\,2^{|S_t|}\,] < 1+\epsilon$. See [PerR04a]. It is especially interesting to notice the difference between the separation and uniform mixing times: by time $C\,n^2 \log^2 n$ with large $C$, the small pieces left out by the SRW on the torus are not enough to make the $p_t(x,\cdot)$-measure of any specific state too small, but they still allow for certain states to carry very high measures.

In this example, (7.10), the relaxation and the mixing times are on the same polynomial order of magnitude; their ratio is just $\log n$, which is $\log\log |V_n|$. This seems to be a rather general phenomenon; e.g., I do not know examples of Markov chains over spin systems where this does not hold. However, I cannot formulate a general result or even conjecture on this, because of counterexamples like expanders, or the ones in Exercise 7.18 (b).

Let us now discuss whether a uniformly positive spectral gap is required for fast mixing. This depends on the exact definition of "fast mixing". First of all, Proposition 7.9 implies that in a sequence of finite chains, we have convergence to stationarity with a uniformly exponential speed if and only if the absolute spectral gaps are uniformly positive. On the other hand, if we define "fast mixing" by the mixing time being small, i.e., $t^{\mathrm{TV}}_{\mathrm{mix}}(1/4) = O(\log |V|)$, then (7.8) does not imply a uniformly positive absolute spectral gap. Indeed, that does not hold in general:

{ex.qexpander}

▷ **Exercise 7.18.** *You may accept here that transitive expanders exist.*

**(a)** *Give a sequence of $d$-regular transitive graphs $G_n = (V_n, E_n)$ with $|V_n| \to \infty$ that mix rapidly, $t^{\mathrm{TV}}_{\mathrm{mix}}(1/4) = O(\log |V_n|)$, but do not form an expander sequence.*

**(b)** *In a similar manner, give a sequence $G_n = (V_n, E_n)$ satisfying $t_{\mathrm{relax}} \asymp t^{\mathrm{TV}}_{\mathrm{mix}}(1/4)^{\alpha} \asymp \log^{\alpha} |V_n|$, with some $0 < \alpha < 1$.*

A nice property of simple random walk on an expander is that the consecutive steps behave similarly in several ways to an i.i.d. sequence, but the walk can be generated using much less randomness. This is useful in pseudorandom generators, derandomization of algorithms, etc. To conclude this section, here is an exact statement of this sort:

{pr.AKSz}

**Proposition 7.11.** *Let $(V, P)$ be a reversible Markov chain with stationary measure $\pi$ and spectral gap $1 - \lambda_2 > 0$, and let $A \subset V$ have stationary measure at most $\beta < 1$. Let $(X_i)_{i=0}^{\infty}$ be the trajectory of the chain in stationarity (i.e., $X_0 \sim \pi$). Then*

$$\mathbf{P}\big[\, X_i \in A \text{ for all } i = 0, 1, \dots, t\,\big] \le C(1-\gamma)^t\,,$$

*where $\gamma = \gamma(\lambda_2, \beta) > 0$ and $C$ is an absolute constant.*

This was first proved by Ajtai, Komlós and Szemerédi [AjKSz87]; see [HooLW06, Section 3] for a proof and applications. I am giving here what I think is the simplest possible proof. Stronger and extended versions of this argument can be found in [AloS00, Section 9.2] and [HamMP12]. We will use the following simple observation:

{ex.suppcontr}

▷ **Exercise 7.19.** *If $(V, P)$ is a reversible Markov chain with stationary measure $\pi$ and spectral gap $1-\lambda_2 > 0$, and $f \in L^2(V)$ has the property that $\pi(\operatorname{supp} f) \le 1 - \epsilon$, then*

$$(Pf, f) \le (1-\delta_1)(f,f) \quad \text{and} \quad (Pf, Pf) \le (1-\delta_2)(f,f)$$

*for some $\delta_i = \delta_i(\lambda_2, \epsilon) > 0$.*

*Proof of Proposition 7.11.* We want to rewrite the exit probability in question in a functional analytic language, since we want to use the notion of spectral gap. So, consider the projection $Q : f \mapsto f\mathbf{1}_A$ for $f \in L^2(V)$, and note that

$$\begin{aligned}
\mathbf{P}\big[\, X_i \in A \text{ for } i = 0, 1, \dots, 2t+1 \,\big] &= \big(Q(PQ)^{2t+1}\mathbf{1}, \mathbf{1}\big) \\
&= \big(P(QP)^t Q\mathbf{1}, (QP)^t Q\mathbf{1}\big)\,, \text{ by self-adjointness of } P \text{ and } Q \\
&\le (1-\delta_1)\,\big((QP)^t Q\mathbf{1}, (QP)^t Q\mathbf{1}\big)\,, \text{ by Exercise 7.19} \\
&\le (1-\delta_1)\,\big(P(QP)^{t-1} Q\mathbf{1}, P(QP)^{t-1} Q\mathbf{1}\big)\,, \text{ by } Q \text{ being a projection} \\
&\le (1-\delta_1)\,(1-\delta_2)\,\big((QP)^{t-1} Q\mathbf{1}, (QP)^{t-1} Q\mathbf{1}\big)\,, \text{ by Exercise 7.19} \\
&\le (1-\delta_1)\,(1-\delta_2)^t\,\big(Q\mathbf{1}, Q\mathbf{1}\big)\big)\,, \text{ by iterating previous step} \\
&\le (1-\delta_1)\,(1-\delta_2)^t\,\beta\,,
\end{aligned}$$

and we are done, at least for odd times. For even times, we can just use monotonicity in $t$. □

## 7.4 From infinite to finite: Kazhdan groups and expanders

{ss.expanders}

From the previous section, one might think that transitive expanders are the finite analogues of non-amenable groups. This is true in some sense, but this does not mean that we can easily produce transitive expanders from non-amenable groups. A simple but crucial obstacle is the following:

{ex.coveringsp

▷ **Exercise 7.20.**

**(a)** *If $G' \longrightarrow G$ is a covering map of infinite graphs, then the spectral radii satisfy $\rho(G') \le \rho(G)$, i.e., the larger graph is more non-amenable. In particular, if $G$ is an infinite $k$-regular graph, then $\rho(G) \ge \rho(\mathbb{T}_k) = \frac{2\sqrt{k-1}}{k}$.*

**(b)** *If $G' \longrightarrow G$ is a covering map of finite graphs, then $\lambda_2(G') \ge \lambda_2(G)$, i.e., the larger graph is a worse expander.*

The main group-theoretic source of expanders is the following notion, introduced in [Kaz67]. See the book [BekdHV08] for a comprehensive treatment.

{d.kazhdan}

**Definition 7.12.** *We say that a countable group $\Gamma$ has* **Kazhdan's property (T)** *if, for any unitary representation $\rho : \Gamma \longrightarrow \mathsf{U}(\mathcal{H})$ on a complex Hilbert space $\mathcal{H}$ without non-zero invariant vectors (fixed by all $g \in \Gamma$), there exists a $\kappa > 0$ and a finite subset $K \subset \Gamma$ such that for every nonzero $v \in \mathcal{H}$ there exists $k \in K$ with*

$$\|\rho(k)v - v\| \ge \kappa \|v\|\,. \tag{7.11}$$

{e.Kaka}

*If there exists a pair $\kappa_\rho, K_\rho$ for each $\rho$, then there also exists one $\kappa, K$ that works for every $\rho$ simultaneously. The reason is that for any sequence $\rho_1, \rho_2, \dots$ one can take the representation $\pi = \bigoplus_n \rho_n$, and then the pair $\kappa_\pi, K_\pi$ will work for each $\rho_i$, as well.*

In other words, property (T) means that having no (non-zero) invariant vectors in a representation always implies that there are no almost invariant vectors either, where having almost-invariant vectors means that for every finite $K \subset \Gamma$ and every $\kappa > 0$ there is some non-zero $v \in \mathcal{H}$ with

$$\sup_{k \in K} \|\rho(k)v - v\| < \kappa \|v\| \,. \tag{7.12}$$ {e.noKaka}

The notation (T) may be considered as a visual representation of the formulation that the trivial representation is isolated in the so-called Fell topology on the unitary representations of the group — a topology we are not going to define here. Also, the original definition is in the generality of topological groups $\Gamma$, so that $K$ stands not for finite, but for compact subsets. Finally, note that if we have an action of $\Gamma$ on a real Hilbert space by orthogonal transformations, then we can also consider it as a unitary action on a complex Hilbert space, hence the Kazhdan property will apply to real actions, too.

For any generating set $S \subset \Gamma$, we can define the Kazhdan constant

$$\kappa(\Gamma, S) := \inf \left\{ \sup_{s \in S} \frac{\|\rho(s)v - v\|}{\|v\|} \;:\; \Gamma \xrightarrow{\rho} \mathsf{U}(\mathcal{H}) \text{ without invariant vectors, } v \in \mathcal{H} \setminus \{0\} \right\}.$$

Obviously, if there is a finite $S$ with $\kappa(\Gamma, S) > 0$, then $\Gamma$ is Kazhdan. Furthermore, if $S_1 \subseteq S_2$, then $\kappa(\Gamma, S_1) \le \kappa(\Gamma, S_2)$. One can easily get sort of a reverse inequality, as well:

▷ **Exercise 7.21.** *For any pair $S_1, S_2$ of finite generating sets of a group $\Gamma$, show that if $\kappa(\Gamma, S_1) > 0$, then $\kappa(\Gamma, S_2) > 0$, too.*

After all these definitions, we can finally say that Kazhdan groups do exist: finite groups are such.

**Example:** If $\Gamma$ is finite, then $\kappa(\Gamma, \Gamma) \ge \sqrt{2}$. Let us assume, on the contrary, that $v$ is a vector in $\mathcal{H}$ with norm 1 with $\|\rho(g)v - v\| < \sqrt{2}$, for all $g \in \Gamma$. Then $\rho(g)v$ is in the open half-space $\mathcal{H}_v^+ := \{w \in \mathcal{H} : (v, w) > 0\}$ for all $g$. If we average over all $g$,

$$v_0 := \frac{1}{|\Gamma|} \sum_g \rho(g)v \,,$$

we obtain an invariant factor which is in the interior of $\mathcal{H}_v^+$, hence it is non-zero.

It is not obvious to produce infinite Kazhdan groups. First of all, infinite amenable groups are not good:

▷ **Exercise 7.22.** *If a group is Kazhdan and amenable, then it is a finite group. (Hint: The indicators of Følner sets are almost invariant vectors in $\ell^2(\Gamma)$.)*

But the "most non-amenable group", the free group $F_2$ with two generators, is not Kazhdan either, since there exists a surjection $F_2 \longrightarrow \mathbb{Z}$ (the obvious projection) to an amenable group, and then $F_2$ acts on $\ell^2(\mathbb{Z})$ just as $\mathbb{Z}$ does.

The main examples are that $\mathsf{SL}_d(\mathbb{Z})$, for $d \ge 3$, are Kazhdan. See the discussion preceding Theorem 14.5 for a tiny bit of intuition why the $d = 2$ case is not good for property (T).

The main motivation for Kazhdan to introduce his property was to show that lattices in groups like $\mathsf{SL}_d(\mathbb{R})$ for $d \ge 3$ are finitely generated. Here is why property (T) helps.

**Proposition 7.13.** *Any countable Kazhdan group $\Gamma$ is finitely generated. Moreover, $\kappa(\Gamma, S) > 0$ for any finite generating set $S$.*

*Proof.* Let $\mathcal{C}$ be the set of finitely generated subgroups of $\Gamma$. For any $H \in \mathcal{C}$, we can take the (say, left) cosets $\Gamma/H := \{gH : g \in \Gamma\}$ (as a set, not as a multiset!), and then $\Gamma$ acts on $\ell^2(\Gamma/H)$ by left translations, a representation we denote by $\lambda_{\Gamma/H}$:

$$\lambda_{\Gamma/H}(g)(f)(xH) := f(g^{-1}xH) .$$

Now, if we take the representation

$$\pi := \bigoplus_{H\in\mathcal{C}} \lambda_{\Gamma/H} : \Gamma \longrightarrow \bigoplus_{H\in\mathcal{C}} \ell^2(\Gamma/H) ,$$

and any finite subset $K \subset \Gamma$, then the Kronecker delta function $\delta_{\langle K\rangle} \in \bigoplus_{H\in\mathcal{C}} \ell^2(\Gamma/H)$ is invariant under any $\pi(g)$, $g \in \langle K\rangle$. That is, these $\delta_{\langle K\rangle}$ are almost-invariant vectors of $\pi$, as defined in (7.12). Since $\Gamma$ is assumed to be Kazhdan, $\pi$ must also have true invariant vectors. But any non-zero coordinate of such a vector, indexed by some $H \in \mathcal{C}$, is invariant in $\ell^2(\Gamma/H)$, and therefore $\Gamma/H$ must be finite. Since $H$ is finitely generated, $\Gamma$ is finitely generated, as well.

Now assume that $\kappa(\Gamma, S) = 0$ for some finite generating set $S$. Then there is a sequence of representations $\rho_n$ and unit vectors $v_n \in \mathcal{H}_n$ such that $\|\rho_n(s)v_n - v_n\| \to 0$ for every $s \in S$. Then, since any $g \in \Gamma$ can be written as a finite product $g = s_1 \cdots s_t$ with $s_i \in S$, we have (writing $\rho = \rho_n$ as a left action)

$$\begin{aligned}
\|\rho_n(g)v_n - v_n\| &= \big\|\rho(s_1\cdots s_{t-1})\rho(s_t)v_n - \rho(s_1\cdots s_{t-1})v_n + \rho(s_1\cdots s_{t-2})\rho(s_{t-1})v_n - \cdots + \rho(s_1)v_n - v_n\big\| \\
&\le \big\|\rho(s_1\cdots s_{t-1})\rho(s_t)v_n - \rho(s_1\cdots s_{t-1})v_n\big\| + \cdots + \big\|\rho(s_1)v_n - v_n\big\| \\
&= \|\rho(s_t)v_n - v_n\| + \cdots + \|\rho(s_1)v_n - v_n\| ,
\end{aligned}$$

since $\rho$ is unitary. Each of the $t$ terms in this sum goes to 0, hence $\|\rho_n(g)v_n - v_n\| \to 0$ for any fixed $g \in \Gamma$. This means that there cannot exist a finite set $K$ that works for every $\rho$ and every $v$ in the simultaneous version of property (T). $\square$

The first reason for property (T) entering probability theory was the construction of expanders. On one hand, we have the following not very difficult theorem, see [Lub94, Proposition 1.2.1] or [HooLW06, Lemma 1.9]:

**Theorem 7.14** ([Pin73],[Pip77]). *Asymptotically almost every $d$-regular graph is an expander.*

However, as with some other combinatorial problems, although there are many expanders, it is hard to find an infinite family explicitly.

▷ **Exercise 7.23** (Margulis 1973). *If $\Gamma$ is Kazhdan and infinite, with finite factor groups $\Gamma_n$, let $G_n = G(\Gamma_n, S)$ be the Cayley graph of $\Gamma_n$ with a fixed generating set $S$ of $G$. Then the $G_n$ are expanders.*

This was the first explicit construction of an expander family, from an infinite Kazhdan group with infinitely many different finite factors. Since then, much easier constructions have been found. An early example that uses only some simple harmonic analysis is [GaGa81]. More recently, a completely elementary construction was found by Reingold, Vadhan and Wigderson (2002), using the so-called zig-zag product; see [HooLW06, Section 9].

Any infinite $k$-regular transitive graph $G$ has spectral radius $\rho(G) \geq \rho(\mathbb{T}_k) = \frac{2\sqrt{k-1}}{k}$, and any finite $k$-regular graph on $n$ vertices has second largest eigenvalue $\lambda_2(G) \geq \frac{2\sqrt{k-1}}{k} - o(1)$ as $n \to \infty$, see [HooLW06, Section 5.2].

A sequence of $k$-regular graphs $G_n$ are called **Ramanujan graphs** if

$$\liminf_{n\to\infty} \lambda_2(G_n) = \frac{2\sqrt{k-1}}{k}.$$

So, the Ramanujan graphs are the ones with largest spectral gap.

Again, for any $\epsilon > 0$, asymptotically almost every $k$-regular graph has $\lambda_2(G) \leq \frac{2\sqrt{k-1}}{k} + \epsilon$ — this is due to Joel Friedman, and is much harder to prove than just being expanders. There are also explicit constructions, first done by Lubotzky, Philips and Sarnak in 1988. They showed that $G(\mathsf{SL}_3(\mathbb{F}_p), S)$, with appropriate generating sets $S$ are Ramanujan graphs. See [HooLW06] and [Lub94].

**Conjecture 7.15** (Question on $k$-regular graphs)**.** *What is the asymptotic distribution of $\lambda_2$ if we choose $k$-regular graphs uniformly at random? The limiting distribution should be*

$$\frac{\lambda_2 - \frac{2\sqrt{k-1}}{k} + f(n)}{g(n)} \to TW_{\beta=1},$$

*for some unknown normalizing functions $f, g$, where $TW_\beta$ stands for the Tracy-Widom distribution for the fluctuations of the largest eigenvalue of the $\beta$-ensemble of random matrices; in particular, for $\beta = 1$, it is the GOE ensemble, the real symmetric Gaussian random matrices.*

The motivation behind this conjecture is that the adjacency matrix of a random $k$-regular graph is a sparse version of a random real symmetric Gaussian matrix. In fact, the *typical* eigenvalue of large random $k$-regular graphs, in the $k \to \infty$ limit, looks like the typical eigenvalue of large random matrices; more precisely, the eigenvalue distribution of large random $k$-regular graphs converges to the spectral measure of $\mathbb{T}_k$ (see Section 14.2), and this spectral measure converges to Wigner's semicircle law as $k \to \infty$, which is the limiting distribution of the eigenvalue distribution of large random matrices. See, for instance, [Dei07] on the random matrix ensembles and their universal appearance throughout science.

# 8 Isoperimetric inequalities and return probabilities in general

{s.return}

The main topic of this chapter is a big generalization of the deduction of exponential heat kernel decay from non-amenability (or fast mixing from positive Cheeger constant): general isoperimetric inequalities also give sharp upper bounds on the return probabilities. This was first discovered by Varopoulos [Var85a], developed further with his students [VarSCC92]. This approach is very much functional analytic, a continuation of the methods encountered in Sections 7.1 and 7.2, proving and using the so-called Nash inequalities. We will sketch this approach in the first section, then will study in more depth the method of evolving sets, a beautiful probabilistic approach developed by Ben Morris and Yuval Peres [MorP05].

An important special case can be formulated as follows:

{t.varopoulos}

**Theorem 8.1.** *$IP_d \Rightarrow p_n(x,y) \leq Cn^{-\frac{d}{2}}$. For transitive graphs, this is an equivalence.*

As can be seen from the second statement of the theorem, there are also bounds going in the other direction. For instance, an $n^{-d/2}$ heat kernel decay implies a so-called $d$-dimensional Faber-Krahn inequality, see [Cou00] for a nice overview, and [CouGP01] for a geometric approach for groups. For general graphs, this is strong enough to deduce an at least $d$-dimensional volume growth, but not $IP_d$. Similarly, lower bounds on the volume growth, without any regularity assumption, do not imply upper bounds on return probabilities. We will not study these questions, but before diving into Nash inequalities and evolving sets, let us give two quick exercises:

▷ **Exercise 8.1.**

**(a)** *Using the Carne-Varopolous bound, Theorem 9.2 below, show that a* $|B_n(x)| = o(n^2/\log n)$ *volume growth in a bounded degree graph implies recurrence.*

**(b)** *Construct a recurrent tree with exponential volume growth.*

## 8.1 Poincaré and Nash inequalities

{ss.PoincareNa

Recall the **Sobolev inequality** from Proposition 7.5: for $f \in \ell_0(V)$,

$$V \text{ satisfies } IP_d(\kappa) \implies \|f\|_{d/(d-1)} \le C(\kappa)\,\|\nabla f\|_1 \,. \tag{8.1}$$ {e.Sobolev}

To compare this inequality for different $d$ values, take $f_n$ to be roughly constant with a support of $\pi$-measure $n$. Then $\|f_n\|_{d/(d-1)} \asymp n^{(d-1)/d}$. So, one way of thinking of this inequality is that the function $f$ becomes smaller by taking the derivative in any direction, but if the space has better isoperimetry, then from each point we have more directions, hence the integral $\|\nabla f\|_1$ loses less mass compared to $\|f\|_1$. In particular, for the non-amenable case ($d = \infty$), we already have $\|f\|_1 \le C(\kappa)\,\|\nabla f\|_1$.

So, in (8.1), to make up for the loss caused by taking the derivative, we are taking different norms on the two sides, depending on the isoperimetry of the space. Another natural way of compensation would be to multiply the right hand side by a factor depending on the size of the support. This is done in the **Poincaré inequalities**, well-known from PDEs [Eva98, Section 5.8]:

If $U \subseteq \mathbb{R}^d$ with Lipschitz boundary, then

$$\big\|f(x) - f_{RU}\big\|_{L^p(RU)} \le C_U\, R\,\|\nabla f\|_{L^p(RU)} \,, \tag{8.2}$$ {e.Poincare}

with

$$f_{RU} = \frac{\int_{RU} f(x)\,dx}{\mathrm{vol}(RU)}.$$

Since the quality of the isoperimetry now does not appear in (8.2), one might hope that this can be generalized to almost arbitrary spaces. Indeed, the following holds:

{ex.SC}

▷ **Exercise 8.2** (Saloff-Coste's Poincaré inequality [Kle10])**.** *Show that if* $f : \Gamma \longrightarrow \mathbb{R}$ *on any group* $\Gamma$*, and* $B_R$ *is a ball of radius* $R$ *in a Cayley graph, then*

$$\big\|f - f_{B_R}\big\|_{\ell^2(B_R)} \le 2\,\frac{\mathrm{vol}(B_{2R})}{\mathrm{vol}(B_R)}\,R\,\|\nabla f\|_{\ell^2(B_{3R})}. \tag{8.3}$$ {e.SC}

*Hints: we have*

$$\big|f(y) - f_{B_R}\big| \le \frac{\sum_{z\in B_R} |f(z) - f(y)|}{\mathrm{vol}(B_R)} \le \frac{\sum_{g\in B_{2R}} |f(yg) - f(y)|}{\mathrm{vol}(B_R)}\,, \tag{8.4}$$ {e.Hint a}

*and if* $g = s_1 \dots s_m$ *with* $m \le 2R$ *and generators* $s_i \in S$*, then*

$$\big|f(yg) - f(y)\big| \le \big|f(ys_1 \dots s_m) - f(ys_1 \dots s_{m-1})\big| + \dots + \big|f(ys_1) - f(y)\big|\,. \tag{8.5}$$ {e.Hint b}

Note that for non-amenable groups, we have the **Dirichlet inequality** $\|f\|_2 \le C\,\|\nabla f\|_2$, see Theorem 7.3, i.e., the factor $R$ on the right hand side of (8.2) can be spared. So, just like in the case of Sobolev inequalities, no compensation is needed.

The following exercise says that harmonic functions show that the Poincaré inequality (8.3) is essentially sharp. The statement can be regarded as a discrete analogue of the classical theorem that functions satisfying the Mean Value Property are smooth. (Such a function cannot go up and down too much: whenever there is an edge contributing something to $\|\nabla f\|$, the harmonicity carries this contribution to a large distance.)

{ex.revPoincar

▷ **Exercise 8.3** (Reverse Poincaré inequality)**.** *Show that there is a constant $c = c(\Gamma, S) > 0$ such that for any harmonic function $f$ on the Cayley graph $G(\Gamma, S)$,*

$$c\,R\,\|\nabla f\|_{\ell^2(B_R)} \le \|f\|_{\ell^2(B_{2R})}\,. \tag{8.6}$$ {e.reverse}

There are several ways how Poincaré or similar inequalities can be used for proving heat kernel estimates. The first work in this direction was by Varopoulos [Var85a]. He proved sharp heat kernel upper bounds by first proving that an isoperimetric inequality $IP_{\mathfrak{F}}(\kappa)$, with $\mathfrak{f} = \frac{t}{\mathfrak{F}(t)}$ increasing, implies the following **Nash inequality**:

$$\|f\|_2^2 \le \mathfrak{g}\big(\|f\|_1^2/\|f\|_2^2\big)\,\mathcal{E}_{P,\pi}(f), \qquad f \in \ell_0(V)\,, \tag{8.7}$$ {e.Nash}

where $\mathfrak{g} = C_\kappa \mathfrak{f}(4t)^2$. The proof is quite similar to the proofs of Proposition 7.5 and Theorem 7.3. To make the result more readable, consider the case of $IP_d(\kappa)$, i.e., $\mathfrak{F}(t) = t^{(d-1)/d}$. Then we have $\mathfrak{g}(t) = C_\kappa t^{2/d}$, and (8.7) reads as

$$\|f\|_2 \le C_\kappa\left(\|f\|_1^2/\|f\|_2^2\right)^{1/d}\|\nabla f\|_2\,,$$

some kind of mixture of the Sobolev and Poincaré inequalities (8.1) and (8.3). For instance, when $f = \mathbf{1}_{B_R}$, we get the usual $\|f\|_2 \le C_\kappa\,R\,\|\nabla f\|_2$. But we will apply (8.7) to some other functions.

The next step is to note that

$$\mathcal{E}_{P,\pi}(f) \le 2(\|f\|_2^2 - \|\overline{P}f\|_2^2)\,,$$

where $\overline{P} = (I+P)/2$ is the Markov operator for the walk made lazy. Therefore, applying the Nash inequality to $f = \overline{P}^n\delta_x$ for each $n$, we arrive at a difference inequality on the sequence

$$u(n) := \|\overline{P}^n\delta_x\|_2^2 = (\overline{P}^{2n}\delta_x, \delta_x) = \overline{p}_{2n}(x,x)C_x\,,$$

namely

$$u(n) \le 2\mathfrak{g}\big(1/u(n)\big)\big(u(n) - u(n+1)\big)\,.$$

This gives an upper bound on the decay rate of $u(n)$ as a solution of a simple differential equation. Then we can use that $p^{2n}(x,x) \le 2\overline{p}^{2n}(x,x)$. See [Woe00, Section 14.A] for more detail.

A different approach, with stronger results, is given in [Del99]. For Markov chains on graphs with nice regular $d$-dimensional volume growth conditions and Poincaré inequalities, he proves that

$$c_1\,n^{-d/2}\exp(-C_1\,d(x,y)^2/n) \le p_n(x,y) \le c_2\,n^{-d/2}\exp(-C_2\,d(x,y)^2/n)\,.$$

Let us note here that a very general Gaussian estimate on the off-diagonal heat kernel decay is given by the Carne-Varopoulos Theorem 9.2 below.

## 8.2 Evolving sets: the results

{ss.evolving}

Recall that a Markov chain is reversible if there is a measure $m$ with $\pi(x)p(x,y) = \pi(y)p(y,x)$; such a $\pi$ is always stationary. This happens iff the transition probabilities can be given by symmetric conductances: $c(x,y) = c(y,x)$ with $p(x,y) = \frac{c(x,y)}{C_x}$, and then $\pi(x) = C_x$ is a good choice.

Even for the non-reversible case, when the stationary measure $\pi(x)$ is not $C_x$, we define $Q(x,y) := \pi(x)p(x,y)$, and the **isoperimetric profile** is

$$\phi(r) := \inf\left\{ \frac{Q(S,S^c)}{\pi(S)} : \pi(S) \le r \right\}.$$

For instance, $IP_d(\kappa)$ implies $\phi(r) \ge \kappa r^{-1/d}$.

On a finite chain, we take $\pi(V) = 1$, and let $\phi(r) = \phi(1/2)$ for $r > 1/2$.

{t.Morris-Pere

**Theorem 8.2** (Morris-Peres [MorP05]). *Suppose* $0 < \gamma < 1/2$, $p(x,x) > \gamma\ \forall x \in V$. *If*

$$n \ge 1 + \frac{(1-\gamma)^2}{\gamma^2} \int_{4\min(\pi(x),\pi(y))}^{4/\epsilon} \frac{2\,du}{u\phi(u)^2}, \tag{8.8}$$

{e.Morris-Pere

*then*

$$p_n(x,y)/\pi(y) < \epsilon \qquad \text{or} \qquad |p_n(x,y)/\pi(y) - 1| < \epsilon\,,$$

*depending on whether the chain is infinite or finite.*

For us, the most important special cases are the following:

1) By the Coulhon-Saloff-Coste isoperimetric inequality, Theorem 5.11, any group of polynomial growth $d$ satisfies $IP_d$. Then the integral in (8.8) becomes

$$C_\gamma \int_4^{4/\epsilon} \frac{4\,du}{u^{1-2/d}} \asymp (1/\epsilon)^{2/d}.$$

That is, the return probability will be less than $\epsilon$ after $n \asymp \epsilon^{-2/d}$ steps, so $p_n(x,x) < Cn^{-d/2}$.

2) $IP_\infty$ turns the integral into $C_\gamma \log(1/\epsilon)$, hence $p_n(x,x) < \exp(-cn)$, as we know from the Kesten-Cheeger-Dodziuk-Mohar Theorem 7.3.

3) For groups of volume growth $\exp(c\,r^\alpha)$ with $\alpha \in (0,1]$, the CSC isoperimetric inequality implies $\phi(r) \ge c/\log^{1/\alpha} r$. Then the integral becomes

$$\int_1^{1/\epsilon} \frac{\log^{2/\alpha} u}{u}\, du \asymp \log^{1+\frac{2}{\alpha}}(1/\epsilon)\,,$$

thus $p_n(x,y) \le \exp(-cn^{\alpha/(\alpha+2)})$. At least for $\alpha = 1$, this is again the best possible, by the following exercise.

▷ **Exercise 8.4.*** *Show that on the lamplighter group* $\mathbb{Z}_2 \wr \mathbb{Z}$ *we have* $p_n(x,x) \ge \exp(-cn^{1/3})$.

As we discussed in the introduction to this section, these bounds were first proved by Varopoulos, using Nash inequalities. The beauty of the Morris-Peres approach is that is completely probabilistic, defining and using the so-called evolving set process. Moreover, it works also for the non-reversible case, unlike the functional analytic tools. (Although, in the non-reversible case, finding the stationary measure and the corresponding isoperimetric profile might be quite difficult.)

The integral (8.8) using the isoperimetric profile was first found by Lovász and Kannan [LovKa99], for finite Markov chains, but they deduced mixing only in total variation distance, not uniformly as Morris and Peres. To demystify the formula a bit, note that for a finite Markov chain on $V$ with uniform stationary distribution, using the bound $\phi(r) \geq \phi(1/2) = h$ for all $r$, where $h$ is the Cheeger constant (since $r = 1/2$ is an infimum over a larger set), the Lovász-Kannan integral bound implies the upper bound

$$\asymp \int_{\frac{1}{|V_n|}}^{1} \frac{1}{rh^2}\, dr = \frac{\log |V_n|}{h^2}$$

on the mixing time, as was shown earlier in Theorem 7.7. The idea for the improvement via the integral is that, especially in geometric or transitive settings, small subsets often have better isoperimetry than the large ones. (Recall here the end of Section 5.3, and also see the next exercise.)

**Example**: Take an $n$-box in $\mathbb{Z}^d$. A sub-box of side-length $t$ has stationary measure $r = t^d/n^d$ and boundary $\asymp t^{d-1}/n^d$, hence the isoperimetric profile is $\phi(r) \asymp 1/t = r^{-1/d}/n$. (Of course, one would need to show that sub-boxes are at least roughly optimal for the isoperimetric problem. This can be done similarly to the infinite case $\mathbb{Z}^d$, see Section 5.4, with some non-trivial extra work.) This is clearly decreasing as $r$ grows. Therefore, the Cheeger constant is $h \asymp 1/n$, achieved at $r \asymp 1$. Using Theorem 7.6, we get that the spectral gap is at least of order $1/n^2$, which is still sharp, but then Theorem 7.7 gives only a uniform mixing time $\leq C_d n^2 \log n$. However, using the Lovász-Kannan integral, the mixing time comes to $C_d n^2$.

Another standard example for random walk mixing is the hypercube $\{0,1\}^d$ with the usual edges.

▷ **Exercise 8.5.** *For $0 \leq m \leq d$, show that the minimal edge-boundary for a subset of $2^m$ vertices in the hypercube $\{0,1\}^d$ is achieved by the $m$-dimensional sub-cubes. (Hint: use induction.)*

The isoperimetric profile of $\{0,1\}^d$ can also be computed, but it is rather irregular and hard to work with in the Lovász-Kannan integral, and it does not yield the optimal bound. But, as we will see, with the evolving sets approach, one needs the isoperimetry only for sets that do arise in the evolving sets process: {ex.hypermix}

▷ **Exercise 8.6.** * *Prove $O(n \log n)$ mixing time for $\{0,1\}^n$ using evolving sets and the standard one-dimensional Central Limit Theorem.*

## 8.3 Evolving sets: the proof {ss.evolproof}

We first need to give the necessary definitions for the evolving sets approach, and state (with or without formal proofs) a few basic lemmas. Then, before embarking on the proof of the full Theorem 8.2, we will show how the method works in the simplest case, by giving a quick proof of Kesten's Theorem 7.3, the exponential decay of return probabilities in the non-amenable case, even for non-reversible chains.

Let $S \subseteq V$. Define: $\widetilde{S} = \{y : Q(S,y) \geq U\pi(y)\}$, where $U \sim \mathsf{Unif}[0,1]$. Remember: $Q(S,y) = \sum_{x \in S} \pi(x)p(x,y)$ with $\pi P = \pi$, a fixed stationary measure. This is one step of the **evolving set process**. Thus

$$\mathbf{P}\big[\, y \in \widetilde{S} \;\big|\; S \,\big] = \mathbf{P}\big[\, Q(S,y) \geq U\pi(y) \,\big] = \frac{Q(S,y)}{\pi(y)}\,, \tag{8.9}$$ {e.evol}

which leads to

$$\mathbf{E}[\, \pi(\widetilde{S}) \mid S \,] = \sum_y \pi(y) \mathbf{P}\big[\, y \in \widetilde{S} \;\big|\; S \,\big] = \sum_y Q(S,y) = \pi(S)\,.$$

Therefore, for the evolving set process $S_{n+1} = \widetilde{S}_n$, the sequence $\{\pi(S_n)\}$ is a martingale.

Now take $S_0 = \{x\}$; then $\mathbf{E}[\, \pi(S_n) \,] = \pi(x)$. Moreover:

**Lemma 8.3.** $\mathbf{P}\big[\, y \in S_n \bigm| S_0 = \{x\} \,\big] \frac{\pi(y)}{\pi(x)} = p_n(x,y)$.

*Proof.* We use induction on $n$. The case $n=0$ is trivial. Then, using the induction hypothesis,

$$p_n(x,y) = \sum_z p_{n-1}(x,z)\, p(z,y) = \sum_z \mathbf{P}_{\{x\}}\big[\, z \in S_{n-1} \,\big] \frac{\pi(z)}{\pi(x)} p(z,y)\,.$$

Since

$$\mathbf{P}_{\{x\}}\big[\, z \in S_{n-1} \,\big] \pi(z) p(z,y) = \mathbf{E}_{\{x\}}\Big[\, \mathbf{1}_{\{z \in S_{n-1}\}} Q(z,y) \,\Big]\,,$$

the above sum over $z$ can be written as

$$\frac{1}{\pi(x)} \mathbf{E}_{\{x\}}\big[\, Q(S_{n-1}, y) \,\big]\,.$$

Then, using (8.9), this is equal to

$$\frac{\pi(y)}{\pi(x)} \mathbf{E}_{\{x\}}\Big[\, \mathbf{P}\big[\, y \in S_n \bigm| S_{n-1} \,\big] \,\Big] = \frac{\pi(y)}{\pi(x)} \mathbf{P}_{\{x\}}\big[\, y \in S_n \,\big]\,,$$

and we are done. $\square$

These two properties, that $\pi(S_n)$ is a martingale and $\mathbf{P}_{\{x\}}\,[y \in S_n]\,\pi(y) = p_n(x,y)\pi(x)$, at least when the stationary measure $\pi$ is uniform, are shared by the Markov chain trajectory $\{X_n\}$. However, since $\mathbf{P}[\, y \in \widetilde{S} \mid S \,] = \frac{Q(S,y)}{\pi(y)}$, the size of $\widetilde{S}$ will have a conditional variance depending on the size of the boundary of $S$: the larger the boundary, the larger the conditional variance that the evolving set has. This makes it extremely useful if we want to study how the isoperimetric profile affects the random walk.

Recall that, if $f$ is a concave function, by Jensen's inequality $\mathbf{E}\big[\, f(X) \,\big] \le f(\mathbf{E}\big[\, X \,\big])$. Moreover, if there is considerable variance, then the inequality is strict. For example, if $f$ is $\sqrt{\ }$:

▷ **Exercise 8.7.** *If* $\mathrm{Var}[X] \ge c\,(\mathbf{E}X)^2$ *then* $\mathbf{E}\big[\, \sqrt{X} \,\big] \le (1-c')\sqrt{\mathbf{E}X}$*, where* $c' > 0$ *depends only on* $c > 0$.

Recall that $Q(x,y) = \pi(x) \cdot p(x,y)$ thus $Q(A,B) = \sum_{a \in A, b \in B} \pi(a) p(a,b)$.

*Proof of Theorem 8.2.* We will focus on the case of an infinite state space. Let us first state a lemma that formalizes the above observations: large boundary for $S$ means large conditional variance for $\widetilde{S}$, and that in turn means a definite decrease in $\mathbf{E}\sqrt{\pi(\widetilde{S})}$ compared to $\mathbf{E}\sqrt{\pi(S)}$. We omit the proof, because it is slightly technical, and we have already explained the main ideas anyway.

{l.teclem1}

**Lemma 8.4.** *Assume* $0 < \gamma \le 1/2$ *and* $p(x,x) \ge \gamma$*, for all* $x \in V$*. If* $\Psi(S) = 1 - \mathbf{E}_S\left[\sqrt{\frac{\pi(\widetilde{S})}{\pi(S)}}\right]$ *then, for each* $S \subset V$*, we have that*

$$\Psi(S) \ge \frac{\gamma^2}{2(1-\gamma)^2} \phi(S)^2,$$

*where*

$$\phi(S) = \frac{Q(S, S^c)}{\pi(S)}.$$

Notice that, if we define $\Psi(r) = \inf\{\Psi(S) \;:\; \pi(S) \le r\}$, then the lemma gives

$$\Psi(r) \ge \frac{\gamma^2}{2(1-\gamma)^2} \phi(r)^2,$$

and the proof of Theorem 8.2 reduces to show that $p_n(x,y)/\pi(y) < \epsilon$ if $n > \int \frac{1}{r\Psi(r)}\, dr$.

As promised, an example of the usefulness of this method is that we can immediately get exponential decay for non-amenable chains:

If $\phi(r) \geq h > 0 \; \forall r$, then Lemma 8.4 implies that $\Psi(r) \geq h' > 0$, i.e.,

$$\mathbf{E}\big[\sqrt{\pi(\widetilde{S})} \mid \pi(S)\big] \leq (1-h')\sqrt{\pi(S)}.$$

So, by iterating from $S_0 = \{x\}$ to $S_n$ we obtain $\mathbf{E}_{\{x\}}\big[\sqrt{\pi(S_n)}\big] \leq \exp(-Cn)$. If we assume $\pi(x) \equiv 1$, for example in the group case, this implies that $p_n(x,y) = \mathbf{P}_{\{x\}}[\, y \in S_n \,] \leq \mathbf{E}_{\{x\}}\sqrt{|S_n|} \leq \exp(-Cn)$, as desired, where the middle inequality comes from the ridiculous bound $\mathbf{P}_{\{x\}}[\, y \in S_n \,] \leq \mathbf{P}_{\{x\}}\big[\sqrt{\pi(S_n)} \geq 1\big]$ and Markov's inequality.

However, for the general case we will have to work harder. The replacement for the last ridiculous bound will be (8.14) below, which handles the case of a non-constant $\pi$ better than a simple application of Markov's inequality. A more serious issue is the non-uniformity of the isoperimetric bound. Recall that usually (e.g., on the lattice $\mathbb{Z}^d$) the boundary-to-volume ratio is smaller for larger sets, hence, if at some point, $\pi(S_n)$ happens to big (which is not very likely for a martingale at any given time $n$, since $\mathbf{E}\pi(S_n) = \mathbf{E}\pi(x)$, but still may happen at some random times), then our bound from Lemma 8.4 becomes weak, and we lose control. It would be much better to have a stronger downward push for larger sets; when $\pi(S_n)$ is small, then we are happy anyway.

For this reason, we introduce some kind of a dual process. Denote the evolving set transition kernel by $K(S,A) = \mathbf{P}_S[\, \tilde{S} = A \,]$, the probability that, being in $S$, the next step is $A$. Now, consider a non-negative function on the state space $X = 2^V$ of the evolving set process, which equals 1 on some $\Omega \subset X$ and equals 0 for another subset $\Lambda \subset X$, and harmonic on $X \setminus (\Omega \cup \Lambda)$. Then, we can take Doob's $h$-transform:

$$\widehat{K}(S,A) := \frac{h(A)}{h(S)} K(S,A)$$

which is indeed a transition kernel: by the harmonicity of $h$, for $A \notin \Omega \cup \Lambda$,

$$\sum_A \widehat{K}(S,A) = \sum_A \frac{h(A)}{h(S)} K(S,A) = \frac{h(S)}{h(S)} = 1.$$

The usual reason for introducing the $h$-transform is that the process given by the kernel $\widehat{K}$ is exactly the original process conditioned to hit $\Omega$ before $\Lambda$ (in particular, it cannot be started in $\Lambda$, and once it reaches $\Omega$ it is killed). See Lemma 6.13 and (6.10) for a simple example.

For the evolving set process, we will apply the Doob transform with $h(S) := \pi(S)$. Recall that $\pi(S_n)$ is a martingale on $X \setminus \{\emptyset\}$ exactly because $\pi(S)$ is harmonic for $S \neq \emptyset$. (We have harmonicity at $S = V$ because that is an absorbing state of the original evolving sets chain.) So, for the infinite state space evolving set process, $\widehat{K}$ is the process conditioned on never becoming empty. We do not prove this carefully, since we will not use this fact.

We now define $Z_n := \frac{1}{\sqrt{\pi(S_n)}}$. Then

$$\frac{1}{\pi(x)} \mathbf{E}_{\{x\}} \sqrt{\pi(S_n)} = \widehat{\mathbf{E}}_{\{x\}} Z_n \,, \tag{8.10}$$ {e.ESEZ}

because, as it is easy to check from the definitions, $\widehat{\mathbf{E}}_S f(S_n) = \mathbf{E}_S\left[\frac{\pi(S_n)}{\pi(S)} f(S_n)\right]$ for any function $f$.

In the hat mean:

$$\widehat{\mathbf{E}}\left[\frac{Z_{n+1}}{Z_n} \;\Big|\; S_n\right] = \widehat{\mathbf{E}}\left[\frac{\sqrt{\pi(S_n)}}{\sqrt{\pi(S_{n+1})}} \;\Big|\; S_n\right] = \mathbf{E}\left[\frac{\sqrt{\pi(S_{n+1})}}{\sqrt{\pi(S_n)}} \;\Big|\; S_n\right] < 1 - \Psi(\pi(S_n)), \tag{8.11}$$ {e.e-psi}

with the last inequality provided by the definition of $\Psi(r)$ after Lemma 8.4.

When $Z_n$ is large then $S_n$ is small and $\Psi$ is good enough to get from (8.11) a significant decrease for $Z_n$ in the new process. And, by (8.10), a fast average decay of $Z_n$ in the new process means a fast decay for $S_n$ in the original process, and we are happy! This sounds like black magic, but is nevertheless correct. It is a version of the technique of taking a "strong stationary dual" in [DiaF90].

Let us state another technical lemma, which formalizes how a control like (8.11) that we have now for $Z_n$ in the $\widehat{K}$-process actually leads to a fast decay in $\widehat{\mathbf{E}}$.

{l.lem-tech2}

**Lemma 8.5.** *Assume $f_0, f : (0,\infty) \longrightarrow [0,1]$ are increasing functions and $Z_n > 0$ is a sequence of random variables with $Z_0$ fixed, and $L_n := \mathbf{E} Z_n$.*

- **(a)** *Suppose that $L_{n+1} \le L_n(1 - f(L_n))$ for all $n \ge 0$. If $n \ge \int_\delta^{L_0} \frac{dz}{z f(z)}$ for some $\delta > 0$, then $L_n \le \delta$.*
- **(b)** *Suppose now that $\mathbf{E}[\, Z_{n+1} \mid Z_n \,] \le Z_n(1 - f_0(Z_n))$ for all $n \ge 0$, and let $f(z) := f_0(z/2)/2$. If $n \ge \int_\delta^{Z_0} \frac{dz}{z f(z)}$ for some $\delta > 0$, then $\mathbf{E} Z_n \le \delta$.*

*Proof.* Notice that the recursive condition in part (a) implies that

$$\int_{L_{k+1}}^{L_k} \frac{dz}{z\, f(z)} \ge \frac{1}{f(L_k)} \int_{L_{k+1}}^{L_k} \frac{dz}{z} = \frac{1}{f(L_k)} \ln \frac{L_k}{L_{k+1}} \ge 1\,.$$

Summing this up for $k = 0, 1, \ldots, n-1$ and using that $\int_\delta^{L_0} \frac{dz}{z\, f(z)} \ge \int_{L_n}^{L_0} \frac{dz}{z\, f(z)}$ if $L_n \le \delta$ gives the result.

To get part (b) from part (a), we can use the following simple inequality:

$$\mathbf{E}[\, Z\, f_0(Z)\,] \ge \mathbf{E}\big[\, Z\, \mathbf{1}_{\{Z > \mathbf{E}Z/2\}}\, f_0(\mathbf{E}Z/2)\,\big] \ge \frac{\mathbf{E}Z}{2} f_0(\mathbf{E}Z/2)\,.$$

Thus, the recursive condition on $Z_n$ with $f_0$ implies the condition of part (a) for $\mathbf{E}Z_n$ with $f$, and hence the result follows. □

In order to apply this lemma, we rewrite (8.11) as $\widehat{\mathbf{E}}\,[Z_{n+1} | Z_n] < Z_n(1 - f_0(Z_n))$, with $f_0(z) := \Psi(1/z^2)$, and get that

$$n \ge \int_\delta^{1/\sqrt{\pi(x)}} \frac{2\, dz}{z \Psi(4/z^2)} \quad \text{implies} \quad \widehat{\mathbf{E}}_{\{x\}} Z_n \le \delta\,. \qquad (8.12)$$ {e.ZnPsi}

By (8.10), this gives a good upper bound on $\mathbf{E}_{\{x\}} \sqrt{\pi(S_n)}$, but if we then used just the primitive argument as before, we would get

$$\frac{p_n(x,y)}{\pi(y)} \le \mathbf{P}_{\{x\}}[\, y \in S_n\,] \frac{1}{\pi(x)} \le \frac{\mathbf{E}_{\{x\}} \sqrt{\pi(S_n)}}{\sqrt{\pi(y)}\, \pi(x)} = \frac{\widehat{\mathbf{E}}_{\{x\}} Z_n}{\sqrt{\pi(y)}} \le \frac{\delta}{\sqrt{\pi(y)}}\,,$$

which is not good enough when $\sqrt{\pi(y)}$ is small. So, we need to find a better way to deduce the smallness of return probabilities from the smallness of $\mathbf{E}_{\{x\}} \sqrt{\pi(S_n)}$. Here it comes.

{d.chi2dist}

**Definition 8.6** (A non-symmetric comparison of measures). *We define $\chi$ as:*

$$\chi^2(\mu, \pi) = \sum_y \pi(y) \frac{\mu(y)^2}{\pi(y)^2}.$$

*For finite chains, we had the version (7.5), but we will not discuss finite chains here in the proof.*

For $\mu(\cdot) = p_n(x,\cdot)$ and $\pi(\cdot) = 1$, we have $\chi(p_n(x,\cdot),1) = \sum_y p_n(x,y)^2$. If $p_n(x,\cdot)$ is roughly uniform on a large set, i.e., $p_n(x,y) \asymp \epsilon$ for $y \in A_n$, $|A_n| \asymp \frac{1}{\epsilon}$, then $\chi \asymp \epsilon$, which suggests that the smallness of $\chi$ w.r.t. the uniform $\pi$ is a reasonable measure of being well-spread. In fact, from the smallness of $\chi$ we can deduce uniform smallness:

$$\begin{aligned} \frac{p_{n_1+n_2}(x,y)}{\pi(y)} &= \frac{1}{\pi(y)} \sum_y p_{n_1}(x,z)\, p_{n_2}(z,y) \\ &= \sum_z \frac{p_{n_1}(x,z)}{\pi(z)}\, \pi(z)\, \frac{\overleftarrow{p}_{n_2}(y,z)}{\pi(z)} \\ &\le \chi(p_{n_1}(x,\cdot),\pi)\, \chi(\overleftarrow{p}_{n_2}(y,\cdot),\pi) \end{aligned} \tag{8.13}$$ {e.p-chi}

using Cauchy-Schwartz, where $\overleftarrow{p}_{n_2}(y,z) = \frac{\pi(z)}{\pi(y)} p(z,y)$ stands for the stationary reversal. Note that this stationary reversal has the same stationary measure as the original chain, and hence it is easy to see that its isoperimetric profile is also the same.

Also, we can compare $p_n$ with $\pi$ using evolving sets:

$$\begin{aligned} \chi^2(p_n(x,\cdot),\pi) &= \sum_y \pi(y) \frac{\mathbf{P}_{\{x\}}[\, y \in S_n\,]^2 \pi(y)^2}{\pi(x)^2 \pi(y)^2} \\ &= \frac{1}{\pi(x)^2} \sum_y \pi(y) \mathbf{P}_{\{x\}}[\, y \in S_n, y \in T_n\,] \\ &= \frac{1}{\pi(x)^2} \mathbf{E}_{\{x\}}[\, \pi(S_n \cap T_n)\,], \end{aligned}$$

where $\{S_n\}$ and $\{T_n\}$ are two independent evolving set processes. Then, applying Cauchy-Schwartz,

$$\frac{1}{\pi(x)^2} \mathbf{E}_{\{x\}}\big[\, \pi(S_n \cap T_n)\,\big] \le \frac{1}{\pi(x)^2} \mathbf{E}_{\{x\}}\big[\, \sqrt{\pi(S_n)\pi(T_n)}\,\big].$$

Thus

$$\chi(p_n(x,\cdot),\pi) \le \frac{1}{\pi(x)} \mathbf{E}_{\{x\}}\big[\, \sqrt{\pi(S_n)}\,\big], \tag{8.14}$$ {e.p-esp}

which is better than what we had before.

So, we now apply (8.12) with $\delta = \sqrt{\epsilon}$, so that, together with the inequalities (8.13) and (8.14), and after a change of variables $u = 4/z^2$, we get that

$$n \ge \int_{4(\pi(x)\wedge\pi(y))}^{4/\epsilon} \frac{du}{u\Psi(u)} \quad \text{implies} \quad \frac{p_n(x,y)}{\pi(y)} \le \epsilon\,.$$

Using the relation between $\Psi$ and $\phi$ in Lemma 8.4, we arrive at Theorem 8.2. □

# 9 Speed, entropy, Liouville property, Poisson boundary

{s.speedetc}

## 9.1 Speed of random walks

{ss.Speed}

We have seen many results about the return probabilities $p_n(x,x)$, i.e., about the on-diagonal heat kernel decay. We now say a few words about off-diagonal behaviour.

{l.offdiag}

**Lemma 9.1.** *Given a reversible Markov chain with a stationary measure $\pi(x)$, we have:*

**(a)**

$$p_n(x,y) \le \sqrt{\frac{\pi(y)}{\pi(x)}} \rho^n\,.$$

**(b)**

$$\sup_{x,y} \frac{p_{2n}(x,y)}{\pi(y)} \le \sup_x \frac{p_{2n}(x,x)}{\pi(x)} .$$

*Proof.* Both inequalities rely on Cauchy-Schwarz. For part (a):

$$\begin{aligned}
\pi(x)p_n(x,y) &= (\delta_x, P^n\delta_y) \\
&\le \|\delta_x\|\|P^n\delta_y\| \\
&\le \sqrt{\pi(x)}\|P^n\|\sqrt{\pi(y)} .
\end{aligned}$$

Dividing by $\pi(x)$ and using $\|P^n\| \le \|P\|^n = \rho(P)^n$, we get the result.

For part (b), we also use the fact that $P$ is self-adjoint:

$$\begin{aligned}
\pi(x)p_{2n}(x,y) &= (\delta_x, P^{2n}\delta_y) \\
&= (P^n\delta_x, P^n\delta_y) \\
&\le (P^n\delta_x, P^n\delta_x)^{1/2}(P^n\delta_y, P^n\delta_y)^{1/2} \\
&= \big((\delta_x, P^{2n}\delta_x)(\delta_y, P^{2n}\delta_y)\big)^{1/2} \\
&= (\pi(x)p_{2n}(x,x)\pi(y)p_{2n}(y,y))^{1/2} .
\end{aligned}$$

Now divide both sides by $\pi(y)\pi(x)$ and take the supremum over all $x$ and $y$:

$$\sup_{x,y} \frac{p_{2n}(x,y)}{\pi(y)} \le \sup_{x,y} \frac{p_{2n}(x,x)^{1/2}p_{2n}(y,y)^{1/2}}{\pi(x)^{1/2}\pi(y)^{1/2}} = \sup_x \frac{p_{2n}(x,x)}{\pi(x)} .$$

□

The following important result strengthens part (a) of the previous lemma by basically saying that the off-diagonal decay is at least as fast as on $\mathbb{Z}$.

{t.carne}

**Theorem 9.2** (Carne-Varopoulos [Car85, Var85b])**.** *Given a reversible Markov chain,*

$$p_n(x,y) \le 2\sqrt{\frac{\pi(y)}{\pi(x)}}\rho^n \exp\left(-\frac{\mathrm{dist}(x,y)^2}{2n}\right),$$

*where the distance is measured in the graph metric given by the graph of the Markov chain, i.e., there is an edge between $x$ and $y$ if and only if $p(x,y) > 0$.*

This form is due to Carne, with a miraculous proof using Chebyshev polynomials, which we sketch below. Varopoulos' result was a bit weaker, with a more complicated proof. There is now also a probabilistic proof, see [Pey08].

*Sketch of proof.* We will prove that

$$p_n(x,y) \le \sqrt{\frac{\pi(y)}{\pi(x)}}\,\rho^n\,\mathbf{P}_0\Big[\,|S_n| \ge d(x,y)\,\Big], \tag{9.1}$$ {e.Carne}

where $S_n$ is simple random walk on $\mathbb{Z}$. Then the Azuma-Hoeffding bound Proposition 1.8 applied to $S_n$ implies the bound we claimed.

Using trigonometric identities, one can easily prove by induction that for every $k \ge 1$ there is a degree $k$ real polynomial $T_k$ such that $\cos(k\theta) = T_k(\cos\theta)$ for every $\theta \in \mathbb{R}$. These are the Chebyshev polynomials.

Note that $|T_k(r)| \le 1$ for $r \in [-1,1]$, hence the spectral theorem implies that $Q_k := T_k\big(P/\|P\|\big)$ is an operator on $L^2(V,\pi)$ with norm at most 1. (If the Reader is uncomfortable with spectral theory, just truncate $P$ to get a finite Markov chain in a large neighbourhood of the starting point $x$.)

Now, the connection to SRW on $\mathbb{Z}$ arises by writing

$$r^n = (\cos\theta)^n = \left(\frac{e^{i\theta}+e^{-i\theta}}{2}\right)^n = \sum_{k=-n}^{n} \mathbf{P}_0[\, S_n = k \,]\, e^{ik\theta} = \sum_{k=-n}^{n} \mathbf{P}_0[\, S_n = k \,]\, T_k(r)\,,$$

where $T_{-k} := T_k$. Therefore, using the norm 1 functions $\varphi_x(z) := \mathbf{1}_{x=y}/\sqrt{\pi(x)}$,

$$\frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}}\, p_n(x,y) = (\varphi_x, P^n\varphi_y) = \|P\|^n \sum_{k=-n}^{n} \mathbf{P}_0[\, S_n = k \,]\, (\varphi_x, Q_k\varphi_y)\,.$$

Since $(\varphi_x, Q_k\varphi_y) = 0$ for $k < d(x,y)$, and $|(\varphi_x, Q_k\varphi_y)| \le 1$ always holds, (9.1) follows. □

{p.speed}

**Proposition 9.3.** *Consider a graph $G$ with degrees bounded by $D$, and a reversible Markov chain on its vertex set $V$, with not necessarily nearest-neighbour transitions. Assume that the reversible measure satisfies a uniform bound $\pi(x)/\pi(y) < B < \infty$ for any $x,y \in V$, and that $\rho < 1$. Then, in the graph metric,*

$$\liminf_{n\to\infty} \frac{\mathrm{dist}(o,X_n)}{n} > 0 \ a.s.$$

*Proof.* Choose $\alpha > 0$ small enough that $D^\alpha < 1/\rho$. Then,

$$\begin{aligned}\mathbf{P}_o[\mathrm{dist}(o,X_n) \le \alpha n] = \sum_{x\in B_{\alpha n}(o)} p_n(o,x) &\le |B_{\alpha n}(o)|\sqrt{B}\rho^n \\ &\le D^{\alpha n}\sqrt{B}\rho^n \le \sqrt{B}\exp(-c_\alpha n)\,.\end{aligned}$$

In the first line, we used part (a) of Lemma 9.1 and the bound on $\pi(\cdot)$. The second inequality holds because of the choice of $\alpha$. Thus, by Borel-Cantelli, $\mathrm{dist}(o,X_n) \le \alpha n$ holds only finitely many times, and $\liminf_{n\to\infty} \frac{\mathrm{dist}(o,X_n)}{n} > \alpha$ almost surely. □

{l.speedexists

**Lemma 9.4.** *For any group-invariant random walk on a group (with not necessarily nearest neighbour jumps in the Cayley graph that gives the metric),*

$$\lim_{n\to\infty} \mathbf{E}\,\frac{\mathrm{dist}(o,X_n)}{n} \ \textit{exists.}$$

*Proof.* First we define the sequence $(a_n)_{n=1}^\infty$ by $a_n = \mathbf{E}\,\mathrm{dist}(o,X_n)$. We can see that this sequence is subadditive:

$$\begin{aligned}\mathbf{E}\,\mathrm{dist}(o,X_{n+m}) &\le \mathbf{E}\,\mathrm{dist}(o,X_n) + \mathbf{E}\,\mathrm{dist}(X_n,X_{n+m}) \\ &\le \mathbf{E}\,\mathrm{dist}(o,X_n) + \mathbf{E}\,\mathrm{dist}(o,X_m)\,.\end{aligned}$$

So we may apply Fekete's lemma, which gives us that $\lim_{n\to\infty} a_n/n$ exists. □

A stronger statement also holds: $\lim_{n\to\infty} \frac{\mathrm{dist}(o,X_n)}{n}$ exists almost surely. This follows from Kingman's subadditive ergodic theorem; see, e.g., [LyPer16, Theorems 14.10 and 14.44].

We have seen in Proposition 9.3 that non-amenability implies positive speed. The converse is false, as the $d \ge 3$ case of the following theorem shows. For the definition of the lamplighter groups, see the end of Section 5.1.

{t.LLspeed}

**Theorem 9.5.** *For the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$, the following bounds hold for $d_n = \mathrm{dist}(o, X_n)$:*

**(a)** $d = 1$ *has* $\mathbf{E}[d_n] \asymp \sqrt{n}$

**(b)** $d = 2$ *has* $\mathbf{E}[d_n] \asymp \frac{n}{\log n}$

**(c)** $d \geq 3$ *has* $\mathbf{E}[d_n] \asymp n$.

*In each case, we will prove here the statement only for some specific generating set.*

{q.speedinv}

**Question 9.6.** *Is it true that positive speed never depends on the finite symmetric generating set? The case for $\mathbb{Z}_2 \wr \mathbb{Z}^d$ is known [Ers04a].*

*Proof.* In each case, we will consider the walk given by staying put with probability 1/4, switching the lamp at the present location with probability 1/4, and moving to one of the $2d$ neighbours with probability 1/(4d) each.

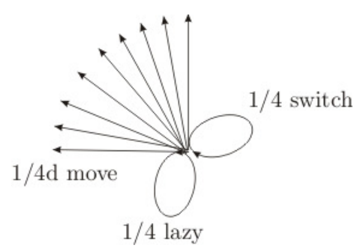



Figure 9.1: The generators for the walk on $\mathbb{Z}_2 \wr \mathbb{Z}^d$. {f.LLd}

Then, if $X_n = (L_n, \phi_n)$, where $L_n$ is the position of the lamplighter and $\phi_n$ is the configuration of the lamps, then it is easy to give some crude bounds on $d_n$: it is at least $|\operatorname{supp} \phi_n|$, the number of lamps on, plus the distance from the origin of the farthest lamp on, and it is at most the number of steps in any procedure that visits all the lamps that are on, switches them off, then goes back to the origin. Moreover, the expected size of $\operatorname{supp} \phi_n$ can be estimated based on the expected size of the range $R_n$ of the movement of the lamplighter, using the following exercise:

{ex.lastvisit}

▷ **Exercise 9.1.** *Show that $\mathbf{P}\big[\, x \in \operatorname{supp} \phi_{n+1} \,\big|\, x \in R_n \,\big] \geq 1/4$.*

Therefore, we have

$$\frac{\mathbf{E}\,|R_n|}{4} \leq \mathbf{E}\,|\operatorname{supp} \phi_{n+1}| \leq \mathbf{E}\,|R_n|\,, \tag{9.2}$$ {e.litrange}

so estimating $\mathbf{E}\,|R_n|$ will certainly be important.

Now, in the $d = 1$ case, we have

$$|\min \operatorname{supp} \phi_n| \vee |\max \operatorname{supp} \phi_n| \leq d_n \leq |L_n| + 3|\min \operatorname{supp} \phi_n| + 3|\max \operatorname{supp} \phi_n|\,, \tag{9.3}$$ {e.LLd1}

using that the number of lamps on is at most $|\min \operatorname{supp} \phi_n| + |\max \operatorname{supp} \phi_n|$. By the Central Limit Theorem, $\mathbf{P}[\,|L_n| > \epsilon\sqrt{n}\,] > \epsilon$ for some absolute constant $\epsilon > 0$. Thus, by Exercise 9.1, we have

$$\mathbf{P}\Big[\operatorname{supp} \phi_n \not\subseteq [-\epsilon\sqrt{n}, \epsilon\sqrt{n}]\Big] > \epsilon/4$$

for $n \geq 2$. Now, by the left side of (9.3), we have

$$\begin{aligned}\mathbf{E}[\, d_n \,] &\geq \mathbf{E}\Big[\,|\min \operatorname{supp} \phi_n| \vee |\max \operatorname{supp} \phi_n|\,\Big] \\ &\geq \mathbf{P}\Big[\operatorname{supp} \phi_n \not\subseteq [-\epsilon\sqrt{n}, \epsilon\sqrt{n}]\Big]\epsilon\sqrt{n} \;\geq\; \epsilon^2\sqrt{n}/4\,,\end{aligned}$$

which is a suitable lower bound.

For an upper bound, let $M_n = \max_{k\le n} |L_k|$. Clearly, $|\min \operatorname{supp} \phi_n| \vee |\max \operatorname{supp} \phi_n| \le M_n$. From the CLT, it looks likely that not only $L_n$, but even $M_n$ cannot be much larger than $\sqrt{n}$, but if we do not want to lose $\log n$-like factors, we have to be slightly clever. In fact, we claim that

$$\mathbf{P}[\, M_n \ge t\,] \le 4\,\mathbf{P}[\, L_n \ge t\,]\,. \tag{9.4}$$

{e.maxreflecti

This follows from a version of the reflection principle. If $T_t$ is the first time for $L_n$ to hit $t \in \mathbb{Z}_+$, then, by the strong Markov property and the symmetry of the walk restarted from $t$,

$$\mathbf{P}[\, T_t \le n, L_n \ge t\,] \ge \frac{1}{2}\,\mathbf{P}[\, T_t \le n\,]\,;$$

we don't have exact equality because of the possibility of $L_n = t$. Since $\{L_n \ge t\} \subset \{T_t \le n\}$, this inequality can be rewritten as

$$2\,\mathbf{P}[\, L_n \ge t\,] \ge \mathbf{P}[\, T_t \le n\,] = \mathbf{P}\big[\max_{k\le n} L_k \ge t\,\big].$$

Since $\mathbf{P}[\, M_n \ge t\,] \le 2\,\mathbf{P}[\,\max_{k\le n} L_k \ge t\,]$, we have proved (9.4).

Summing up (9.4) over $t$, we get $\mathbf{E}[\,|M_n|\,] \le 2\,\mathbf{E}[\,|L_n|\,] \le C\,\sqrt{n}$ for some large $C < \infty$, from the CLT. Using (9.3), we see that $\mathbf{E}[\, d_n\,] \le C'\,\sqrt{n}$, and we are done.

Now consider the case $d = 2$. For the lamplighter, we have $p_k(0,0) \asymp 1/k$. (This comes from the general fact that $p_k(0,0) \asymp k^{-d/2}$ on $\mathbb{Z}^d$). Thus, $\sum_{k=1}^n p_k(0,0) \asymp \log n$. But $\sum_{k=1}^n p_k(0,0)$ is exactly the expected number of visits to 0 by time $n$. This suggests that once a point is visited, it is typically visited roughly $\log n$ times in expectation, and thus the walk visited roughly $n/\log n$ different points. In fact, it is not hard to prove that

$$\mathbf{E}\,|R_n| \asymp \frac{n}{\log n}\,. \tag{9.5}$$

{e.2dRn}

▷ **Exercise 9.2.** *Prove the above statement.*

The precise asymptotics in the above statement is also known, a classical theorem by Erdős and Dvoretzky.

Since, to get from $X_n$ to the origin, we have to switch off all the lamps, by (9.2) we have $\mathbf{E}\,|R_n|/4 \le \mathbf{E}\,d_n$. On the other hand, $R_n$ is a connected subset of $\mathbb{Z}^2$ containing the origin, so we can take a spanning tree in it, and wherever $L_n \in R_n$ is, can go around the tree to switch off all the lamps and return to the origin. This can be done in less than $3|R_n|$ steps in $\mathbb{Z}^2$, while the switches take at most $|R_n|$ steps, hence $d_n \le 4|R_n|$. Altogether, $\mathbf{E}\,d_n \asymp \mathbf{E}\,|R_n|$, so the $d = 2$ case follows from (9.5).

For the case $d \ge 3$, consider the following exercise:

▷ **Exercise 9.3.** *For simple random walk on a transitive graph, we have:*

$$\lim_{n\to\infty} \frac{\mathbf{E}\,|R_n|}{n} = q := \mathbf{P}_o[\textit{never return to } o]\,.$$

From this exercise and (9.2), we can deduce that:

$$n \ge \mathbf{E}[d_n] \ge \mathbf{E}\,|\operatorname{supp}\phi_n| \ge \frac{\mathbf{E}\,|R_n|}{4} \ge \frac{qn}{4},$$

thus proving the case $d \ge 3$. □

The distinction between positive and zero speed will be a central topic of the present chapter. And, of course, there are also finer questions about the rate of escape than just being linear or sublinear, which will be addressed in later chapters.

## 9.2 The Liouville and strong Liouville properties for harmonic functions

{ss.Liouville}

In this section we will start to examine what kind of discrete harmonic functions may exist for infinite Markov chains. As we saw in Section 6.3 (with an application in Section 8.3), harmonic functions are intimately related to martingales, hence we will make use of the following:

{d.suMG}

**Definition 9.7.** *A submartingale is a sequence* $(M_n)_{n=1}^{\infty}$ *such that the following two conditions hold:*

$$\forall n \ \mathbf{E}\,|M_n| < \infty\,,$$

$$\forall n \ \mathbf{E}[\, M_{n+1} \mid M_1, \ldots, M_n \,] \geq M_n\,.$$

*A supermartingale is a similar sequence, except that the second condition is replaced by*

$$\forall n \ \mathbf{E}[\, M_{n+1} \mid M_1, \ldots, M_n \,] \leq M_n\,.$$

{t.MGconverge}

**Theorem 9.8** (Martingale Convergence Theorem)**.** *Let* $(M_n)_{n=1}^{\infty}$ *be a submartingale such that* $\sup_n \mathbf{E}\big[\, M_n \mathbf{1}_{\{M_n>0\}}\,\big] < \infty$*. (For instance, there exists some constant $B$ with $M_n \leq B$ almost surely.). Then there exists a random variable $M_\infty$ with*

$$M_n \to M_\infty \ a.s.$$

For a proof, see [Dur10, Theorem 5.2.8].

{c.nobddharm}

**Corollary 9.9.** *A recurrent Markov chain has no nonconstant bounded harmonic functions, not even harmonic functions bounded from one direction (equivalently, positive ones).*

*Proof.* If $f$ is a positive harmonic function, then $M_n = f(X_n)$ is a martingale (in particular, a supermartingale) bounded from below. If $f$, in addition, is nonconstant, then there are $x$ and $y$ states such that $f(x) \neq f(y)$. The walk is recurrent, so returns to these states infinitely often. So $\forall N\ \exists m, n > N$ such that $M_m = f(x)$ and $M_n = f(y)$. So $M_n$ cannot converge to any value, which contradicts Theorem 9.8. So no such $f$ can exist. □

{t.LiouProp}

**Theorem 9.10.** *In any Markov chain, if $x, y \in V$ are such that there is a coupling $(X_n, Y_n)_{n\geq 0}$ of random walks starting from $(x, y)$ such that $\mathbf{P}[\, X_n \neq Y_n\,] \to 0$, then any bounded harmonic function has the same value on $x$ and $y$. In particular, if every pair $x, y \in V$ has this coupling property, then the chain has the Liouville property, i.e., every bounded harmonic function is constant.*

*Proof.* Suppose $f$ is a bounded harmonic function, say $|f| < B$. Then $f(X_n)$ and $f(Y_n)$ are both martingales, with $\mathbf{E}[\, f(X_n)\,] = f(x)$ and $\mathbf{E}[\, f(Y_n)\,] = f(y)$. Thus,

$$\begin{aligned} \forall n \ |f(x) - f(y)| &= |\mathbf{E} f(X_n) - \mathbf{E} f(Y_n)| \\ &\leq \mathbf{E}\,|f(X_n) - f(Y_n)| \\ &\leq \mathbf{P}[\, X_n \neq Y_n\,]\, 2B\,. \end{aligned}$$

Therefore, $|f(x) - f(y)| \leq \lim_{n\to\infty} \mathbf{P}[\, X_n \neq Y_n\,]\, 2B = 0$, so $f$ must be constant. □

{t.Blackwell}

**Theorem 9.11** (Blackwell 1955)**.** *$\mathbb{Z}^d$ has the Liouville property.*

*Proof.* Clearly, $f$ is harmonic with respect to $P$ if and only if it is harmonic with respect to $\frac{I+P}{2}$, the lazy version of $P$. So, consider the lazy random walk in $\mathbb{Z}^d$ given by flipping a $d$-sided coin to determine which coordinate to move in and then using the lazy simple random walk on $\mathbb{Z}$ in that coordinate. This is the product chain of lazy walks coordinatewise. Now consider the following coupling: $X_n$ and $Y_n$ always move in the same coordinate as one another. If their distance in that coordinate is zero, then they move (or remain still) together in that coordinate. If not, then when $X$ moves, $Y$ stays still and when $X$ stays still, $Y$ moves. Each of $X$ and $Y$ when considered independently is still using the lazy random walk described.

Now, considering one coordinate at a time, if $X_n$ and $Y_n$ have not yet come together in that coordinate, then whenever that coordinate is chosen, the distance goes up by one with probability 1/2 and down by one with probability 1/2. This is equivalent to a simple random walk on $\mathbb{Z}$, which is recurrent, so with probability 1, $X_n$ and $Y_n$ will eventually have distance zero in that coordinate, and then will remain together for the rest of the coupling. This happens in each coordinate, so with probability 1, $\exists N$ such that $X_n = Y_n \ \forall n \geq N$. So by Theorem 9.10, this chain has the Liouville property. $\square$

▷ **Exercise 9.4.*** *By studying how badly the coupling may fail, show that any harmonic function $f$ on $\mathbb{Z}^d$ with sublinear growth, i.e., satisfying $\lim_{\|x\|_2\to\infty} f(x)/\|x\|_2 = 0$, must be constant.* {ex.sublin}

▷ **Exercise 9.5.** *Using couplings, show that any random walk on $\mathbb{Z}^d$ with symmetric bounded jumps has the Liouville property.*

In fact, the Choquet-Deny theorem [ChoD60] says that *any* generating measure on any Abelian group has the Liouville property. See Theorem 9.30 in Section 9.6 below for a proof that can be generalized for nilpotent groups. Generalizing in a different direction, the Choquet-Deny proof also implies the following; see [Saw97]:

{t.ZdMartin}

**Theorem 9.12** (Strong Liouville property for $\mathbb{Z}^d$)**.** *For the random walk generated by any probability measure $\mu$ on $\mathbb{Z}^d$ whose support generates all of $\mathbb{Z}^d$ and has the properties that*

$$\sum_{x\in\mathbb{Z}^d} \mu(x)\,|x| < \infty \qquad \text{and} \qquad \sum_{x\in\mathbb{Z}^d} \mu(x)\,x = 0 \in \mathbb{R}^d\,, \tag{9.6}$$ {e.ChDsymm}

*any nonnegative harmonic function is constant.*

*Proof.* The main tool will be the classical Krein-Milman theorem (1940): any compact convex subset $K$ of a locally convex Hausdorff topological vector space $X$ is the closed convex hull of its extreme points: $K = \overline{\mathrm{conv}(\mathrm{ext}(K))}$. Based on our intuition from finite-dimensional vector spaces, this sounds like a triviality, but we can see its power from the fact how easily it implies this result, stronger than Theorem 9.11.

Let $X$ be the vector space of real $\mu$-harmonic functions on $\mathbb{Z}^d$, with the topology of pointwise convergence, and let

$$K := \{h \in X : h \geq 0, \text{ and } h(0) = 1\}\,.$$

This is obviously convex. Furthermore, observe that, since the support of $\mu$ generates $\mathbb{Z}^d$, harmonicity and the conditions $h \geq 0$ and $h(0) = 1$ imply that for each $x \in \mathbb{Z}^d$ there exists a bound $H_\mu(x)$ such that $h(x) \leq H_\mu(x)$ for every $h \in K$. Thus $K$ is compact by Tychonoff's theorem. Let $h$ be an extreme point of $K$; by the Krein-Milman theorem, it is enough to prove that $h$ is constant. Again using that the support of $\mu$ generates $\mathbb{Z}^d$, the maximum principle implies that $h(x) > 0$ for all $x \in \mathbb{Z}^d$. Then, the definition of harmonicity can be written as

$$h(x) = \sum_{s\in\mathbb{Z}^d} \mu(s)h(s)\frac{h(x+s)}{h(s)}\,, \tag{9.7}$$ {e.hconvcomb}

for any $x \in \mathbb{Z}^d$. Since $\mathbb{Z}^d$ is Abelian, the function $x \mapsto h_s(x) := h(x+s)/h(s)$ is again harmonic, with $h_s(0) = 1$, and hence the RHS of (9.7) is a convex combination of elements of $K$. (Note here that (9.7) with $x = 0$ says that $\sum_s \mu(s)h(s) = 1$.) Since $h$ is extremal, we must have $h_s(x) = h(x)$ for each $s$ with $\mu(s) > 0$. By $\mu$ being a generating measure of $\mathbb{Z}^d$, this implies that $h(x+y) = h(x)h(y)$ for any $x, y \in \mathbb{Z}^d$, and thus $h$ is of the form $h(x) = \exp(\alpha \cdot x)$ for some $\alpha \in \mathbb{Z}^d$.

Now, assuming (9.6), we apply Jensen's inequality to the $x = 0$ case of (9.7):

$$1 = \sum_{s \in \mathbb{Z}^d} \mu(s) \exp(\alpha \cdot s) \geq \exp\left( \alpha \cdot \sum_{s \in \mathbb{Z}^d} \mu(s) s \right) = 1\,.$$

Since $x \mapsto h(x) = \exp(\alpha \cdot x)$ is strictly convex, we can have the above equality in Jensen only if $h(s)$ is constant 1 on the entire support of $\mu(s)$. Thus $\alpha \cdot s = 0$ for every $s$ in the support, while the support must span $\mathbb{R}^d$ if it is a generating set, and hence $\alpha = 0$, and $h(x) = 1$ for all $x \in \mathbb{Z}^d$. □

In Theorem 9.10 and its application Theorem 9.11, we deduced the Liouville property from a coupling property of the random walk. Can we use something similar instead of the Krein-Milman theorem to prove Theorem 9.12? For smooth positive harmonic functions on $\mathbb{R}^d$, the task would be easy: harmonicity of $f$ implies that $f(x) = \mathbf{E}_x f(B_t)$ holds for any $t \geq 0$, for Brownian motion $B_t$ started at $B_0 = x$; on the other hand, the densities of $B_t$ for $B_0 = x$ and $B_0 = y$ are within a factor of $1 + o(1)$ from each other as $t \to \infty$ (starting from $t \gg \|x - y\|_2^2$), hence the same holds for the expectations, and we get $f(x) = f(y)$ in the limit. For $\mathbb{Z}^d$, the discreteness causes some technical difficulties: e.g., it is not true that $\sup_{z \in \mathbb{Z}^d} \big(p_n(x,z))/p_n(y,z) - 1\big) \to 0$ as $n \to \infty$, since for any $n$ there is a $z$ for which $p_n(x,z) = 0 < p_n(y,z)$. One possibility to resolve this issue is to use the Local Central Limit Theorem for the random walk at times $n$ and $n(1+\epsilon)$; another one is to use the divisible sandpile model, Theorem 3.3 of [LevP09].

▷ **Exercise 9.6.** *Use either of the above suggestions to give a probabilistic proof of Theorem 9.12.*

{ex.LiouSLiou}

▷ **Exercise 9.7.** *Consider an irreducible Markov chain $(V, P)$.*

**(a)** *Assume that $d_{\mathrm{TV}}\big(p_n(x,\cdot)), p_n(y,\cdot)\big) \to 0$ as $n \to \infty$, for any $x, y \in V$. Show that $(V, P)$ has the Liouville property.*

**(b)** *Show that biased nearest-neighbor random walk on $\mathbb{Z}$ has the property of part (a), but nevertheless it does not have the strong Liouville property: it has non-constant positive harmonic functions.*

The asymmetry of the walk in part (b) of the previous exercise made it easy to distinguish between the Liouville and strong Liouville properties. We will see an example below where even the symmetric simple random walk on a group has the Liouville but not the strong Liouville property. But, before that, let us see the first example that does not have the Liouville property: the $d$-regular tree $\mathbb{T}_d$ with $d \geq 3$.

Consider Figure 9.2, and let $f(y) = \mathbf{P}_y[\{x_n\}$ ends up in $A]$. By transience and by any vertex being a cutpoint of the tree, $\lim_{n\to\infty} \mathbf{1}_{x_n \in A}$ exists a.s, so this definition makes sense. It is obviously a bounded harmonic function. Let $p = \mathbf{P}_x[\text{never hit } o] \in (0,1)$, then $f(x) = (1-p)f(o)$. This shows that $f$ is nonconstant.

▷ **Exercise 9.8.** *Show that the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}^d$ with $d \leq 2$ has the Liouville property.*

▷ **Exercise 9.9.** *Show that the lamplighter group with $d \geq 3$ does not have the Liouville property.*

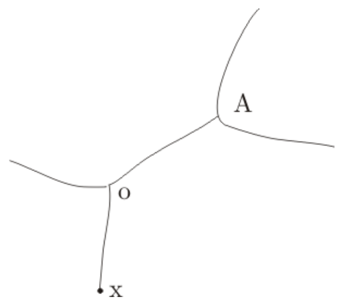



Figure 9.2: Portion of a $d$-regular tree ($d = 3$) {f.dreg}

A hint for these two exercises is that positive speed and the existence of non-trivial bounded harmonic functions will turn out to be intimately related issues, so you may look at the proof of Theorem 9.5 to get an idea how to construct good couplings in the case $d \le 2$ and how to construct nonconstant bounded harmonic functions in the case $d \ge 3$.

{t.LLLipHarm}

**Theorem 9.13** ([BenDCKY]). *Consider SRW on the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}$ with one of the usual finite generating sets. Then the strong Liouville property completely fails: there are infinitely many linearly independent positive Lipschitz harmonic functions.*

*Idea of proof.* As in Theorem 9.5, the random walk trajectory will be denoted by $X_n = (M_n, \phi_n)$. Consider a fixed configuration $\chi$ of lamps on $-\mathbb{N}$ with finitely many lamps on; for simplicity, we will take now all the lamps to be off. Let $\tau_R$ be the hitting time of $-R$ or $R$ by the lamplighter, and consider

$$g_R^\chi(x) := \mathbf{P}_x\left[\, \phi_{\tau_R}\big|_{-\mathbb{N}} = \chi \,\right].$$

It is not hard to see that for $R$ large enough, $g_R^\chi(x)$ is comparable (up to constant factors depending on $\chi$) to the probability of not visiting before $\tau_R$ any negative integer that is less than $\min \operatorname{supp} \chi$. This probability is on the order of $1/R$ (with constant factors depending on $\chi$ and $x$), and it should not be surprising that the limit

$$g^\chi(x) := \lim_{R\to\infty} R \cdot g_R^\chi(x)$$

exists, non-constant, and Lipschitz in $x$. And, of course, it is harmonic. Furthermore, one can show that for different configurations $\chi$, these functions are all linearly independent. $\square$

▷ **Exercise 9.10.** *Fill in some of the gaps in the above proof.*

{ex.sublinLL}

▷ **Exercise 9.11.** *Similarly to Exercise 9.4 about $\mathbb{Z}^d$, show that the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}$ has no non-constant harmonic functions of sublinear growth.*

In Exercises 9.4 and 9.11, the triviality of sublinear growth harmonic functions followed from good couplings of the random walk. One may speculate that the existence of such good couplings is intimately related to the walk not being too much spread out. This motivates the following courageous conjecture, with both sides being open (beyond the case of almost nilpotent groups, where both the diffusive behaviour and the triviality of sublinear growth harmonic functions are known [HeSC93, Theorem 6.1]):

**Conjecture 9.14** (Gady Kozma). *SRW on a group has no non-constant harmonic functions of sublinear growth iff the walk is diffusive, i.e., $\mathbf{E}\operatorname{dist}(X_0, X_n) \asymp \sqrt{n}$.*

It is important to mention here that the diffusive $\sqrt{n}$ rate of escape is the slowest possible on transitive graphs; see Section 10.2.

## 9.3 Entropy, and the main equivalence theorem

{ss.SpEnt}

As we have seen in the previous two sections, on the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$ there is a strong relationship between positive speed and the existence of non-trivial bounded harmonic functions. The main result of the current section is that this equivalence holds on all groups, see Theorem 9.17 below. However, a transparent probabilistic reason currently exists only on the lamplighter groups. For the general case, the proof goes through more ergodic/information theoretic notions like entropy and tail-$\sigma$-fields. We start by discussing entropy.

Let $\mu$ be a finitely supported probability measure on a group $\Gamma$ such that $\operatorname{supp}\mu$ generates $\Gamma$. Then we get a nearest neighbour random walk on the directed right Cayley graph of $\Gamma$ given by the generating set $\operatorname{supp}\mu$:

$$\mathbf{P}[\, x_{n+1} = h \mid x_n = g \,] = \mu(g^{-1}h)\,,$$

where $h, g \in \Gamma$. Then the law of $X_n$ is the $n$-fold convolution $\mu^n = \mu^{*n}$, where $(\mu * \nu)(g) := \sum_{h\in\Gamma} \mu(h)\nu(h^{-1}g)$. This walk is reversible iff $\mu$ is symmetric, i.e., $\mu(g^{-1}) = \mu(g)$ for all $g \in \Gamma$. All our previous random walks on groups were examples of such walks.

Recall the notion of entropy $H(\mu)$ from Definition 1.10 in Section 1.2.

**Definition 9.15.** *The* **asymptotic entropy** $h(\mu)$ *of the random walk generated by* $\mu$ *is defined by:*

$$h(\mu) := \lim_{n\to\infty} \frac{H(\mu^n)}{n} = \lim_{n\to\infty} \frac{H(X_n)}{n}\,.$$

▷ **Exercise 9.12.** *Show that* $h(\mu)$ *exists for any* $\mu$ *on a group* $\Gamma$ *that satisfies* $H(\mu) < \infty$.

As a corollary to the fact $H(\mu) \le \log|\operatorname{supp}\mu|$, if $\Gamma$ has sub-exponential volume growth (in the directed Cayley graph given by the not necessarily symmetric $\operatorname{supp}\mu$), then $h(\mu) = 0$.

Here is a fundamental theorem that helps comprehend what asymptotic entropy means:

{t.SMB}

**Theorem 9.16** (Shannon-McMillan-Breiman, version by Kaimanovich-Vershik)**.** *For almost every random walk trajectory* $\{X_n\}$,

$$\lim_{n\to\infty} \frac{-\log\mu^n(X_n)}{n} = h(\mu).$$

To see why this theorem should be expected to hold, note that the expectation of $-\log\mu^n(X_n)$ is exactly $H(\mu^n)$, hence the sequence converges in expectation to $h(\mu)$ by definition. So, this theorem is an analogue of the Law of Large Numbers. For a proof, see [KaiV83].

▷ **Exercise 9.13.** $h(\mu) = 0$ *iff* $\forall\epsilon > 0$ *there exists a sequence* $\{A_n\}$ *with* $\mu^n(A_n) > 1-\epsilon$ *and* $\log|A_n| = o(n)$. *In words, zero entropy means that the random walk is mostly confined to a sub-exponentially growing set.*

We now state a central theorem of the theory of random walks on groups. We will define the invariant $\sigma$-field (also called the Poisson boundary) only in the next section, so let us give here an intuitive meaning: it is the set of different places where infinite random walk trajectories can escape (with the natural measure induced by the random walk, called the harmonic measure at infinity). A great introduction to the Poisson boundary is the seminal paper [KaiV83].

{t.SEHI}

**Theorem 9.17.** *For any symmetric finitely supported random walk on a group, the following are equivalent:*

- **(S)** *positive speed $\sigma(\mu) > 0$,*
- **(E)** *positive asymptotic entropy $h(\mu) > 0$,*
- **(H)** *existence of non-constant bounded harmonic functions (the non-Liouville property),*
- **(I)** *non-triviality of the invariant $\sigma$-field (Poisson boundary).*

Here are some two-sentence summaries of the proofs:

The meaning of (S) $\iff$ (E) is that positive speed for a reversible walk is equivalent to the walk being very much spread out. This will be proved in this section. Note that the reversibility of the walk (in other words, the symmetry of $\mu$) is important: consider, e.g., biased random walk on $\mathbb{Z}$, which has positive speed but zero entropy.

The equivalence (I) $\iff$ (H) holds for any Markov chain, and will be proved in Section 9.4: a bounded harmonic function evaluated along a random walk trajectory started at some vertex $x$ is a bounded martingale, which has an almost sure limit whose expectation is the value at $x$. Therefore, from a bounded harmonic function we can construct a function "at infinity", and vice versa.

Finally, (E) $\iff$ (I) will be proved in Section 9.5, by expressing the asymptotic entropy as the amount of information gained about the first step by knowing the limit of the random walk trajectory, see (9.12). Hence positive entropy means that non-trivial information is contained in the boundary. This equivalence holds also for non-symmetric measures $\mu$, but it is important for the walk to be group-invariant, as will be shown by an example of a non-amenable bounded degree graph that has positive speed and entropy, but has the Liouville property and trivial Poisson boundary.

Beyond the case of lamplighter groups discussed in the previous sections, the following is open:

▷ **Exercise 9.14.** *** *Find a direct probabilistic proof of (S) $\iff$ (H) for symmetric bounded walks on any group. Is there a quantitive relationship between the speed and the amount of bounded harmonic functions, e.g., their von Neumann dimension, defined later? (It is clear from the discussion of the different equivalences above that both the symmetry and the group invariance of the walk is needed.)*

For symmetric finitely supported measures with support generating a non-amenable group, from Proposition 9.3 we know that the walk has positive speed and hence non-trivial Poisson boundary. The importance of the lamplighter group examples was pointed out first by [KaiV83]: $\mathbb{Z}_2 \wr \mathbb{Z}^d$ shows that exponential volume growth is not enough ($d \le 2$) and non-amenability is not needed ($d \ge 3$) for positive speed. Nevertheless, it is worth comparing non-amenability and non-trivial Poisson boundary: the first means that $\mu^n(1)$ decays exponentially (Theorem 7.3), while the second means that $\mu^n(X_n)$ decays exponentially (by Theorem 9.16). We will see a characterization of amenability using the Poisson boundary in Theorem 9.29 below.

Now, we prove (S) $\iff$ (E) by giving the following quantitative version:

{t.VVSE}

**Theorem 9.18** (Varopoulos [Var85b] and Vershik [Ver00])**.** *For any symmetric finitely supported measure $\mu$ on a group,*

$$\log \frac{1}{\rho(\mu)} + \sigma(\mu)^2/2 \le h(\mu) \le \nu(\mu)\,\sigma(\mu)\,,$$

*where $\rho(\mu)$ is the spectral radius, $\sigma(\mu)$ is the speed of the walk, and $\nu(\mu) := (\lim_n \log |B_n^\mu|)/n$ is the exponential rate of volume growth of the graph (where the limit exists because $|B_{n+m}| \le |B_n| \cdot |B_m|$.) Both for the speed and the volume, distances are given by the generating set* supp *$\mu$.*

*Proof.* For the lower bound, the key ingredient is the Carne-Varopoulos bound, Theorem 9.2, which implies:

$$\frac{-\log p_n(o, X_n)}{n} \geq \log\frac{1}{\rho} + o(1) + \frac{\mathrm{dist}(o, X_n)^2}{2n^2}\,.$$

Taking expectation w.r.t. $X_n$, we get

$$\frac{H(X_n)}{n} \geq \log\frac{1}{\rho} + \frac{\mathbf{E}\big[\,\mathrm{dist}(o, X_n)^2\,\big]}{2n^2} + o(1) \geq \log\frac{1}{\rho} + o(1) + \frac{\big(\mathbf{E}\,\mathrm{dist}(o, X_n)\big)^2}{2n^2}\,.$$

The LHS tends to $h(\mu)$, the RHS tends to $\log\frac{1}{\rho} + \sigma(\mu)^2/2$, hence we are done.

For the upper bound, the key idea is that entropy on a finite set is maximized by the uniform measure. Thus, using that the sphere $|\partial B_r(o)|$ has size at most $\exp\big((\nu + o(1))r\big)$ as $r \to \infty$, and assuming that $\mathrm{dist}(o, X_n) \to \infty$ almost surely,

$$H(X_n) = H(\mathrm{dist}(o, X_n)) + H(X_n \,|\, \mathrm{dist}(o, X_n)) \leq \log n + \mathbf{E}\log\exp\big((\nu + o(1))\mathrm{dist}(o, X_n)\big)\,.$$

Dividing by $n$, the LHS converges to $h(\mu)$, while the RHS converges to $\nu(\mu)\,\sigma(\mu)$. In case $\mathrm{dist}(o, X_n) \to \infty$ does not hold, i.e., the walk is recurrent, then the group has at most two-dimensional volume growth, so $H(X_n)$ can grow at most logarithmically, and $h(\mu) = 0$ follows. □

{ex.funda}

▷ **Exercise 9.15** (Vershik).*** *The upper bound in the previous theorem is called "the fundamental inequality" in [Ver00]. Does there exist for every finitely generated group a finitely supported $\mu$ with equality here, $h(\mu) = \nu(\mu)\,\sigma(\mu)$? The trouble with a strict inequality is that then the random walk measure is far from uniform on the sphere where it is typically located, i.e., sampling from the group using the random walk is not a good idea.*

As we mentioned in Question 9.6, it is not known if the positivity of speed on a group is independent of the Cayley graph, or more generally, invariant under quasi-isometries. On general graphs, the speed does not necessarily exist, but the Liouville property can clearly be defined. However, for the class of bounded degree graphs it is known that the Liouville property is *not* quasi-isometry invariant [Lyo87, Ben91]. A simple example is given in [BenS96a], which we now describe:

The set $Y = \{0,1\}^*$ of finite 0-1-words can naturally be viewed as an infinite binary tree. Let $A$ be the subset of words in which the ratio of 1's among the letters is more than $2/3$. Now consider the lattice $\mathbb{Z}^4$, take a bijection between the vertices along the $x$-axis of $\mathbb{Z}^4$ and the vertices in $A$, and put an edge between each pair of vertices given by this bijection. This graph will be $G$. It easily follows from the law of large numbers that SRW on $Y$ spends only finite time in $A$, and hence SRW on $G$, started from any $y \in Y$, will ever visit the $\mathbb{Z}^4$ part only with a probability bounded away from 1. Whenever the walk visits $\mathbb{Z}^4$, by the transience of the $\mathbb{Z}^3$ direction orthogonal to the $x$-axis, with positive probability it will eventually stay in $\mathbb{Z}^4$. So, the function $v \mapsto \mathbf{P}_v[\,\text{the walk ends up in } \mathbb{Z}^4\,]$ for $v \in G$ is a non-trivial bounded harmonic function. Now, let $Y'$ be the tree where each edge $(\underline{w}, \underline{w}0)$ of $Y$ is replaced by a path of length $k$, for some large but fixed $k \in \mathbb{N}$, and $G'$ is the same join of $Y'$ and $\mathbb{Z}^4$ as before. If $k$ is large enough, then SRW on $Y'$ will visit $A$ infinitely often, and hence it will enter $\mathbb{Z}^4$ infinitely often, and hence will almost surely end up in $\mathbb{Z}^4$. By the coupling results Theorem 9.10 and Theorem 9.11 for the triviality of bounded harmonic functions, this means that $G'$ has the Liouville property. However, $G$ and $G'$ are obviously quasi-isometric to each other.

## 9.4 Liouville and Poisson

{ss.LioPoi}

Having proved (S) $\iff$ (E), we now turn to harmonic functions, the invariant sigma-field, and the equivalence (I) $\iff$ (H). Let us start with a basic result of probability theory:

{t.Levy01}

**Theorem 9.19** (Lévy's Zero-One Law)**.** *Given a filtration $\mathcal{F}_n \uparrow \mathcal{F}_\infty$, and $X$ almost surely bounded, we have that:*

$$\mathbf{E}[\, X \mid \mathcal{F}_n \,] \to \mathbf{E}[\, X \mid \mathcal{F}_\infty \,] \ a.s.$$

This is called a zero-one law since, for the case $X = \mathbf{1}_A$ with $A \in \mathcal{F}_\infty$, it says

$$\mathbf{P}[\, A \mid \mathcal{F}_n \,] \to \mathbf{1}_A \text{ a.s.},$$

hence the conditional probabilities $\mathbf{P}[\, A \mid \mathcal{F}_n \,]$ converge to 0 or 1.

*Proof.* This is simply a special case of the Martingale Convergence Theorem 9.8. To see this, recall from Section 6.3 that $M_n := \mathbf{E}[X|\mathcal{F}_n]$ is a bounded martingale, and then apply the theorem. $\square$

Although this result might seem obvious, it is quite powerful: for example, it implies the following fundamental theorem, to be used later in this section and in Section 12.1.

{t.Kol01}

**Theorem 9.20** (Kolmogorov's 0-1 law)**.** *Let $X_1, X_2, \dots$ be independent random variables and $A$ a tail event (i.e., it is contained in the* **tail-$\sigma$-field** *$\bigcap_{n\geq 1} \sigma\{X_n, X_{n+1}, \dots\}$, or in plain words, its occurrence is not influenced by any finite subset of the variables). Then $\mathbf{P}[\, A \,] = 0$ or $1$.*

*Proof using martingales.* For each $n$, $A$ is independent of $\mathscr{F}_n$, so $\mathbf{E}\big[\, 1_A \bigm| \mathscr{F}_n \,\big] = \mathbf{P}[\, A \,]$. As $n \to \infty$, the left-hand side converges to $1_A$ almost surely, by Lévy's theorem. So $\mathbf{P}[\, A \,] = 1_A$ almost surely, and it follows that $\mathbf{P}[\, A \,] \in \{0, 1\}$. $\square$

*A down-to-Earth proof.* For each $n$, $A$ is measurable in the product sigma-algebra $\sigma\{X_n, X_{n+1}, \dots\}$, hence, for any $\epsilon > 0$ there exists a $k = k(n, \epsilon)$ and a cylinder event $A_{n,\epsilon}$ measurable in $\sigma\{X_n, X_{n+1}, \dots, X_{n+k}\}$ such that $\mathbf{P}[\, A \triangle A_{n,\epsilon} \,] < \epsilon$. Now consider $\tilde{n} > n + k$, and then the event $A_{\tilde{n},\epsilon}$. Clearly, $A_{n,\epsilon}$ and $A_{\tilde{n},\epsilon}$ are independent. Thus,

$$\begin{aligned}
\Big|\mathbf{P}[\, A \,] - \mathbf{P}[\, A \,]^2\Big| &\leq \Big|\mathbf{P}[\, A \,] - \mathbf{P}[\, A_{n,\epsilon} \,]\, \mathbf{P}[\, A_{\tilde{n},\epsilon} \,]\Big| + \mathbf{P}[\, A \triangle A_{n,\epsilon} \,] + \mathbf{P}[\, A \triangle A_{\tilde{n},\epsilon} \,] \\
&\leq \Big|\mathbf{P}[\, A \,] - \mathbf{P}\big[\, A_{n,\epsilon} \cap A_{\tilde{n},\epsilon} \,\big]\Big| + 2\epsilon \\
&\leq \Big|\mathbf{P}[\, A \,] - \mathbf{P}\big[\, A \cap A \,\big]\Big| + \mathbf{P}[\, A \triangle A_{n,\epsilon} \,] + \mathbf{P}[\, A \triangle A_{\tilde{n},\epsilon} \,] + 2\epsilon \\
&\leq 4\epsilon\,.
\end{aligned}$$

This holds for every $\epsilon > 0$, hence $A$ is independent of itself, $\mathbf{P}[\, A \,] \in \{0, 1\}$. $\square$

Let $P$ be a transition matrix for a Markov chain on a countable state space $\mathcal{S}$. Let $\Omega$ be the measure space

$$\Omega = \Big\{ \{y_j\}_{j=0}^\infty : P(y_j, y_{j+1}) > 0,\ \forall j \Big\}\,,$$

with the usual Borel $\sigma$-field (the minimal $\sigma$-field that contains all the cylinders $\{\{y_j\}_{j=0}^\infty : y_0 = x_0, \dots, y_k = x_k\}$, for all $k$ and $x_0, \dots, x_k \in \mathcal{S}$), and the natural measure generated by $P$. Also, for any $x \in \mathcal{S}$, define the measure space

$$\Omega(x) = \Big\{ \{y_j\}_{j=0}^\infty : y_0 = x,\ P(y_j, y_{j+1}) > 0,\ \forall j \Big\}\,,$$

with the Borel $\sigma$-field. Now the measure generated by $P$ is a probability measure: the Markov chain trajectories started from $x$.

We define two equivalence classes. For $y, z \in \Omega$ let

$$y \overset{\mathcal{T}}{\sim} z \iff \exists n \quad \forall m \geq n \quad y_m = z_m ,$$

and

$$y \overset{\mathcal{I}}{\sim} z \iff \exists k, n \quad \forall m \geq n \quad y_m = z_{m+k} .$$

We identically define the equivalence for $\Omega(x)$ for any $x \in \mathcal{S}$. Using these equivalence relations, we now give an alternative definition of the **tail $\sigma$-field** $\mathcal{T}$ and also define the **invariant $\sigma$-field** $\mathcal{I}$ on $\Omega$ (and identically for $\Omega(x)$ for any $x \in \mathcal{S}$):

{d.tailinv}

**Definition 9.21.** *$A \in \mathcal{T}$ if and only if*

*(i) $A$ is Borel-measurable.*

*(ii) $y \in A$, $y \overset{\mathcal{T}}{\sim} z \implies z \in A$, i.e., $A$ is a union of tail equivalence classes.*

*$A \in \mathcal{I}$ if and only if*

*(i) $A$ is Borel-measurable.*

*(ii) $y \in A$, $y \overset{\mathcal{I}}{\sim} z \implies z \in A$, i.e., $A$ is a union of invariant equivalence classes.*

**Definition 9.22.** *A function $F : \Omega \longrightarrow \mathbb{R}$ is called a tail function if it is $\mathcal{T}$-measureable (i.e., $F$ is Borel measureable and $y \overset{\mathcal{T}}{\sim} z$ implies $F(y) = F(z)$). Similarly, $F$ is called an invariant function if it is $\mathcal{I}$-measureable.*

Of course, being an invariant function is a stronger condition than being a tail function. A key example of strict inequality is simple random walk on a bipartite graph $G(V, E)$ with parts $V = V_1 \cup V_2$: the event $A = \{y_{2n} \in V_1 \text{ for all } n\}$ is in the tail field but is not invariant.

It is also important to notice that although the jumps in the Markov chain are independent, Kolmogorov's 0-1 law does not say that tail or invariant functions are always trivial, since the $y_n$'s themselves are not independent. For instance, if the chain has three states, $\{0, 1, 2\}$, where 1 and 2 are absorbing, and $p(0, i) = 1/3$ for all $i$, then $B = \{y_n = 1 \text{ eventually}\}$ is an invariant event, with $\mathbf{P}_0[B] = 1/2$.

Call two tail or invariant functions $f, g : \Omega \longrightarrow \mathbb{R}$ equivalent if $\mathbf{P}_x[f = g] = 1$ for any $x \in \mathcal{S}$. Accordingly, we say that $\mathcal{T}$ or $\mathcal{I}$ is trivial if for any $x \in \mathcal{S}$ and any event $A$ in the $\sigma$-field, $\mathbf{P}_x[A] \in \{0, 1\}$. Note that if we consider only the measures $\mathbf{P}_x$, then the distinction between $\mathcal{T}$ and $\mathcal{I}$ in the above example of simple random walk on a bipartite graph disappears: for $x \in V_1$, we have $\mathbf{P}_x[\,A\,] = 1$, hence a $\mathbf{P}_x$-measure zero set can be added to $A$ to put it into $\mathcal{I}$, while for $x \in V_2$, we have $\mathbf{P}_x[\,A\,] = 0$, hence a $\mathbf{P}_x$-measure zero set can be subtracted from $A$ to put it into $\mathcal{I}$. The following theorem shows that for SRW on groups, a similar collapse always happens; on the other hand, the exercise afterwards shows that this is not a triviality.

{t.TI}

**Theorem 9.23.**

**(i)** *For simple random walk on a finitely generated group, with starting point $x \in \Gamma$, the invariant and tail sigma-fields $\mathcal{I}$ and $\mathcal{T}$ coincide up to $\mathbf{P}_x$-measure zero sets.*

**(ii)** *For any countable irreducible Markov chain $(V, P)$, the lazy version $\tilde{P} = (I + P)/2$ has invariant and tail sigma-fields $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{T}}$ coinciding up to $\mathbf{P}_x$-measure zero sets, for any $x \in V$.*

▷ **Exercise 9.16.** *Give an example of a graph $G = (V, E)$ and a vertex $x \in V$ such that for simple random walk started at $x$, the $\sigma$-fields $\mathcal{T}$ and $\mathcal{I}$ are not the same up to $\mathbf{P}_x$-measure zero sets.*

Part (ii) of Theorem 9.23 is the simpler to prove: the key idea is that the number of actual steps taken in $\tilde{P}^n$ has distribution $\mathsf{Binom}(n, 1/2)$, whose TV-distance from $\mathsf{Binom}(n+1, 1/2)$ converges to 0, hence $d_{\mathrm{TV}}\big(\tilde{p}_n(x,\cdot), \tilde{p}_{n+1}(x,\cdot)\big) \to 0$, as well. As one can easily imagine, this implies that the exact time that would make the difference between $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{T}}$ does not actually matter, hence $\tilde{\mathcal{I}} = \tilde{\mathcal{T}}$ under $\mathbf{P}_x$. The analogous ingredient for part (i) is the following result, whose proof we omit, although the Reader is invited to think about it a little bit:

{t.D02}

**Theorem 9.24** (Derriennic's 0-2 law [Der76])**.** *For SRW on a finitely generated group, for any $k \in \mathbb{N}$,*

$$\|\mu^n - \mu^{n+k}\|_1 = 2\, d_{\mathrm{TV}}(\mu^n, \mu^{n+k}) = \begin{cases} 2 & \textit{for all } n\,, \ \textit{or} \\ o(1) & \textit{as } n \to \infty\,. \end{cases}$$

▷ **Exercise 9.17.** *Prove Theorem 9.23 from Theorem 9.24 and the hints above. Consult [LyPer16, Sections 14.2 and 14.6] if needed.*

We now have the following connection between the invariant $\sigma$-field and bounded harmonic functions for arbitrary Markov chains on a state space $\mathcal{S}$. In some sense, it is a generalization of Theorem 9.10, which roughly said that if there is one possible limiting behavior of a Markov chain, then there are only trivial bounded harmonic functions.

{t.harmonicinv

**Theorem 9.25.** *There is an invertible correspondence between bounded harmonic functions on $\mathcal{S}$ and equivalence classes of bounded invariant functions on $\Omega$ .*

*Proof.* Let $f : \Omega \longrightarrow \mathbb{R}$ be an invariant function representing an equivalence class, then the corresponding harmonic function $u : \mathcal{S} \longrightarrow \mathbb{R}$ is

$$u(x) = \mathbf{E}_x f(X_0, X_1, \ldots)\,,$$

where $\mathbf{E}_x$ is the expectation operator with respect to $\mathbf{P}_x$. From the other direction, let $u : \mathcal{S} \longrightarrow \mathbb{R}$ be harmonic, then the corresponding equivalence class is the one represented by the invariant function $f : \Omega \longrightarrow \mathbb{R}$ defined by

$$f(y_0, y_1, \ldots) = \limsup_{n\to\infty} u(y_n)\,.$$

We now prove this correspondence is invertible. For one direction, let $u : \mathcal{S} \longrightarrow \mathbb{R}$; we need to show

$$u(x) = \mathbf{E}_x \limsup u(X_n)\,.$$

This follows easily from the fact that $u(X_n)$ is a bounded martingale, and so by the Martingale Convergence Theorem 9.8 and the Dominated Convergence Theorem, we have that $\mathbf{E}_x \limsup u(X_n) = \mathbf{E}_x(u(X_0)) = u(x)$.

For the other direction, let $f : \Omega \longrightarrow \mathbb{R}$ be invariant and represent an equivalence class, then we need to prove that for any $x \in \mathcal{S}$ the following equality holds $\mathbf{P}_x$-a.s.:

$$f(y_0, y_1, \ldots) = \limsup_{n\to\infty} \mathbf{E}_{y_n} f(X_0, X_1, \ldots)\,.$$

In words, for almost every random walk trajectory $\{y_n\}$ started from $y_0 = x$, the end-result is well-approximated by the average end-result of a new random walk started at a large $y_n$. Indeed, for any $x \in \mathcal{S}$ we have by Lévy's 0-1 law Theorem 9.19 that $\mathbf{P}_x$-a.s.

$$f(y_0, y_1, \dots) = \lim_{n\to\infty} \mathbf{E}_x \Big[ f(y_0, \dots, y_n, X_{n+1}, \dots) \,\Big|\, X_0 = y_0, \dots, X_n = y_n \Big] ,$$

but since $f$ is an invariant function,

$$\mathbf{E}_x \Big[ f(y_0, \dots, y_n, X_{n+1}, \dots) \,\Big|\, X_0 = y_0, \dots, X_n = y_n \Big] = \mathbf{E}_{y_n} f(X_0, X_1, \dots) ,$$

and we are done. □

By looking at indicators of invariant events, we get the following immediate important consequence, a much more general version of the (I) $\Longleftrightarrow$ (H) equivalence in Theorem 9.17.

{c.harmbound}

**Corollary 9.26.** *The invariant $\sigma$-field of $\Omega$ is trivial if and only if all bounded harmonic functions on $\mathcal{S}$ are constant.*

The invariant $\sigma$-field of a Markov chain (or equivalently, the space of bounded harmonic functions) is called the **Poisson boundary**. We can already see why it is a boundary: it is the space of possible different behaviours of the Markov chain trajectories at infinity. The name Poisson comes from the following analogue.

If $\mathbb{U} \subset \mathbb{C}$ is the open unit disk, and $f : \partial\mathbb{U} \longrightarrow \mathbb{R}$ is a bounded Lebesgue-measurable function, then we can define its harmonic extension $\overline{f} : \mathbb{U} \longrightarrow \mathbb{R}$ by the Poisson formula

$$\overline{f}(z) = \int_0^1 \frac{1 - |z|^2}{|e^{2\pi i\theta} - z|^2} f(\theta)\, d\theta ,$$

which is nothing else but integration against the harmonic measure from $x$, i.e., $\mathbf{E}_x f(\tau)$, where $\tau$ is the first hitting time of $\partial\mathbb{U}$ by Brownian motion in $\mathbb{U}$. That is, $\partial\mathbb{U}$ plays the role of the Poisson boundary for Brownian motion on $\mathbb{U}$.

This analogy between random walks on groups and complex analysis can be made even closer by equipping $\mathbb{U}$ with the hyperbolic metric, then thinking of this hyperbolic plane as $\mathsf{SL}_2(\mathbb{R})/\mathsf{SO}(2)$, via the Möbius transformations of the upper half plane model, as around (3.3). Now, harmonic measure on $\partial\mathbb{U}$ w.r.t. hyperbolic Brownian motion is the same as w.r.t. the Euclidean one, except that the hyperbolic BM never hits the ideal boundary $\partial\mathbb{U}$, only converges to it. It turns out that simple random walk on cocompact discrete subgroups of $\mathsf{PSL}_2(\mathbb{R})$ (equivalently, on hyperbolic tilings) approximates well this hyperbolic Brownian motion w.r.t. harmonic measure, hence the Poisson boundary of these groups can be realized geometrically as $\partial\mathbb{U}$. This is true in much greater generality: the Poisson boundary of Gromov-hyperbolic groups can be naturally identified with their usual topological Gromov-boundary [Kai00].

## 9.5 The Poisson boundary and entropy. The importance of group-invariance

{ss.PoiEnt}

We now prove the (E) $\Longleftrightarrow$ (I) part of Theorem 9.17, in a more general form: for arbitrary finite entropy measures instead of just symmetric finitely supported ones.

{t.PoiEnt}

**Theorem 9.27** ([KaiV83]). *For any countable group $\Gamma$ and a measure $\mu$ on it with finite entropy $H(\mu) < \infty$, the Poisson boundary of $\mu$ is trivial iff the asymptotic entropy vanishes, $h(\mu) = 0$.*

*Proof.* We will need the notion of **conditional entropy**: if $X$ and $Y$ are two variables on the same probability space, taking values in some countable sets, then the **conditional entropy** $H(X\,|\,Y)$ is the expected entropy of the conditioned variable $X$, given $Y$:

$$\begin{aligned}H(X\,|\,Y) &= \sum_y \mathbf{P}[\,Y=y\,]\cdot\Big(-\sum_x \mathbf{P}[\,X=x\mid Y=y\,]\log\mathbf{P}[\,X=x\mid Y=y\,]\Big)\\ &= -\sum_{x,y}\mathbf{P}[\,X=x,Y=y\,]\Big(\log\mathbf{P}[\,X=x,Y=y\,]-\log\mathbf{P}[\,Y=y\,]\Big)\\ &= H(X,Y)-H(Y).\end{aligned}$$

The concavity of $-x\log x$ and Jensen's inequality imply that, for any possible value $x$ of $X$,

$$\sum_y\Big(-\mathbf{P}[\,X=x\mid Y=y\,]\log\mathbf{P}[\,X=x\mid Y=y\,]\Big)\mathbf{P}[\,Y=y\,]\le-\mathbf{P}[\,X=x\,]\log\mathbf{P}[\,X=x\,]\,,$$

and thus $H(X\,|\,Y)\le H(X)$, with equality if and only if $\mathbf{P}[\,X=x\mid Y=y\,]=\mathbf{P}[\,X=x\,]$ for all possible $x,y$, that is, if $X$ and $Y$ are independent. It follows that

$$H(X,Y)\le H(X)+H(Y)\,, \tag{9.8}$$ {e.Hjoint}

with equality if and only if $X$ and $Y$ are independent, which shows that entropy is a very sensible measure of randomness.

With this definition, using the notation $x_i=y_{i-1}^{-1}y_i$ for $i\ge1$ on the trajectory space $\Omega(y_0)$, from

$$\mathbf{P}_{y_0}\big[\,X_i=y_i \text{ for } i=1,\dots,k \;\big|\; X_n=y_n\,\big]=\frac{\mu(x_1)\cdots\mu(x_k)\,\mu^{n-k}(y_k^{-1}y_n)}{\mu^n(y_n)} \tag{9.9}$$ {e.9}

we easily get

$$H(X_1,\dots,X_k\mid X_n)=kh_1+h_{n-k}-h_n\,, \tag{9.10}$$ {e.7}

where $h_i:=H(\mu^i)=H(X_i)$. Now notice that Markovianity implies that $H(X_1,\dots,X_k\mid X_n,X_{n+1})=H(X_1,\dots,X_k\mid X_n)$, and (9.8) implies that $H(X_1,\dots,X_k\mid X_n,X_{n+1})\le H(X_1,\dots,X_k\mid X_{n+1})$. Combined with the $k=1$ case of (9.10), we get that

$$h_n-h_{n-1}\ge h_{n+1}-h_n\,.$$

From this and the definition of $h(\mu)$, we get that $h_{n+1}-h_n$ decreases monotonically to $h(\mu)$.

Now, take the limit $n\to\infty$ in (9.10) to get

$$\begin{aligned}\lim_{n\to\infty}H(X_1,\dots,X_k\mid X_n)&=kh_1-kh(\mu)\\ &=H(X_1,\dots,X_k)-kh(\mu)\,,\end{aligned} \tag{9.11}$$ {e.12}

where $H(X_1,\dots,X_k)=kH(\mu)=kh_1$ is similar to (9.9). Note that, in particular, we get the following formula for the asymptotic entropy:

$$h(\mu)=H(X_1)-\lim_{n\to\infty}H(X_1\mid X_n)\,. \tag{9.12}$$ {e.13}

By (9.8), we can reinterpret (9.11) as follows: $h(\mu)=0$ iff $(X_1,\dots,X_k)$ is asymptotically independent of $X_n$ for all $k\ge1$, under any $\mathbf{P}_{y_0}$. We will show that this is equivalent to the triviality of the tail $\sigma$-field $\mathcal{T}$. Since Theorem 9.23 tells us that $\mathcal{T}=\mathcal{I}$ up to $\mathbf{P}_{y_0}$-measure zero sets, this will finish the proof.

Conditioned on $X_n$, any tail event $A \in \mathcal{T}$ is independent of $(X_1, \dots, X_k)$ by the Markov property. Thus, if $(X_1, \dots, X_k)$ is also asymptotically independent of $X_n$, then $A$ is independent of $(X_1, \dots, X_k)$. That means that $A$ is measurable w.r.t. the jumps made after $X_k$, for any $k$, hence it is in the tail $\sigma$-field of the jumps. So, $\mathcal{T}$ is trivial by Kolmogorov's 0-1 law (Theorem 9.20). Conversely, if $\mathcal{T}$ is trivial, then Lévy's 0-1 law (Theorem 9.19) tells us that the distribution of $(X_1 \mid X_n) = (X_1 \mid X_n, X_{n+1}, \dots)$ converges to that of $X_1$. Fatou's lemma implies that $\liminf_{n\to\infty} H(X_1 \mid X_n) \ge H(X_1)$, and hence in (9.12) we must have $h(\mu) = 0$. □

In the above proof, and hence in the equivalence between positive speed and the non-Liouville property, it is important for the walk to be group-invariant, as shown by the following example, a non-amenable bounded degree graph that has positive speed and entropy, but has the Liouville property and trivial Poisson boundary. The main place where the above proof breaks down is in (9.9), where we wrote $\mathbf{P}[\, X_n = y_n \mid X_k = y_k \,] = \mu^{n-k}(y_k^{-1} y_n)$, and hence got $h_{n-k}$ instead of $H(X_n \mid X_k)$ in (9.10).

**Example: a non-amenable Liouville graph [BenK10].** Take an infinite binary tree, and on the vertex set of each level $L_n$, place a 3-regular expander. Clearly, simple random walk on this graph $G = (V, E)$ has positive speed (namely, $2/6 - 1/6 = 1/6$) and positive entropy (after $n$ steps, the walk is close to level $n/6$ with huge probability, and given that this level is $k$, the distribution is uniform on this set of size $2^k$). However, this graph is Liouville. The main idea is that the expanders mix the random walk fast enough within the levels, so that there can never be a specific direction of escape. Here is the outline of the actual argument.

Let $h : V \longrightarrow \mathbb{R}$ be a bounded harmonic function. Let $u, v \in V$, and let $\nu_n^u$ denote the harmonic measure on level $L_n$ of simple random walk (i.e., the distribution of the first vertex hit), for $n$ larger than the level of $u$. Then, by the Optional Stopping Theorem (see Section 6.3), for all large $n$,

$$h(u) - h(v) = \sum_{w \in L_n} h(w)(\nu_n^u(w) - \nu_n^v(w)) \le \max_{x \in G} h(x)\, 2\, d_{\mathrm{TV}}(\nu_n^u, \nu_n^v)\,.$$

So, we need to show that the total variation distance between the harmonic measures converges to zero. The idea for this is that when first getting from $L_n$ to $L_{n+1}$, the walk with positive probability makes at least one step inside the expander on $L_n$, delivering some strict contractive effect in the $L^2$-norm. (All $L^p$ norms are understood w.r.t. the uniform probability measure on $L_n$.) It is also not hard to show that moving in the other directions are at least non-expanding. Formally, define the operator $T_n : L^2(L_n) \longrightarrow L^2(L_{n+1})$ by

$$T_n \phi(y) := 2 \sum_{x \in L_n} \phi(x) \nu_{n+1}^x(y)\,,$$

where the factor $2 = |L_{n+1}|/|L_n|$ ensures that $T_n \mathbf{1} = \mathbf{1}$, and, in fact, one can show that $\|T_n\|_{1\to 1} \le 1$ and $\|T_n\|_{\infty\to\infty} \le 1$, so, by the Riesz-Thorin interpolation theorem, also $\|T_n\|_{2\to 2} \le 1$. Then, by looking at the possible first steps of the walk before hitting $L_{n+1}$, we can write $T_n$ as a weighted sum of composition of other operators, all with $L^2$-norm at most 1, while the one encoding the step within $L_n$ is a strict $L^2$-contraction on functions orthogonal to constants. Altogether, we get that for any function $\phi \in L^2(L_n)$ with $\sum_{x \in L_n} \phi(x) = 0$, we have $\|T_n \phi\|_2 \le (1 - c)\|\phi\|_2$ with some $c > 0$. This implies that $f_n^u := 2^n \nu_n^u - 1 = T_n T_{n-1} \cdots T_{k+1}(2^k \delta_u - \mathbf{1})$, where $u \in L_k$, satisfies $\|f_n^u\|_2 \to 0$ as $n \to \infty$. This also easily implies that

$$2\, d_{\mathrm{TV}}(\nu_n^u, \nu_n^v) \le \|f_n^u\|_1 + \|f_n^v\|_1 \to 0\,,$$

and we are done. □

A key motivation for [BenK10] was the following conjecture:

{ex.infexp}

▷ **Exercise 9.18** (Itai Benjamini). *** *Show that there exists* **no infinite expander***: this would be a bounded degree infinite graph such that every ball $B_n(x)$ around every vertex $x$ is itself an expander, with a uniform Cheeger constant $c > 0$.*

*Show at least that this property is invariant under quasi-isometries, or at least independent of the choice of a finite generating set of a group.*

Note that the graph in the above example has expander balls around the root of the binary tree. One may hope that the above proof could be generalized to show that simple random walk on an infinite expander would always have trivial Poisson boundary, and thus this could not be a Cayley graph, at least. However, such a generalization seems problematic. It is important in this proof that the harmonic measure from the root on the levels $L_n$ is uniform; in fact, if we start a random walk uniformly at $x \in L_n$, then the averaged harmonic measure on $L_{n+1}$ is uniform again. Formally, this is used in proving $\|T_n\|_{\infty\to\infty} \le 1$ w.r.t. the uniform measures on the levels. Without this uniformity, say, if we place expanders on the levels on an irregular infinite tree, one could work with non-uniform measures on the levels in the definition of the operators $T_n$ or in the operator norms, and still could have $\|T_n\|_2 \le 1$ for all $n$. However, the expanders on the levels are expanders w.r.t. uniform measure, and if the harmonic measure on a level is concentrated on a very small subset, then the walk might not feel the contractive effect in the horizontal direction. See, e.g., the following exercise.

▷ **Exercise 9.19** ([BenK10]). * *Consider an imbalanced binary tree: from each vertex, put a double edge to the right child and a single edge to the left child. Then place a regular expander on each level in a way that the resulting graph has a non-trivial Poisson boundary.*

This issue calls for the following question:

{ex.unifharm}

▷ **Exercise 9.20.** *** *Does every finitely generated group have a generating set in which the harmonic measures $\nu_n^o$ on the spheres $L_n := \partial_V^{\mathrm{in}} B_n(o)$ are roughly uniform in the sense that there exist $0 < c, C < \infty$ such that for each $n$ there is $U_n \subset L_n$ with $\nu_n^o(U_n) > c$ and $c < \nu_n^o(x)/\nu_n^o(y) < C$ for all $x, y \in U_n$? This is very similar to Vershik's question, Exercise 9.15.*

An affirmative answer to this question, together with an affirmative answer to the question of invariance of the infinite expander property under a change of generators would give a proof of Benjamini's conjecture, Exercise 9.18, for Cayley graphs. On the other hand, one could try to find counterexamples to Exercises 9.20 and 9.15 among non-amenable torsion groups, say (see Section 15.2).

## 9.6 Unbounded measures

{ss.unbound}

There are a lot of nice results showing that there are good reasons for leaving sometimes our usual world of finitely supported measures.

The following two theorems from [KaiV83] make a direct connection between non-trivial Poisson boundary and non-amenability. Recall that a measure $\mu$ on a group $\Gamma$ with unit element $e$ is called aperiodic if the largest common divisor of the set $\{n : \mu^n(e) > 0\}$ is 1.

{t.PoissonMean

**Theorem 9.28.** *For any aperiodic measure $\mu$ whose (not necessarily finite) support generates the group $\Gamma$, the Poisson boundary of $\mu$ is trivial if and only if the convolutions $\mu^n$ converge weakly to a left-invariant mean on $L^\infty(\Gamma)$, see Definition 5.4; i.e., if for any bounded function $f : \Gamma \longrightarrow \mathbb{R}$ and any $g \in \Gamma$, we have $\sum_{x\in\Gamma} \big(\mu^n(x) - \mu^n(g^{-1}x)\big) f(x) \to 0$.*

*Proof.* Note that weak convergence to a left-invariant mean is a bit weaker than the condition $d_{\mathrm{TV}}\big(\mu^n(\cdot),\mu^n(g^{-1}\cdot)\big) \to 0$ that we used in the coupling proofs of Section 9.2. Nevertheless, it still implies a trivial Poisson boundary: if $f$ is any bounded harmonic function, then $f(g) = \sum_{x\in\Gamma}\mu_n(g^{-1}x)f(x)$ for any $n \in \mathbb{N}$, and subtracting from this the $g = e$ case and letting $n \to \infty$, we get that $f(g) = f(e)$, for any $g \in \Gamma$.

On the other hand, in the proof of Theorem 9.27 we saw that a trivial Poisson boundary implies that $\mathbf{P}_e[\,X_1 = g \mid X_n\,]$ converges to $\mu(g)$ for almost every trajectory $X_0, X_1, \dots$. Thus, for any $g \in \operatorname{supp}\mu$, we have

$$\lim_{n\to\infty} \frac{\mu^{n-1}(g^{-1}X_n)}{\mu^n(X_n)} = 1 \quad \text{almost surely.}$$

If $\mu$ is aperiodic, then by passing to a large enough convolution power we may assume that $\mu(e) > 0$. Taking $g = e$ in the displayed formula, we get that

$$\lim_{n\to\infty} \frac{\mu^{n-1}(X_n)}{\mu^n(X_n)} = 1 \quad \text{almost surely.}$$

Comparing the two displayed formulas, for $g \in \operatorname{supp}\mu$,

$$1 = \lim_{n\to\infty} \frac{\mu^{n-1}(g^{-1}X_n)}{\mu^n(X_n)} = \lim_{n\to\infty} \frac{\mu^{n-1}(g^{-1}X_n)}{\mu^{n-1}(X_n)} = \lim_{n\to\infty} \frac{\mu^{n}(g^{-1}X_n)}{\mu^n(X_n)} \quad \text{almost surely,}$$

where we got the last equality by changing the index $n$ to $n+1$ and conditioning on the last step to be $X_n = X_{n+1}$, an event of positive probability. This implies that, for any $\epsilon > 0$,

$$\mu^n\left\{x \in \Gamma : \left|1 - \frac{\mu^n(g^{-1}x)}{\mu^n(x)}\right| > \epsilon\right\} \to 0$$

for any $g \in \operatorname{supp}\mu$. Since $\operatorname{supp}\mu$ generates $\Gamma$, we obtain the desired convergence to an invariant mean. □

The previous theorem immediately implies that on a non-amenable group any non-degenerate $\mu$ has non-trivial Poisson boundary. (For symmetric finitely supported measures we knew this from Proposition 9.3.) Conversely, Theorem 9.28 can also be used to produce a symmetric measure with trivial Poisson boundary in any amenable group. The idea is to take an average of the uniform measures on larger and larger Følner sets. This establishes the following result, conjectured by Furstenberg.

{t.FurstConj}

**Theorem 9.29.** *A group $\Gamma$ is amenable iff there is a measure $\mu$ supported on the entire $\Gamma$ whose Poisson boundary is trivial.*

In this theorem, the support of the measure $\mu$ that shows amenability might indeed need to be large: it is shown in [Ers04a, Theorem 3.1] that, on the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$ with $d \geq 3$, any measure with finite entropy has non-trivial Poisson boundary.

For non-amenable groups, the theorem says that positive speed cannot be ruined by strange large generating sets. Can the spectral radius being less than 1 be ruined? It turns out that there exist non-amenable groups where the spectral radius of finite symmetric walks can be arbitrary close to 1 [Osi02, ArBLRSV05], hence I expect that an infinite support can produce a spectral radius 1.

▷ **Exercise 9.21.** ** *Can there exist a symmetric measure $\mu$ whose infinite support generates a finitely generated non-amenable group $\Gamma$ such that the spectral radius is $\rho(\mu) = 1$?*

We have already mentioned the Choquet-Deny theorem [ChoD60], saying that any irreducible measure on any Abelian group has the Liouville property. How small does a group have to be for this to remain true? It was proved by Dynkin-Malyutov (1961) and Margulis (1966) that all virtually nilpotent groups have this property, and this was recently proved to be an equivalence [FrHTV18]. The following very simple proof for the direction that random walks on Abelian groups are always Liouville was found by Raugi, also extended to nilpotent groups by him.

{t.Raugi}

**Theorem 9.30** ([Rau04]). *For any probability measure $\mu$ on a countable nilpotent group $\Gamma$ that generates the entire group, the associated Poisson boundary is trivial.*

*Proof for the Abelian case.* Let $h : \Gamma \longrightarrow \mathbb{R}$ be a bounded harmonic function, $T_1, T_2, \dots$ independent steps w.r.t. $\mu$, and $X_n = X_0 + T_1 + \cdots + T_n$ the random walk. For all $n \geq 1$, define

$$\begin{aligned} u_n(x) :&= \mathbf{E}_x\big[ \big(h(X_n) - h(X_{n-1})\big)^2 \big] \\ &= \sum_{t_1,\dots,t_n \in \Gamma} \big(h(x+t_1+\cdots+t_n) - h(x+t_1+\cdots+t_{n-1})\big)^2 \mu(t_1)\cdots\mu(t_n)\,. \end{aligned}$$

Now, for $n \geq 2$,

$$\begin{aligned} u_n(x) &= \mathbf{E}\Big[ \mathbf{E}_x\big[ \big(h(X_n) - h(X_{n-1})\big)^2 \bigm| T_2,\dots,T_n \big] \Big] \\ &\geq \mathbf{E}\Big[ \Big( \mathbf{E}_x\big[ h(X_n) - h(X_{n-1}) \bigm| T_2,\dots,T_n \big] \Big)^2 \Big] \text{ by Jensen or Cauchy-Schwarz} \\ &= \mathbf{E}\Big[ \big( h(x+T_2+\cdots+T_n) - h(x+T_2+\cdots+T_{n-1}) \big)^2 \Big] \text{ by harmonicity and commutativity} \\ &= \mathbf{E}_x\big[ \big(h(X_{n-1}) - h(X_{n-2})\big)^2 \big] = u_{n-1}(x)\,. \end{aligned}$$

On the other hand, $u_n(x) = \mathbf{E}_x\big[h(X_n)^2\big] - \mathbf{E}_x\big[h(X_{n-1})^2\big]$ by the orthogonality of martingale increments (Pythagoras for martingales), hence $\sum_{n=1}^N u_n(x) = \mathbf{E}_x\big[h(X_N)^2\big] - h(X_0)^2$. This is a sum of non-decreasing non-negative terms that remains bounded as $N \to \infty$, hence all the terms must be zero, which means that $h$ is constant along any possible trajectory and hence on the entire group. □

We know from the entropy criterion for the Poisson boundary that a finitely supported measure on any group with subexponential growth has trivial boundary. On the other hand, Anna Erschler proved in [Ers04b] that there exist a group with subexponential growth (an example of Grigorchuk, see Section 15.1) and some infinitely supported measure on it with finite entropy that has a non-trivial Poisson boundary.

The following question from [KaiV83] is still open: *Is it true that on any group of exponential growth there is a $\mu$ with non-trivial Poisson boundary?* For solvable groups, this was shown (with symmetric $\mu$ with finite entropy) in [Ers04a, Theorem 4.1]. On the other hand, Bartholdi and Erschler have recently shown that there exists a group of exponential growth on which any finitely supported $\mu$ has a trivial boundary [BartE11].

# 10 Growth of groups, of harmonic functions and of random walks

## 10.1 A proof of Gromov's theorem

{ss.Gromov}

We now switch to the theorem characterizing groups of polynomial growth due to Gromov [Gro81] and give a brief sketch of the new proof due to Kleiner [Kle10], also using ingredients and insights from [LeeP13] and [Tao10]. In fact, [Tao10] contains a self-contained and mostly elementary proof of Gromov's theorem. We will borrow some parts of the presentation there.

{t.Gromov}

**Theorem 10.1** (Gromov's theorem [Gro81]). *A finitely generated group has polynomial growth if and only if it is virtually nilpotent.*

We have already proved the "if" direction, in Theorem 4.12. For the harder "only if" direction, we will need several facts which we provide here without proper proofs but at least with a sketch of their main ideas. The first ingredient is the following:

{t.harmfindim}

**Theorem 10.2** ([ColM97], [Kle10])**.** *For fixed positive integers $d$ and $\ell$ there exists some constant $f(d,\ell)$, such that for any Cayley graph $G$ of any finitely generated group $\Gamma$ of polynomial growth with degree $\leq d$, the space of harmonic functions on $G$ with growth degree $\leq \ell$ has dimension $\leq f(d,\ell)$ (so in particular is finite-dimensional).*

For a tiny bit of intuition to start with, recall Exercise 9.4, saying that sublinear harmonic functions on $\mathbb{Z}^d$ are constant. But there, in the coupling proof, we used the product structure of $\mathbb{Z}^d$ heavily, while here we do not have any algebraic information — this is exactly the point. The Colding-Minicozzi proof used Gromov's theorem, so it was not good enough for Kleiner's purposes, but it did motivate his proof.

*Sketch of proof of Theorem 10.2.* We give Kleiner's proof for the case when the Cayley graph satisfies the so-called doubling condition: there is an absolute constant $D < \infty$ such that $|B(2R)| \leq D\,|B(R)|$ for any radius $R$. This does not follow easily from being a group of polynomial growth, so this is an annoying but important technicality all along Kleiner's proof. Now, the key lemma is the following:

{l.elliptic}

**Lemma 10.3.** *Cover the ball $B_R$ by balls $B^i$ of radius $\epsilon R$, and suppose that a harmonic function $f : G \longrightarrow \mathbb{R}$ has mean zero on each $B^i$. Then $\|f\|_{\ell^2(B_R)} \leq C\,\epsilon\,\|f\|_{\ell^2(B_{4R})}$.*

*Proof.* Shrink each $B^i$ by a factor of two, and take a maximal subset of them in which all of them are disjoint. Then the corresponding original balls $B^i$ still cover $B_R$. Let the set of these $B^i$s be $\mathcal{B}$. Enlarge now each $B^i \in \mathcal{B}$ by a factor of three. We claim that each point in $B_{2R}$ is covered by at most some $D'$ of these $3B^i$'s. This is because of the doubling property: if there were more than $D'$ of them covering a point $x$, then $B_{3.5\epsilon R}(x)$ would contain $D'$ disjoint balls $B^i/2$, so we would have $D'|B_{\epsilon R/2}| \leq |B_{3.5\epsilon R}|$, which cannot be for large enough $D'$.

Now apply Saloff-Coste's Poincaré inequality Exercise 8.2 to each $B^i$, sum over the $B^i$'s, use the existence of $D'$, then apply the reverse Poincaré inequality Exercise 8.3 to $B_{2R}$ to get

$$\begin{aligned}\sum_{x\in B_R} |f(x)|^2 \leq \sum_{B^i\in\mathcal{B}}\sum_{x\in B^i} |f(x)|^2 &\leq O(1)\,\epsilon^2 R^2 \sum_{B^i\in\mathcal{B}}\sum_{x\in 3B^i} |\nabla f(x)|^2 \\ &\leq O(1)\,\epsilon^2 R^2 \sum_{x\in B_{2R}} |\nabla f(x)|^2 \leq O(1)\,\epsilon^2 \sum_{x\in B_{4R}} |f(x)|^2\,,\end{aligned}$$

which proves the lemma. □

This lemma implies that by imposing relatively few constraints on a harmonic function we can make it grow quite rapidly. To finish the proof of Theorem 10.2, the idea is that if we take a vector space spanned by many independent harmonic functions, then this lemma implies that some of the functions in this space would grow quickly. So, the space of harmonic functions with moderate growth cannot be large. To realize this idea, the trick is to consider the Gram determinant

$$\det\left((u_i,u_j)_{\ell^2(B_R)}\right)_{i,j=1}^N,$$

where $\{u_i : i = 1,\dots,N\}$ is a basis for the harmonic functions of growth degree $\leq \ell$ such that $\{u_i : 1 \leq i \leq M\}$ span the subspace where the mean on each $B^i \in \mathcal{B}$ is zero. This determinant is the squared volume of the parallelepiped spanned by the $u_i$'s. Now, on one hand, obviously $\|u_i\|_{\ell^2(B_R)} \leq \|u_i\|_{\ell^2(B_{4R})}$

for all $i$, while, for $1 \le i \le M = N - |\mathcal{B}|$, we also have $\|u_i\|_{\ell^2(B_R)} \le C\,\epsilon\,\|u_i\|_{\ell^2(B_{4R})}$ by Lemma 10.3. By the construction of $\mathcal{B}$ and the doubling property, $|\mathcal{B}| = O(\epsilon)$. Altogether, the squared volume of the parallelepiped is

$$\begin{aligned}\det\big((u_i,u_j)_{\ell^2(B_R)}\big)_{i,j=1}^N &\le O(\epsilon^2)^{N-O(\epsilon)}\,\det\big((u_i,u_j)_{\ell^2(B_{4R})}\big)_{i,j=1}^N\\ &\le O(\epsilon^2)^{N-O(\epsilon)}\,O(1)\,4^{d+2\ell}\det\big((u_i,u_j)_{\ell^2(B_R)}\big)_{i,j=1}^N \text{ by the growth conditions}\\ &\le \frac{1}{2}\,\det\big((u_i,u_j)_{\ell^2(B_R)}\big)_{i,j=1}^N \text{ if } \epsilon \text{ is small and } N \text{ is large}\,.\end{aligned}$$

This means that the determinant is zero for all $R$ whenever $N$ is larger than some $f(d,\ell)$. This proves Theorem 10.2. $\square$

The second ingredient for the proof of Gromov's theorem is complementary to the previous one: it will imply that for groups of polynomial growth there do exist non-trivial harmonic functions of moderate growth. Combining the two ingredients, we will be able to construct a non-trivial representation of our group over a finite-dimensional vector space, and then use the better understanding of linear groups.

{t.Mok}

**Theorem 10.4** ([Mok95], [KorS97], [Kle10], [LeeP13], [ShaT10])**.** *Let $\Gamma$ be a finitely generated group without Kazhdan's property (T); in particular, anything infinite amenable is fine. Then there exists an isometric (linear) action of $\Gamma$ on some real Hilbert space $\mathcal{H}$ without fixed points and a non-constant $\Gamma$-equivariant harmonic function $\Psi : \Gamma \longrightarrow \mathcal{H}$. Equivariance means that $\Psi(gh) = g(\Psi(h))$ for all $g, h \in \Gamma$, and harmonicity is understood w.r.t. some symmetric finite generating set.*

The first three of the five proofs listed above use ultrafilters to obtain a limit of almost harmonic functions into different Hilbert spaces. The last two proofs are constructive and quite similar to each other (done independently), inspired by random walks on amenable groups. We will briefly discuss now the proof of Lee and Peres [LeeP13].

Let us focus on the amenable case. (The trick for the general non-Kazhdan case is to consider the right Hilbert space action instead of the regular representation on $\ell^2(\Gamma)$ that is so closely related to random walks.) It works for all amenable transitive graphs, not only for groups. The key lemma is the following:

{l.yafoo}

**Lemma 10.5.** *For simple random walk on any transitive amenable graph $G(V,E)$, we have*

$$\inf_{\varphi\in\ell^2(V)} \frac{\|(I-P)\varphi\|^2}{\big(\varphi,(I-P)\varphi\big)} = 0\,.$$

As we have seen in Theorem 7.3 and Lemma 7.2, both $\|(I-P)\varphi\|^2/\|\varphi\|^2$ and $\big(\varphi,(I-P)\varphi\big)/\|\varphi\|^2$ are small if we take $\varphi$ to be the indicator function $\mathbf{1}_A$ of a large Følner set $A \subset V$. But the lemma concerns the ratio of these two quantities, which would be of constant order for typical Følner indicator functions $\mathbf{1}_A$. Instead, the right functions to take will be the smoothened out versions $\sum_{i=0}^{k-1} \mathbf{P}^i \mathbf{1}_A(x) = \mathbf{E}_x\big|\{0 \le i \le k-1 : X_i \in A\}\big|$, which are some truncated Green's functions.

{ex.smoothing}

▷ **Exercise 10.1.**

**(a)** *Show that if $(V,P)$ is transient or null-recurrent, then $P^i f \to 0$ pointwise for any $f \in \ell^2(V)$.*

**(b)** *Let $\varphi_k = \varphi_k(f) := \sum_{i=0}^{k-1} P^i f$. Show that $\frac{1}{k}(\varphi_k, f) \to 0$.*

**(c)** *Show that $\|(I-P)\varphi_k\|^2 \le 4\|f\|^2$ and $(\varphi_k,(I-P)\varphi_k) = (2\varphi_k - \varphi_{2k}, f)$.*

{ex.boosting}

▷ **Exercise 10.2.*** *Show that if $\|(I-P)f\|/\|f\| < \theta$ is small (e.g., for the indicator function of a large Følner set), then there is a $k \in \mathbb{N}$ such that $\big(2\varphi_k(f) - \varphi_{2k}(f), f\big) \ge L(\theta)$ is large. (Hint: first show that there is an $\ell$ such that $(\varphi_\ell(f), f)$ is large, then use part (b) of the previous exercise.)*

The combination of Exercise 10.1 (c) and Exercise 10.2 proves Lemma 10.5. Let us note that the proof would have looked much simpler if we had worked with $\varphi = \varphi_\infty(f) = \sum_{i=0}^\infty P^i f$, with $f = \mathbf{1}_A$ for a large Følner set $A$. Namely, we would have $(I-P)\varphi = \mathbf{1}_A$, hence $\|(I-P)\varphi\|^2 = |A|$, while $(\varphi,(I-P)\varphi) = \sum_{x\in A}\varphi(x) \geq r(|A| - d^r|\partial_V^{\mathrm{in}} A|)$ for any given $r$, where $G$ is $d$-regular, since we stay in $A$ for at least $r$ steps of the walk if we start at least distance $r$ away from the boundary. The above identity and the inequality together give the required small ratio in Lemma 10.5. However, we would need to prove here that this $\varphi$ exists (which is just transience) and is in $\ell^2(V)$, which is in fact true whenever the volume growth of $G$ is at least degree 5, but it is not clear how to show that without relying on Gromov's polynomial growth theorem, which we obviously want to avoid here.

We now want to prove Theorem 10.4 for the amenable case. Choose a sequence $\psi_j \in \ell^2(\Gamma)$ giving the infimum 0 in Lemma 10.5, and define $\Psi_j : G \longrightarrow \ell^2(\Gamma)$ by

$$\Psi_j(x) : \, g \mapsto \frac{\psi_j(g^{-1}x)}{\sqrt{2\big(\psi_j,(I-P)\psi_j\big)}}\,.$$

These are clearly equivariant functions on $\Gamma$. It is easy to see that

$$\sum_{y\in\Gamma} p(x,y)\|\Psi_j(x)-\Psi_j(y)\|^2 = 1 \tag{10.1}$$ {e.sumone}

for every $x\in\Gamma$, and

$$\Big\|\Psi_j(x) - \sum_{y\in\Gamma} p(x,y)\Psi_j(y)\Big\|^2 = \frac{\|(I-P)\psi_j\|^2}{2\big(\psi_j,(I-P)\psi_j\big)} \to 0\,, \tag{10.2}$$ {e.almostharm}

as $j\to\infty$. Using (10.1), the $\Psi_j$'s are uniformly Lipschitz, and by (10.2), they are almost harmonic. Then one can use a compactness argument to extract a limit, a harmonic equivariant function that is non-constant because of (10.1). This finishes the construction.

As the final ingredient for Gromov's theorem, here is what we mean by linear groups being better understood:

{t.Tits}

**Theorem 10.6** (Tits' alternative [Tit72]). *Any finitely generated linear group (i.e., a subgroup of some* $\mathsf{GL}(n,\mathbb{F})$*) is either almost solvable or contains a subgroup isomorphic to* $F_2$*.*

Actually, the special case of a finitely generated subgroup of a *compact* linear *Lie* group $H \subset \mathsf{GL}(n,\mathbb{C})$ will suffice for us, which is already a statement that can be proved in a quite elementary way, as I learnt from [Tao10]. First of all, $H$ is isomorphic to a subgroup of $\mathsf{U}(n)$, since any inner product on $\mathbb{C}^n$ can be averaged by the Haar measure of $H$ to get an $H$-invariant inner product on $\mathbb{C}^n$, thereby $H$ becoming a subgroup of the unitary group associated to this inner product. Then, the key observation is that if $g,h\in\mathsf{U}(n)$ are close to the identity (in the operator norm), then their commutator is even closer:

$$\begin{aligned}\|[g,h]-1\|_{op} &= \|gh-hg\|_{op}\\ &= \|(g-1)(h-1)-(h-1)(g-1)\|_{op}\\ &\leq 2\,\|g-1\|_{op}\,\|h-1\|_{op}\,,\end{aligned} \tag{10.3}$$ {e.commclose}

where the first line used that $g,h$ are unitary and the last line is by the triangle inequality. Using this, one can already easily believe *Jordan's theorem*: any finite subgroup $\Gamma$ of $\mathsf{U}(n)$ contains an Abelian subgroup $\Gamma^*$ of index at most $C_n$. Why? The elements $\Gamma_\epsilon$ of $\Gamma$ lying in a small enough $\epsilon$-neighbourhood of the identity will generate a good candidate for $\Gamma^*$: the finite index comes from the compactness of $\mathsf{U}(n)$ and the positivity of $\epsilon$, while a good amount of commutativity comes from taking the element $g$ of $\Gamma_\epsilon$ closest to the identity, and then using (10.3) to show that $[g,h] = 1$ for any $h \in \Gamma_\epsilon$. Then one can use some sort of induction. Similarly to this and to the proof of Proposition 4.9, one can prove that any finitely generated subgroup of $\mathsf{U}(n)$ of subexponential growth is almost Abelian.

*Proof of Gromov's Theorem 10.1.* We now use these facts to give a proof of the "only if" direction, by induction on the growth degree of $\Gamma$. The base case is clear: groups of growth degree 0 are precisely the finite ones, and they are almost nilpotent, since the trivial group is nilpotent. Suppose we already proved the result for groups of degree $\leq d-1$, and let $\Gamma$ be finitely generated of polynomial growth with degree $\leq d$. $\Gamma$ is amenable, hence from Theorem 10.4 we have a non-trivial equivariant harmonic embedding $\psi$ into a real Hilbert space $\mathcal{H}$. It is Lipschitz, because $\psi(gs) - \psi(g) = g(\psi(s) - \psi(e))$. Let $V$ be the vector space of harmonic real-valued Lipschitz functions on $\Gamma$ — it is finite dimensional by Theorem 10.2. Since $\psi : G \longrightarrow \mathcal{H}$ is non-constant, there is a bounded linear functional $\pi : \mathcal{H} \longrightarrow \mathbb{R}$ such that $\psi_0 := \pi \circ \psi \in V$ is a non-constant harmonic function on $\Gamma$. This $\psi_0$ cannot attain its maximum by the maximum principle, so it takes infinitely many values.

Now, $\Gamma$ acts on $V$ via $g : u \mapsto u_g$, where $u_g(x) = u(g^{-1}x)$. Moreover, it preserves the Lipschitz norm, and also acts on the vector space $W = V/\mathbb{C}$, where $u(x) \sim u(x) + c$ for any $c \in \mathbb{C}$. On $W$, the Lipschitz norm is a genuine norm, and on a finite dimensional vector space any two norms are equivalent (up to constant factor bounds), hence the action of $\Gamma$ preserves a Euclidean structure up to constant factors. Thus, we get a representation $\rho : \Gamma \longrightarrow \mathsf{GL}(W)$ with a precompact image. This image is infinite, because by applying elements of $\Gamma$ we can get infinitely many different functions from $\psi_0$. So, the group $B = \operatorname{Im} \rho$ is finitely generated, infinite, and of polynomial growth, inside a compact Lie group, so by the compact Lie case of the Tits alternative Theorem 10.6, it is almost solvable (and, in fact, almost Abelian, but let us use only solvability). Let $A_0 = A$ be the finite index solvable subgroup of $B$, and $A_k = [A_{k-1}, A_{k-1}]$. Since $A$ is infinite and solvable, there is a smallest index $\ell$ such that $A_\ell$ has infinite index in $A_{\ell-1}$. Then $A_{\ell-1}$ has finite index in $B$, so $\Gamma_1 := \rho^{-1}(A_{\ell-1})$ has finite index in $\Gamma$, so is also finitely generated of polynomial growth with degree $\leq d$, by Corollary 3.8. It is enough to show that $\Gamma_1$ is almost nilpotent. The group $A_{\ell-1}/A_\ell$ is Abelian, infinite and finitely generated, hence it can be projected onto $\mathbb{Z}$. So, we get a projection $\psi : A_{\ell-1} \longrightarrow \mathbb{Z}$, and then an exact sequence

$$1 \longrightarrow N \longrightarrow \Gamma_1 \xrightarrow{\psi \circ \rho} \mathbb{Z} \longrightarrow 1\,.$$

We now apply the results of Section 4.3: by Proposition 4.9, $N$ is finitely generated and of polynomial growth with degree $\leq d-1$, so, it is almost nilpotent by the inductive assumption. Then, by Proposition 4.8, $\Gamma_1$ is almost nilpotent. $\square$

This proof was made as quantitative as possible [ShaT10]; one of the many kinds of things that Terry Tao likes to do. In particular, they showed that there is some small $c > 0$ such that a $|B(R)| \leq R^{c(\log\log R)^c}$ volume growth implies almost nilpotency and hence polynomial growth. This is the best result to date towards verifying the conjectured gap between polynomial and $\exp(c\sqrt{n})$ volume growth.

We will see in Chapter 16 that there are transitive graphs that are very different from Cayley graphs: not quasi-isometric to any of them. But this does not happen in the polynomial growth regime: there is an extension of Gromov's theorem by [Tro85] and [Los87], see also [Woe00, Theorem 5.11], that any quasi-transitive graph of polynomial growth is quasi-isometric to some Cayley graph (of an almost nilpotent group, of course). In particular, it has a well-defined integer growth degree.

## 10.2 Random walks on groups are at least diffusive

{ss.diffusive}

We have seen that non-amenability implies positive speed of escape. However, we may ask for other rates of escape.

▷ **Exercise 10.3.** *Using the results seen for return probabilities in Chapter 8, show that:*
**(a)** *Any group with polynomial growth satisfies* $\mathbf{E}\big[\, d(X_0, X_n) \,\big] \geq c\sqrt{n}$. *(Notice that this is sharp for* $\mathbb{Z}^d$, $d \geq 1$, *using the Central Limit Theorem.)*
**(b)** *If the group has exponential growth, then* $\mathbf{E}\big[\, d(X_0, X_n) \,\big] \geq c\, n^{1/3}$.

Although it may seem that a random walk on a group with exponential growth should be further away from the origin than on a group with polynomial growth, the walk with exponential growth also has more places to go which are not actually further away, or in other words, it is more spread-out than random walk on a group with polynomial growth. Moreover, there can be many dead ends in the Cayley graph: vertices from which all steps take us closer to the origin. Therefore, it is not at all obvious that the rate of escape is at least $c\sqrt{n}$ on any group, and this was in fact unknown for quite a while.

{ex.AnnaMok}

▷ **Exercise 10.4** (Anna Erschler)**.** *Using the equivariant harmonic function* $\Psi$ *into a Hilbert space* $\mathcal{H}$ *claimed to exist for any amenable group in Theorem 10.4, show that*

$$\mathbf{E}\big[\, d(X_0, X_n)^2 \,\big] \geq c \cdot n\,,$$

*which is slightly weaker than the rate* $c\sqrt{n}$ *for the expectation in the previous exercise.* ***Hint:*** *If* $M_n$ *is martingale in* $\mathcal{H}$, *then* $\mathbf{E}\big[\, \|M_n - M_0\|^2 \,\big] = \sum_{k=0}^{n-1} \mathbf{E}\big[\, \|M_{k+1} - M_k\|^2 \,\big]$, *a Pythagorean Theorem for martingales (the orthogonality of martingale incremements) in Hilbert spaces.*

If we want to improve the linear bound on the second moment to a square root bound on the first moment, we need to show that $d(X_0, X_n)$ is somewhat concentrated. One way to do that is to show that higher moments do not grow rapidly.

▷ **Exercise 10.5.** * *Show that* $\mathbf{E}\big[\, \|\Psi(X_n) - \Psi(X_0)\|^4 \,\big] \leq Cn^2$, *using the orthogonality of martingale incremements. Then deduce that* $\mathbf{E}\big[\, d(X_0, X_n) \,\big] \geq c\sqrt{n}$. *(This improvement is due to Bálint Virág.* ***Hint:*** *do not be afraid to consider the time-reversal of the random walk when you need to condition on the future.)*

Even stronger concentration is true:

{t.LeePeres}

**Theorem 10.7** (James Lee and Yuval Peres [LeeP13])**.** *There is an absolute constant* $C < \infty$ *such that, for SRW on any amenable transitive* $d$*-regular graph, for any* $\epsilon > 0$ *and all* $n \geq 1/\epsilon^2$, *we have*

$$\frac{1}{n} \sum_{k=0}^{n} \mathbf{P}\big[\, d(X_0, X_k) < \epsilon\sqrt{n/d} \,\big] \leq C\epsilon\,.$$

Let us point out that one does not really need an actual harmonic Lipschitz embedding of the transitive amenable graph $G$ into Hilbert space to bound the rate of escape from below, but can use the approximate harmonic functions $\varphi_\infty(\mathbf{1}_A)$ or $\varphi_k(\mathbf{1}_A)$ for large Følner sets $A$ directly, and get quantitative bounds: $\mathbf{E}[\, d(X_0, X_n)^2 \,] \geq n/d$, where $G$ is $d$-regular. There is a similar bound for finite transitive graphs: $\mathbf{E}[\, d(X_0, X_n)^2 \,] \geq n/(2d)$ for all $n < 1/(1 - \lambda_2)$, where $\lambda_2$ is the second largest eigenvalue of $G$. See [LeeP13].

▷ **Exercise 10.6.** ** *The tail of the first return time: is it true that* $\mathbf{P}_x[\tau_x^+ > n] \geq c/\sqrt{n}$ *for all groups? (This was a question from Alex Bloemendal in class, and I did not know how hard it was. Later this turned out to be a recent theorem of Ori Gurel-Gurevich and Asaf Nachmias.)*

# 11 Harmonic Dirichlet functions and Uniform Spanning Forests

{s.HDandUSF}

## 11.1 Harmonic Dirichlet functions

{ss.HD}

Recall that the Dirichlet energy for functions $f : V(G) \longrightarrow \mathbb{R}$ is defined as $\sum_{x\sim y} |f(x) - f(y)|^2$. We have seen that the question whether there are non-constant bounded harmonic functions is interesting. It is equally interesting whether a graph has non-constant harmonic Dirichlet functions (i.e., harmonic functions with finite Dirichlet energy, denoted by $\mathsf{HD}(G)$ from now on). This is not to be confused with our earlier result that there is a finite energy flow from $x$ to $\infty$ iff the network is transient: such a flow gives a harmonic function only off $x$. Note, however, that the difference of two different finite energy flows from $x$ with the same inflow at $x$ is a non-constant harmonic Dirichlet function, hence the non-existence of non-constant $\mathsf{HD}$ functions is equivalent to a certain uniqueness of currents in the graph. We will explain this more in a later version of these notes, but, very briefly, the non-trivial harmonic Dirichlet functions lie between the "extremal" currents, the free and wired ones. This connection was first noted by [Doy88]. See [LyPer16, Chapter 9] for a complete account.

The property of having non-constant $\mathsf{HD}$ functions (i.e., $\mathsf{HD} > \mathbb{R}$) is a quasi-isometry invariant, which is due to Soardi (1993), and can be proved somewhat similarly to Kanai's Theorem 6.11. Thus we can talk about a group having $\mathsf{HD} > \mathbb{R}$ or not.

In amenable groups, all harmonic Dirichlet functions are constants, $\mathsf{HD} = \mathbb{R}$. One proof is by [BLPS01], using uniform spanning forests, see the next section.

▷ **Exercise 11.1.** *The free groups $F_k$ with $k \geq 2$ have $\mathsf{HD} > \mathbb{R}$.*

▷ **Exercise 11.2.*** *Any non-amenable group with at least two ends (and hence with a continuum of ends, by Stallings' Theorem 3.3) has $\mathsf{HD} > \mathbb{R}$.*

Although hyperbolic planar tilings are not quasi-isometric to regular trees, being a non-amenable planar graph is a good indication for having $\mathsf{HD} > \mathbb{R}$:

{t.circlepacki

**Theorem 11.1** (He-Schramm [HeS95], Benjamini-Schramm [BenS96a, BenS96b]). *For any locally finite infinite triangulated planar graph, there exists a circle packing representation, i.e., the vertices are represented by disks, and edges are represented by two disks touching. In such a circle packing representation, the union of the triangular faces is either the entire plane, in which case the graph is recurrent, or it is a domain conformally equivalent to a disk, in which case the graph is transient. If the case is the latter (conformally equivalent to a disk), then $\mathsf{HD} > \mathbb{R}$.*

For the circle packing representation, see the original paper [HeS95] or [Woe00, Section 6.D]. For the $\mathsf{HD}$ results, see the papers or [LyPer16, Sections 9.3 and 9.4].

It is worth mentioning another theorem which shows where Kazhdan groups fits.

{t.BeVaKazhdan

**Theorem 11.2** ([BekV97]). *If $\Gamma$ is a Kazhdan group, then $\mathsf{HD} = \mathbb{R}$ for any Cayley graph.*

As $\mathsf{HD}$ is a vector space, we can try to compute its dimension. Of course, when $\mathsf{HD} > \mathbb{R}$, this dimension is infinite, due to the action of $\Gamma$ on the Cayley graph. But there is a notion, the **von Neumann dimension** of Hilbert spaces with group action, denoted by $\dim_\Gamma \mathcal{H}$, which gives dimensions relative to $L^2(\Gamma)$: that is, $\dim_\Gamma(L^2(\Gamma)) = 1$, and in general the values can be in $\mathbb{R}_{\geq 0}$.

{t.BeVaHDbeta}

**Theorem 11.3** ([BekV97]). *We have that* $\dim_\Gamma \mathsf{HD} = \beta_1^{(2)}(\Gamma)$ *for any Cayley graph of* $\Gamma$*, where* $\beta_1^{(2)}(\Gamma)$ *is the first Betti number of the* $L^2$*-cohomology of the group. In particular,* $\dim_\Gamma \mathsf{HD}$ *does not depend on the Cayley graph.*

We will see a probabilistic definition of $\beta_1^{(2)}(\Gamma)$ in the next section, giving a way of measuring the difference between free and wired currents.

## 11.2 Loop-erased random walk, uniform spanning trees and forests

{ss.USF}

For this section, [LyPer16, Chapters 4 and 10] and the fantastic [BLPS01] are our main references.

On a finite graph, we can use the **loop-erased random walk** $\mathsf{LERW}$ to produce a **uniformly random spanning tree** $\mathsf{UST}$ with Wilson's algorithm [Wil96]:

In a connected finite graph $G$, we choose an order of the vertices $x_0, x_1, \dots, x_n$, then produce a path from $x_1$ to $x_0$ by starting a random walk at $x_1$ and stopping it when we hit $x_0$. If the random walk produces some loops, we erase them in the order of appearance. Then we start a walk from $x_2$ till we hit the path between $x_0$ and $x_1$, take the loop-erasure of it, and so on, always walking from $x_i$ till we hit the already existing tree, until all the vertices become part of the tree.

**Theorem 11.4** (Wilson's algorithm). *If $G$ is a finite graph, then Wilson's algorithm (defined above) samples a spanning tree with the uniform distribution (from now on,* $\mathsf{UST}$*, uniform spanning tree).*

This result probably appears to be an absolute miracle. We will not prove it in detail, but here is the main idea. For each vertex $x \in V(G)$, let us generate an i.i.d. random sequence $S_x(i)$, $i = 1, 2, \dots$ of uniform random neighbors of $x$: a sequence of possible next steps the random walk can take from $x$, with the right distribution. We think of this infinite sequence $S_x$ as a stack, where $S_x(1)$ is the topmost element, the only one that is visible, and below $S_x(i)$ is $S_x(i+1)$, ad infinitum. Now, given these stacks, and the sequence of vertices $x_0, x_1, \dots, x_n$, we can generate the $\mathsf{LERW}$ deterministically: starting the walk from $x_0$, we always move according to the visible element of the stack of the vertex where we are currently, and having used it, we remove that element from the stack. Now the key statement is that the final spanning tree is in fact independent of the ordering $x_0, x_1, \dots, x_n$, and can be obtained from the stacks simply by repeatedly removing visible directed cycles from the topmost layer, in any possible order, until no directed cycles are left, which means that the topmost layer gives us a spanning tree. It is now hopefully less surprising that this resulting tree is in fact independent of the cycles removed from above of it, and it is just a uniform random spanning tree.

There is a natural more general version of the algorithm: for any finite irreducible Markov chain $p(x,y)_{x,y \in V}$, if $T$ is a spanning tree on $V$ with a distinguished root vertex $r$, we can orient each edge of $T$ towards $r$, and consider the weight $\Psi(T) = \prod_{(x,y) \in E(T)} p(x,y)$ with the above orientation $(x,y)$. Then the theorem is that Wilson's algorithm, with the $\mathsf{LERW}$ using of course the Markov chain transition probabilities, produces a random tree with distribution proportional to $\Psi$.

This algorithm can be useful to study $\mathsf{UST}$ even on the complete graph $K_n$; on more complicated graphs, such as subgraphs of $\mathbb{Z}^d$, where $\mathsf{LERW}$ can be understood well, it is often the main tool.

▷ **Exercise 11.3** ([LyPer16]). *Use Wilson's algorithm on the complete graph $K_n$ to prove Cayley's formula: the number of trees on n labeled vertices is $n^{n-2}$. (Hint: given any spanning tree $t_{n-1}$ of $K_n$, consider a sequence $t_1 \subset t_2 \subset \cdots \subset t_{n-2} \subset t_{n-1}$ of subtrees such that $t_i$ has $i$ edges. Find recursively the probability that Wilson's algorithm builds this sequence of trees.)*

▷ **Exercise 11.4** ([PerR04b]). * *Using Wilson's algorithm on the complete graph $K_n$, prove the following Rayleigh limit law for distances in a uniform random tree $\mathcal{T}_n$ on $n$ labeled vertices: if $x \neq y$ are two uniformly chosen random vertices, then their graph distance satisfies*

$$\lim_{n\to\infty} \mathbf{P}\big[\, d_{\mathcal{T}_n}(x,y) > t\sqrt{n} \,\big] = \exp(-t^2/2)\,.$$

*(Hint: a bit similarly to the previous exercise, compute $\mathbf{P}[\, d_{\mathcal{T}_n}(x,y) > k \mid d_{\mathcal{T}_n}(x,y) > k-1 \,]$ for each $k$.)*

One corollary to Wilson's algorithm (but it was proved first by Kirchhoff in 1847) is that edge marginals in the $\mathsf{UST}$ have random walk and electric network interpretations. Moreover, the joint probability for $k$ edges being in the $\mathsf{UST}$ is given by a $k \times k$ determinant involving currents. This is the **Transfer-Current Theorem** of Burton and Pemantle [BurtP93] that we will include in a later version of these notes.

In an infinite graph $G$, we can take an exhaustion $G_n$ of it. In each step of the exhaustion, we obtain a $\mathsf{UST}$, denoted by $\mathsf{UST}_n$. Using the electric interpretations, it can be shown that if $S \subset E(G_n)$, then we have $\mathbf{P}\big[\, S \subset \mathsf{UST}_n \,\big] \geq \mathbf{P}\big[\, S \subset \mathsf{UST}_{n+1} \,\big]$, hence the limit exists for all finite $S$. These limit probabilities form a consistent family, since any consistency condition will be satisfied in some large enough $G_n$, hence, the Kolmogorov extension theorem gives us a measure, the free uniform spanning forest, denoted by $\mathsf{FUSF}$. It cannot have cycles, but it is not necessarily connected, since a $\mathsf{UST}_n$ restricted to a finite subgraph has these properties, as well.

If we have two exhaustions, $G_n$ and $G'_n$, then for any $n$ there is an $m$ such that $G_n \subset G'_m$, and vice versa, hence the limit $\lim_n \mathbf{P}\big[\, S \subset \mathsf{UST}_n \,\big] = \mathbf{P}\big[\, S \subset \mathsf{FUSF} \,\big]$ does not depend on the exhaustion. This also implies that the $\mathsf{FUSF}$ of a Cayley graph has a group-invariant law: if we translate a finite edge-set $S$ by some $g \in \Gamma$, and want to show $\mathbf{P}\big[\, S \subset \mathsf{FUSF} \,\big] = \mathbf{P}\big[\, g(S) \subset \mathsf{FUSF} \,\big]$, we can just translate the entire exhaustion we used in the definition.

Another possible approach for a limit object could be that, at each step in the exhaustion $G_n \subset G$, we "wire" all the boundary vertices of $G_n$ into a single vertex, obtaining $G_n^*$. (We can keep or delete the resulting loops, this will not matter.) In this wired $G_n$ we pick a $\mathsf{UST}$ and call it $\mathsf{UST}_n^*$. It can now be shown that if $S \subset E(G_n^*)$, then $\mathbf{P}\big[\, S \subset \mathsf{UST}_n^* \,\big] \leq \mathbf{P}\big[\, S \subset \mathsf{UST}_{n+1}^* \,\big]$, hence we again have a limit, called $\mathsf{WUSF}$: wired uniform spanning forest. This again does not depend on the exhaustion, and on a Cayley graph has a group-invariant law.

Since $G_n^*$ can be considered as having the same edge set as $G_n$, but less vertices, it is intuitively clear that $\mathsf{UST}_n$ stochastically dominates $\mathsf{UST}_n^*$, i.e., they can be coupled so that $\mathsf{UST}_n \supseteq \mathsf{UST}_n^*$. (This domination can be proved using electric networks again.) This implies that, in the limit, $\mathsf{FUSF}$ stochastically dominates $\mathsf{WUSF}$. However, there seems to be no unique canonical coupling on the finite level, hence there does not seem to be a unique coupling in the limit, and the above proof of the group invariance breaks down.

{ex.WFUSF}

▷ **Exercise 11.5.** *** *Does there exist a group invariant monotone coupling between* $\mathsf{FUSF}$ *and* $\mathsf{WUSF}$*?*

A recent result of R. Lyons and A. Thom is that this is the case for sofic groups (see Question 14.2 below). On the other hand, the following theorem shows that the situation in general is not simple.

{t.mester}

**Theorem 11.5** ([Mes13]). *There are two* $\mathrm{Aut}(G)$*-invariant processes* $\mathscr{F}, \mathscr{G} \subseteq E(G)$ *on* $G = \mathbb{T}_3 \times \mathbb{Z}$ *such that there exists a monotone coupling* $\mathscr{F} \subseteq \mathscr{G}$ *(i.e.,* $\mathscr{G}$ *stochastically dominates* $\mathscr{F}$*), but there exists no invariant coupling* $\mathscr{F} \subseteq \mathscr{G}$*.*

The group invariance of $\mathsf{WUSF}$ also follows from Wilson's algorithm rooted at infinity, which is also a reason for considering it to be a very natural object:

**Theorem 11.6.** *On transient graphs, the* WUSF *can be generated by Wilson's algorithm rooted at infinity: Order the vertices of the network as* $\{x_1, x_2, \dots\}$. *Start a loop-erased random walk at* $x_1$. *This walk will escape to infinity. Now start a loop-erased random walk at* $x_2$. *Either this second walk will go to infinity, or it will intersect the first walk at some point (possibly* $x_2$*). If it intersects, end the walk there. Repeat ad infinitum.*

No algorithm is known for generating the FUSF in a similar way, but there are special cases where the WUSF is the same as the FUSF, for example, on all amenable transitive graphs [Häg95], as follows (and see Theorem 11.8 below for the exact condition for equality):

{p.amenUSF}

**Proposition 11.7.** *On any amenable transitive graph,* FUSF = WUSF *almost surely.*

*Proof.* Both FUSF and WUSF consists of infinite trees only. For any such forest $\mathscr{F}$, if $F_n$ is a Følner sequence, then the average degree along $F_n$ is 2. Hence we also have $\mathbf{E}\deg_{\mathsf{WUSF}}(x) = \mathbf{E}\deg_{\mathsf{FUSF}}(x) = 2$ for all $x$. On the other hand, we have the stochastic domination. Hence, by Exercise 13.3 (b), we must have equality. □

The edge marginals of WUSF are given by wired currents, while those of the FUSF are given by free currents. More generally, the Transfer-Current Theorem mentioned above implies that both the FUSF and the WUSF are determinantal processes. Now, since the difference between free and wired currents are given by the non-trivial harmonic Dirichlet functions, see Section 11.1, we get the following result, with credits going to [Doy88, Häg95, BLPS01, Lyo09]:

{t.WFUSFHD}

**Theorem 11.8.** WUSF = FUSF *on a Cayley graph if and only if* HD = $\mathbb{R}$. *More quantitively:*
**(1)** $\mathbf{E}\deg_{\mathsf{WUSF}}(x) = 2$.
**(2)** $\mathbf{E}\deg_{\mathsf{FUSF}}(x) = 2 + 2\beta_1^{(2)}(\Gamma)$.

Due to Wilson's algorithm rooted at infinity, WUSF is much better understood than FUSF, as shown below by a few nice results. On the other hand, the connection of FUSF to $\beta_1^{(2)}(\Gamma)$ makes the free one more interesting. For instance:

▷ **Exercise 11.6** (Gaboriau).*** *Let* $\Gamma$ *be a group, and* $\epsilon$ *be given. Does there exist an invariant percolation* $P_\epsilon$ *with edge marginal less than or equal to* $\epsilon$ *such that* FUSF $\cup P_\epsilon$ *is connected?*

If the answer is yes, it would imply that $\mathrm{cost}(\Gamma) = 1+\beta_1^{(2)}(\Gamma)$, see Section 13.5 below. The corresponding question for WUSF has the following answer:

**Theorem 11.9** ([BLPS01])**.** WUSF $\cup P_{\mathsf{Ber}(\epsilon)}$ *is connected for all* $\epsilon$ *if and only if* $\Gamma$ *is amenable.*

Wilson's algorithm has the following important, but due to the loop erasures, not at all immediate, corollary:

{t.rwcollision

**Theorem 11.10** ([LyPS03])**.** *Let* $G$ *be any network. The* WUSF *is a single tree a.s. iff two independent random walks started at any different states intersect with probability 1.*

The following exercise shows that, on a transitive graph, when the WUSF is not a single tree, then it has infinitely many components. This is not known for the FUSF.

▷ **Exercise 11.7.**

**(a)** *[She06] Let $G(V,E)$ be a transitive graph, $\mathcal{U} \subset V$ an invariant random subset, and $\{X_n\}_{n=0}^{\infty}$ a simple random walk started at $X_0 = x \in V$. Prove that $\{X_n\}$ hits $\mathcal{U}$ with probability 1. (Hint: let $f(x,\mathcal{U})$ be the hitting probability given $\mathcal{U}$, and $F(x)$ be its expectation. What are the harmonicity properties of these functions?)*

**(b)** *Deduce from the previous part that the number of trees in the* WUSF *is more than one but finite with probability zero.*

On $\mathbb{Z}^d$, the following very precise and beautiful result has been proved:

{t.numTrees}

**Theorem 11.11** ([BenKPS04])**.** *Consider* USF = WUSF = FUSF *on $\mathbb{Z}^d$. If $d \leq 4$, then the* USF *is a single tree. If $4 < d \leq 8$ then the* USF *contains infinitely many spanning trees, but all are neighbours, in the sense that given any two of these trees, there is a a vertex in the first, and a vertex in the second which are neighbours. If $8 < d \leq 12$ then the* USF *contains infinitely many spanning trees. It may be the case that there exist two such trees which are not neighbours, but they will have a common neighbour. In general, if $4n < d < 4(n+1)$ then the* USF *contains infinitely many spanning trees, and given any two such trees, there is a chain of trees, each the neighbour of the next, connecting the first to the second, with at most $(n-1)$ trees in the chain, not including the original two trees.*

# 12 Percolation theory

{s.perc}

Percolation theory discusses the connected components (clusters) in random subgraphs of a given graph. The simplest example of this is **Bernoulli**($p$) **bond percolation**, where each edge of a graph is erased (gets closed) with probability $1-p$, and kept (remains open) with probability $p$. Another version is **site percolation**, where we keep and delete vertices, and look at the components induced by the vertices kept. We will usually consider bond percolation, but the results are always very similar in the two cases. (In fact, bond percolation is just site percolation on the so called line graph of the original graph, hence site percolation is more general.) Bernoulli percolation on $\mathbb{Z}^d$ was introduced by Hammersley [Ham57], who gave the name "percolation" because he thought of this as a model of water percolating through a porous stone. By now, it has become a classical subject, one of the main examples of statistical mechanics; the standard textbook is [Gri99]. In particular, percolation in the plane is a key example in contemporary probability, in the study of critical phenomena, see [Wer07]. Percolation on groups beyond $\mathbb{Z}^d$ was initiated by Benjamini and Schramm in 1996 [BenS96c], for which the standard reference is [LyPer16]. In a slightly different direction, $\mathsf{Ber}(p)$ bond percolation on the complete graph $K_n$ on $n$ vertices is the classical Erdős-Rényi (1960) model $G(n,p)$ of probabilistic combinatorics, see [AloS00].

The most important feature of Bernoulli percolation on most infinite graphs is a simple **phase transition**: the existence of a non-trivial **critical density** $p_c \in (0,1)$, above which there is an infinite connected component, and below which there is not. In statistical mechanics in general, the name "phase transition" usually means that there is an important external parameter (most often the temperature), and when this passes a critical point, the system undergoes a dramatic change — usually, some global order appears (e.g., water molecules start getting aligned, and water freezes). The existence of an infinite cluster in percolation may sound like a very simple-minded analogue of global order, but, as we will see in Section 13.1, it is actually not far from being physically relevant.

## 12.1 Basic definitions, examples and tools

{ss.percbasics

Here is the precise definition of the critical density mentioned above:

{d.pc}

**Definition 12.1.** *For* $\mathsf{Ber}(p)$ *percolation on any infinite connected graph* $G(V,E)$*, define* $p_c := \inf\{p : \mathbf{P}_p[\,\exists$ *an infinite cluster*$\,] > 0\} \in [0,1]$*. We will write* $p_c(G, bond)$ *or* $p_c(G, site)$ *if we want to emphasize that we are talking about bond or site percolation. If* $p_c \in (0,1)$*, then we talk about the existence of a phase transition, and the intervals* $[0,p_c)$ *and* $(p_c,1]$ *are called the* **sub-** *and* **supercritical phases***, respectively.*

As a trivial example, note that $p_c(\mathbb{Z}) = 1$, both for bond and site percolation, hence there is no phase transition in this case. As a less trivial example, we will see in a second that

$$1/3 \le p_c(\mathbb{Z}^2, \text{bond}) \le 2/3\,, \tag{12.1}$$ {e.thirds}

so that we do have a non-trivial phase transition here. More generally, we will see that for "non-1-dimensional graphs" one expects $p_c \in (0,1)$, and will prove, among other things, that $p_c(\mathbb{T}_{d+1}, \text{bond}) = p_c(\mathbb{T}_{d+1}, \text{site}) = 1/d$ for the $d+1$-regular tree, and $p_c(\mathbb{Z}^2, \text{bond}) = 1/2$. However, before proving anything, let us examine Definition 12.1 a bit.

First of all, here are some measure theoretic basics, in case you are worried if the definition makes sense. The probability space we have here is the infinite product space $\{0,1\}^{E(G)}$, where 0 or 1 encodes if an edge is closed or open, respectively. In this infinite product space, the sigma-algebra is the one generated by finite cylinder events (defined in general as events of the form $\{\omega : \omega_i \in A_i \text{ for all } i \in I\}$, where $I$ is a finite subset of the coordinates, and each $A_i$ is some fixed measurable set in the $i^{\text{th}}$ coordinate); in other words, it consists of the Borel sets w.r.t. the product topology. Moreover, there is a unique probability measure $\mathbf{P}_p$ on this space that coincides with the obvious measure on finite cylinders, and for every event $\mathcal{A}$ and every $\epsilon > 0$ there exists an event $\mathcal{A}_\epsilon$ depending only on finitely many coordinates (i.e., a finite union of finite cylinder events) such that $\mathbf{P}[\,\mathcal{A} \triangle \mathcal{A}_\epsilon\,] < \epsilon$. This follows from Caratheodory's and Kolmogorov's extension theorems (see [Dur10, Sections A.1, A.2., A.3]). For instance, $\{\exists$ an infinite cluster$\}$ is indeed an event, since it is equal to $\bigcup_{x \in V} \bigcap_{n \ge 1} \{x \longleftrightarrow \partial_V^{\text{out}} B_n(x)\}$. Here, $\{x \longleftrightarrow \partial_V^{\text{out}} B_n(x)\}$ is the event (depending only on finitely many variables) that there exists an open path between $x$ and the outer vertex boundary of the ball $B_n(x)$, as defined at the beginning of Subsection 5.1.

Next, is it obvious that $\mathbf{P}_p[\,\exists$ an infinite cluster$\,]$ is monotone in $p$? Well, a simple proof is by the **standard coupling** of all $\mathsf{Ber}(p)$ percolation measures for $p \in [0,1]$: to each edge $e$ (in the case of bond percolation) assign an independent $U(e) \sim \mathsf{Unif}[0,1]$ variable, and then $\omega_p := \{e \in E(G) : U(e) \le p\}$ is a $\mathsf{Ber}(p)$ bond percolation configuration for each $p$, while $\omega_p \subseteq \omega_{p'}$ for $p \le p'$. In this coupling, $\{\exists \infty \text{ cluster in } \omega_p\} \subseteq \{\exists \infty \text{ cluster in } \omega_{p'}\}$, hence the probability is monotone. Moreover, whether or not an infinite cluster exists does not depend on the states of any finite set of edges, so this is a tail event. As a result, Kolmogorov's 0-1 law (see Theorem 9.20) implies that $\mathbf{P}_p[\,\exists$ an infinite cluster$\,] = 1$ for all $p > p_c$. As we will see, whether we have an infinite cluster exactly at $p_c$ is a very interesting question.

Let us now turn to the proof of (12.1). For the lower bound, note that if the cluster $\mathscr{C}(o)$ of the origin $o = (0,0)$ is infinite, then it contains arbitrary long self-avoiding paths starting from $o$ (by a compactness argument, even contains an infinite one), while the number of such paths of length $n$ in $\mathbb{Z}^2$ is at most $4 \cdot 3^{n-1}$, hence

$$\begin{aligned}
\mathbf{P}_p[\,o \longleftrightarrow \infty\,] &\le \mathbf{P}_p[\exists \text{ open self-avoiding path of length } n \text{ from } o] \\
&\le \mathbf{E}_p[\text{number of open self-avoiding paths of length } n \text{ from } o] \\
&\le 4(3p)^n \le C \exp(-cn)
\end{aligned}$$

for some $c = c(p) > 0$ and $C = C(p) < \infty$ for any $p < 1/3$. This tends to zero as $n \to \infty$, hence there is no infinite cluster almost surely.

For the upper bound, we consider the dual percolation configuration on the planar dual of $\mathbb{Z}^2$, where a dual edge is dual-open if the corresponding primal edge is closed. The idea is that $\mathscr{C}(o)$ can be finite only if $o$ is surrounded by a closed (i.e., dual-open) dual circuit, and for $p$ close to 1, long dual circuits are unlikely. See Figure 12.1. Indeed, for $p > 2/3$,

$$\mathbf{P}_p[\exists \text{ closed self-avoiding dual circuit of length } n \text{ surrounding } o] \leq n\,(3(1-p))^n \leq C\exp(-cn)\,,$$

where the factor $n$ comes from the fact that any such dual circuit must intersect the segment $[0,n]$ on the real axis of the plane. This is summable in $n$, hence, for $N$ large enough, the sum over $n \geq N$ is less than 1, hence with positive probability there is no closed dual circuit longer than $N$ surrounding the origin. This implies that the outer vertex boundary $\partial_V^{\mathrm{out}} B_{N/8}(o)$ must be connected to infinity, hence we have an infinite cluster with positive probability. (Proof: If the union $U$ of the open clusters intersecting $B_{N/8}(o)$ is finite, then take its exterior edge boundary $\partial_E^+ U$, i.e., the boundary edges with an endpoint in the infinite component of $\mathbb{Z}^2 \setminus U$. The edges dual to these edges contain a closed circuit around $o$, with length larger than $N$.) I.e., there is an infinite cluster with positive probability, and we are done. (This counting of dual circuits and using the first moment method is called the **Peierls contour argument**.)

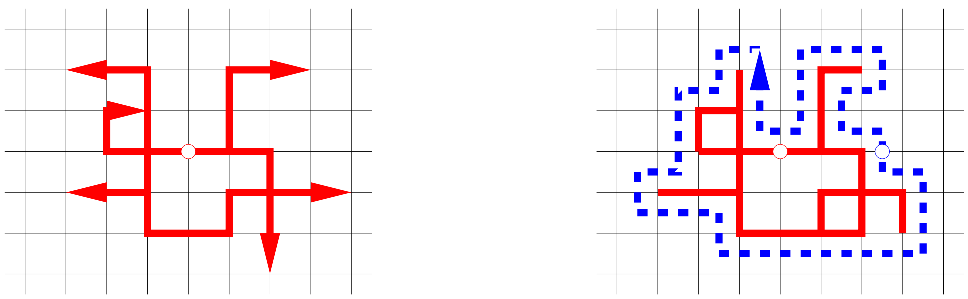


Figure 12.1: Counting primal self-avoiding paths and dual circuits. {f.Peierls}

{l.pc}

**Lemma 12.2.** *Let $\theta_x(p) := \mathbf{P}_p[\,|\mathscr{C}(x)| = \infty\,]$, where $\mathscr{C}(x)$ is the cluster of the vertex $x$. Then, the following definitions for $p_c$ are equivalent:*

$$\begin{aligned} p_c &:= \inf\big\{p : \mathbf{P}_p[\exists \textit{ an infinite cluster}\,] > 0\big\} = \inf\big\{p : \mathbf{P}_p[\exists \textit{ an infinite cluster}] = 1\big\} \\ &:= \inf\big\{p : \exists\, x \in V \textit{ with } \theta_x(p) > 0\big\} = \inf\big\{p : \forall\, x \in V,\ \theta_x(p) > 0\big\}. \end{aligned}$$

*Proof.* We have already discussed that Kolmogorov's 0-1 law, Theorem 9.20, implies that the first two definitions are equivalent. Next, if the probability that an infinite cluster exists is 0 for some $p$, then the probability that any given $x$ is part of an infinite cluster must also be 0. On the other hand, if $\theta_x(p) = 0$ holds for all $x \in V$, then the union bound gives that $\mathbf{P}_p[\exists \text{ an infinite cluster}\,] \leq \sum_x \theta_x(p) = 0$. This yields the equivalence with the third definition. For the equivalence with the fourth definitions, we need that $\theta_x(p) > 0$ happens or fails simultaneously for all $x \in V$ depending only on $p$, which is not clear in a non-transitive graph. We are going to show this in three different ways, just to introduce some basic techniques that will often be useful.

*First proof of $\mathbf{P}_p[\exists\, \infty \text{ cluster}\,] = 1$ implying $\mathbf{P}_p[\,|\mathscr{C}_x| = \infty\,] > 0$ for any $x \in V(G)$.* On any connected graph, the events $E_n := \{B_n(x) \text{ intersects an infinite cluster}\}$ increase to $\{\exists\, \infty \text{ cluster}\}$ as $n \to \infty$, hence $\mathbf{P}_p[\exists\, \infty \text{ cluster}\,] = 1$ implies that there exists some $n = n(x)$ such that

$$\mathbf{P}_p[\,\partial_V^{\mathrm{in}} B_n(x) \longleftrightarrow \infty\,] > 0\,.$$

On the other hand, we also have

$$\mathbf{P}_p[\, x \longleftrightarrow y \text{ inside } B_n(x) \text{ for all } y \in \partial_V^{\text{in}} B_n(x)\,] > 0\,,$$

since this requires just finitely many bits to be open. These two events of positive probability depend on different bits, hence are independent of each other, hence their intersection also has positive probability. Since the intersection implies that $\{x \longleftrightarrow \infty\}$, we are done.

*Second proof.* Bernoulli percolation has a basic property called **finite energy** or **insertion and deletion tolerance**. For the case of bond percolation, for any event $A$ and $e \in E(G)$, let

$$A \cup \{e\} := \{\omega \cup \{e\} : \omega \in A\},$$

$$A \setminus \{e\} := \{\omega \setminus \{e\} : \omega \in A\}.$$

Then the property is that $\mathbf{P}[\,A\,] > 0$ implies that $\mathbf{P}[\,A \cup \{e\}\,] > 0$ and $\mathbf{P}[\,A \setminus \{e\}\,] > 0$, as well, for any $e \in E(G)$. In fact, for $\mathsf{Ber}(p)$ percolation, we have

$$\mathbf{P}_p[\,A \cup \{e\}\,] \geq p\,\mathbf{P}_p[\,A\,] \qquad \text{and} \qquad \mathbf{P}_p[\,A \setminus \{e\}\,] \geq (1-p)\,\mathbf{P}_p[\,A\,]\,.$$

Why? First of all, we can focus only on finite cylinder events $A = \big\{\omega \in \{0,1\}^{E(G)} : \omega(e_i) = 1,\ i = 1\dots,k,\ \omega(f_j) = 0,\ j = 1,\dots,\ell\big\}$, since any measurable event can be approximated by a finite union of such events. And then, $\mathbf{P}_p[\,A \cup \{e\}\,] = \mathbf{P}_p[\,A\,]$ if $e \in \{e_i\}$, $\mathbf{P}_p[\,A \cup \{e\}\,] = p/(1-p)\,\mathbf{P}_p[\,A\,]$ if $e \in \{f_j\}$, and $\mathbf{P}_p[\,A \cup \{e\}\,] = p\mathbf{P}_p[\,A\,]$ if $e \notin \{e_i\} \cup \{f_j\}$. Similarly, $\mathbf{P}_p[\,A \setminus \{e\}\,] \geq (1-p)\mathbf{P}_p[\,A\,]$.

Why is this property called finite energy? One often wants to look at not just Bernoulli percolation, but more general percolation processes: a random subset of edges or vertices. For instance, the Uniform Spanning Tree and Forests from Section 11.2 are bond percolation processes, while the Ising model of magnetization, where the states of the vertices (spins) have a tendency to agree with their neighbours (with a stronger tendency if the temperature is lower) is a site percolation process; see Section 13.1. Such models are often defined by assigning some energy to each (finite) configuration, then giving larger probabilities to smaller energy configurations in some way (typically using a negative exponential of the energy). Then the finite energy property is that any finite modification of a configuration changes the energy by some finite additive constant, and hence the probability by a positive factor. If changing one edge or site changes the probability by a uniformly positive factor, independently of $A$, then the model has uniformly finite energy or uniform deletion/insertion tolerance. Bernoulli percolation and the Ising model are such examples, while the $\mathsf{UST}$ is neither deletion nor insertion tolerant.

We can now easily establish that if $\mathbf{P}_p[|\mathscr{C}(x)| = \infty] > 0$ holds at some $p$ for some $x \in V(G)$, then the same holds for any $y \in V(G)$ in place of $x$. Take a finite path $\gamma$ between $x$ and $y$, and then

$$\mathbf{P}_p[y \longleftrightarrow \infty] \geq \mathbf{P}_p[\gamma \text{ is open},\ x \longleftrightarrow \infty] \geq p^{|\gamma|}\,\mathbf{P}_p[x \longleftrightarrow \infty],$$

by repeated applications of the insertion tolerance shown above.

**Third proof.** Given a partially ordered set $\Omega$, we say that an event $A$ is **increasing** or **upward closed** if $1_A(\omega_1) \leq 1_A(\omega_2)$ whenever $\omega_1 \leq \omega_2$, where $1_A$ is the indicator of the event $A$. For bond percolation on $G$, the natural partially ordered set $\Omega$ is, of course, the set of all subsets of $E(G)$ ordered by inclusion, or $\{0,1\}^{E(G)}$ with coordinate-wise ordering.

{t.HarrisFKG}

**Theorem 12.3** (Harris-FKG inequality)**.** *Increasing events in the Bernoulli product measure are positively correlated:* $\mathbf{P}[\,A \cap B\,] \geq \mathbf{P}[\,A\,]\,\mathbf{P}[\,B\,]$*, or* $\mathbf{E}[\,fg\,] \geq \mathbf{E}[\,f\,]\,\mathbf{E}[\,g\,]$ *if $f$ and $g$ are increasing functions with finite second moments.*

Before proving this, let us go back to the equivalence with the fourth definition in Lemma 12.2 as an example: for any $x, y \in V(G)$, the events $\{x \longleftrightarrow y\}$ and $\{y \longleftrightarrow \infty\}$ are both monotone increasing, hence, by the Harris inequality,

$$\mathbf{P}_p[y \longleftrightarrow \infty] \geq \mathbf{P}_p[x \longleftrightarrow y,\ x \longleftrightarrow \infty] \geq \mathbf{P}_p[x \longleftrightarrow y]\, \mathbf{P}_p[x \longleftrightarrow \infty],$$

finishing our *third proof of Lemma 12.2.* □

*Proof of Theorem 12.3.* This is classical for the case when we have only one bit. Even more generally than just $\mathsf{Ber}(p)$ measure on $\{0,1\}$, one can consider two increasing functions on $\mathbb{R}$ with any probability measure $\mu$, and then the statement is one of Chebyshev's inequalities:

$$\int_{\mathbb{R}} f(x)g(x)\, d\mu(x) \geq \int_{\mathbb{R}} f(x) d\mu(x) \int_{\mathbb{R}} g(x) d\mu(x)\,. \tag{12.2}$$ {e.Cheb}

The proof of this is very easy:

$$\int_{\mathbb{R}} \int_{\mathbb{R}} [f(x) - f(y)]\, [g(x) - g(y)]\, d\mu(x)\, d\mu(y) \geq 0\,,$$

since the integrand is always nonnegative, and then rearranging gives (12.2). The general case, namely a product probability measure $\mu_1 \times \cdots \times \mu_d$ on $\mathbb{R}^d$, can be proved by induction on the dimension. Namely, fixing the first $d-1$ coordinates in any way, we get two increasing functions in the last coordinate, hence can apply the 1-dimensional case, and this averaging out w.r.t. the last coordinate yields a $(d-1)$-dimensional problem that we know by induction. If you cannot see the details for yourself, look it up in [LyPer16, Section 7.2]. □

The explanation for the name Harris-FKG is that Harris proved it for product measures (just what we stated) in [Har60], a fundamental paper for percolation theory. But there is a generalization of Theorem 12.3 for dependent percolation processes where the energy of a configuration is given by certain nice local functions favouring agreement: this is the **FKG inequality**, due to Fortuin, Kasteleyn and Ginibre (1971). For instance, the Ising model of magnetization does satisfy this inequality. The proof of this must be very different from what is sketched above, since the measure does not have the product structure anymore. Indeed, we will do this in Theorem 13.1, via a beautiful Markov chain coupling argument, due to Holley. For now, see [Wik10b] for a very short introduction to the general FKG inequality, and [GeHM01] for a proper one. At the opposite end from the FKG inequality, in the $\mathsf{UST}$, as usual for determinantal measures, increasing events supported on disjoint sets are negatively correlated [LyPer16, Section 4.2].

Before going further, here are two exercises to make sure you understand what the measurability of having an infinite cluster means:

▷ **Exercise 12.1.** *Let $G(V, E)$ be any bounded degree infinite graph, and $S_n \nearrow V$ an exhaustion by finite connected subsets. Is it true that, for $p > p_c(G)$, we have*

$$\lim_{n \to \infty} \mathbf{P}_p[\, \textit{largest cluster for percolation inside } S_n \textit{ is the subset of an infinite cluster}\,] = 1\,?$$

▷ **Exercise 12.2.** *Show that for percolation on any infinite graph, the event $\{$there are exactly three infinite clusters$\}$ is Borel measurable.*

We will often consider random subgraphs that are not given by independent Bernoulli variables. On the one hand, there are a lot of natural processes of this type, such as the Uniform Spanning Forests from Section 11.2 and the Ising model from Section 13.1. On other hand, as we will see later on, studying percolation in this generality will also turn out to be useful for Bernoulli percolation itself.

When looking at general percolation processes on transitive graphs $G$, one usually wants that it is an **invariant percolation** process: a random subset of edges or vertices whose law is invariant under the automorphism group of $G$, or at least under a subgroup of $\mathrm{Aut}(G)$ that still acts transitively on $G$. (If $G$ is a Cayley graph of a group $\Gamma$, then one often considers $\Gamma$-invariance, invariance under the translations by the group elements, even if this is only a proper subgroup of $\mathrm{Aut}(G)$. E.g., the graph $\mathbb{Z}^2$ has more automorphisms than just translations.) Intuitively, invariance means that the process "looks the same" from all vertices of $G$. For a formal definition, we first have to say how the automorphism group acts on configurations and events. Thinking of a site percolation configuration $\omega$ as a function on the vertices, it is natural to pull it back via automorphisms: for any $\gamma \in \mathrm{Aut}(G)$ and $x \in V(G)$, we let $\omega^\gamma(x) := \omega(\gamma(x))$. Note that if $\delta$ is another automorphism, then $\omega^{\delta\circ\gamma}(x) = \omega(\delta(\gamma(x))) = (\omega^\delta)(\gamma(x)) = (\omega^\delta)^\gamma(x)$; thus, while $\mathrm{Aut}(G)$ with the composition $\circ$ as group operation acts naturally from the left on $G$, it will act from the right on configurations (and that is why we chose to write $\omega^\gamma$). An event $\mathcal{A}$ can be considered as a set of configurations, hence $\mathcal{A}^\gamma := \{\omega^\gamma : \omega \in \mathcal{A}\}$. E.g., for the event $\mathcal{A} = \{|\mathscr{C}_x| > 100\}$, we have $\mathcal{A}^\gamma = \{|\mathscr{C}_{\gamma^{-1}(x)}| > 100\}$. And now we can finally say that a percolation measure $\mathbf{P}$ is invariant under a group $\Gamma$ of automorphisms if $\mathbf{P}[\,\mathcal{A}\,] = \mathbf{P}[\,\mathcal{A}^\gamma\,]$ for any event $\mathcal{A}$ and any $\gamma \in \Gamma$. For the event $\mathcal{B} = \{\exists\, x : |\mathscr{C}_x| > 100\}$, we have $\mathcal{B}^\gamma = \mathcal{B}$, hence this is an invariant event in any percolation process. In general, we say that an event $\mathcal{A}$ is $\Gamma$-invariant for a percolation measure $\mathbf{P}$ if $\mathbf{P}[\,\mathcal{A} \triangle \mathcal{A}^\gamma\,] = 0$ for all $\gamma \in \Gamma$.

The next exercise gives a non-trivial example of an invariant bond percolation on the 3-regular tree $\mathbb{T}_3$:

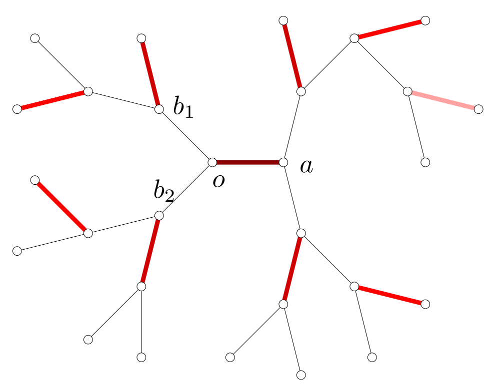



Figure 12.2: Constructing the unique invariant random perfect matching on $\mathbb{T}_3$. {f.invmatch}

{ex.invmatch}

▷ **Exercise 12.3.** *We are going to define a random perfect matching $\omega$ of $\mathbb{T}_3$. Fix a root $o \in V(\mathbb{T}_3)$. Choose uniformly at random one of the three edges emanating from $o$ to be in $\omega$. Denote this chosen edge by $(o, a)$, while the other two by $(o, b_i)$, $i = 1, 2$. Now, for each $b_i$, choose one of the two edges emanating from $b_i$ not in the direction of $o$ uniformly at random to be in $\omega$. The two edges emanating from $a$ but not containing $o$ can of course not be in $\omega$. Continue inductively for each vertex, for the pair of edges emanating from the vertex not in the direction of $o$. See Figure 12.2.*

*Prove that the resulting perfect matching has a distribution that is invariant under all the automorphisms* $\mathrm{Aut}(\mathbb{T}_3)$*, despite the fact that the construction itself used a specific vertex as the origin. In fact, this is the unique* $\mathrm{Aut}(\mathbb{T}_3)$*-invariant distribution on perfect matchings of* $\mathbb{T}_3$*.*

A basic property that an invariant percolation process might have is ergodicity:

{l.ergodic}

**Lemma 12.4.** Ber$(p)$ *bond (or site) percolation on any infinite transitive graph $G$ is* **ergodic**: *any invariant event has probability 0 or 1. In fact, instead of transitivity, it is enough that there is an edge with an infinite orbit (and then it is easy to see that every edge has an infinite orbit).*

*Proof.* As discussed after Definition 12.1, any measurable event $\mathcal{A}$ can be approximated by a finite union of finite cylinder events, i.e., for any $\epsilon > 0$ there is an event $\mathcal{A}_\epsilon$ depending only on finitely many coordinates such that $\mathbf{P}_p[\mathcal{A} \triangle \mathcal{A}_\epsilon] < \epsilon$. Then, there exists a "large enough" $\gamma = \gamma_\epsilon \in \mathrm{Aut}(G)$ such that the two sets of relevant coordinates for $\mathcal{A}_\epsilon$ and $\mathcal{A}_\epsilon^\gamma$ are disjoint, hence $\mathbf{P}_p[\mathcal{A}_\epsilon \cap \mathcal{A}_\epsilon^\gamma] = \mathbf{P}_p[\mathcal{A}_\epsilon]^2$. On the other hand, $\mathbf{P}_p[\mathcal{A} \cap \mathcal{A}^\gamma] = \mathbf{P}_p[\mathcal{A}]$ by invariance. Altogether, $\mathbf{P}_p[\mathcal{A}]$ can be approximated well by $\mathbf{P}_p[\mathcal{A}]^2$:

$$\left|\mathbf{P}_p[\mathcal{A}] - \mathbf{P}_p[\mathcal{A}]^2\right| = \left|\mathbf{P}_p[\mathcal{A} \cap \mathcal{A}^\gamma] - \mathbf{P}_p[\mathcal{A}]^2\right| \le \left|\mathbf{P}_p[\mathcal{A}_\epsilon \cap \mathcal{A}_\epsilon^\gamma] - \mathbf{P}_p[\mathcal{A}_\epsilon]^2\right| + 2\epsilon + 2\epsilon = 4\epsilon\,.$$

By letting $\epsilon \to 0$, we get that $\mathbf{P}_p[\mathcal{A}]^2 = \mathbf{P}_p[\mathcal{A}]$, hence $\mathbf{P}_p[\mathcal{A}] \in \{0,1\}$. □

There are natural invariant percolations that are not ergodic: a trivial example is taking the empty or the full vertex set with probability 1/2 each; a less trivial one is the Ising model on $\mathbb{Z}^2$ at low temperature with a free boundary condition, which gives a mixture of the plus and minus phases (see Section 13.1). We will be concerned here mainly with ergodic measures, although in Theorem 12.19 for instance, the difference between ergodicity and non-ergodicity will be the main point.

Similarly to invariance, ergodicity can also be understood w.r.t. not the entire automorphism group of the transitive graph, but a subgroup of it. If the group gets smaller, then invariance becomes a weaker condition, while ergodicity becomes stronger. For instance, on $\mathbb{Z}^2$, consider the bond percolation that is the collection of all the vertical lines, or all the horizontal ones, with probability 1/2 each. This is $\mathrm{Aut}(\mathbb{Z}^2)$-invariant and $\mathrm{Aut}(\mathbb{Z}^2)$-ergodic, but not ergodic under the $\mathbb{Z}^2$ translations. The bond percolation that is simply the collection of all the vertical lines deterministically is not $\mathrm{Aut}(\mathbb{Z}^2)$-invariant, but $\mathbb{Z}^2$-invariant and $\mathbb{Z}^2$-ergodic. Another example is the site percolation process on $\mathbb{Z}^2$ in which, with probability 1/2, the even vertices (i.e., $(x,y) \in \mathbb{Z}^2$ with $x+y$ being even) are open, the odd vertices are closed, and vice versa with probability 1/2. Even though this measure looks degenerate the same way as the previous horizontal/vertical example, it is both $\mathrm{Aut}(\mathbb{Z}^2)$-invariant and $\mathbb{Z}^2$-ergodic. The property it fails to possess is **total ergodicity**: there is a finite index subgroup under which it is not ergodic. In other words, it is **periodic**. The following exercise also utilizes this anomaly:

{ex.ErgodUnion

▷ **Exercise 12.4.** *Show that the union of two independent ergodic percolation processes may not be ergodic.*

A strengthening of ergodicity is **mixing**. For an invariant percolation $\mathbf{P}$ on an infinite transitive graph $G$, the definition is that for any two events $A, B$, and any $\epsilon > 0$, there is a finite set $K_\epsilon \subset \mathrm{Aut}(G)$ such that if $\gamma \notin K_\epsilon$, then $\left|\mathbf{P}[A \cap \gamma(B)] - \mathbf{P}[A]\mathbf{P}[B]\right| < \epsilon$. For general graphs, an appropriate definition is that, for any $k \ge 1$, if $A$ and $B_n$ are cylinder events such that each $B_n$ depends on at most $k$ variables (sites or bonds), and all the variables on which $B_n$ depends are at distance at least $n$ from the variables on which $A$ depends, then $\lim_{n\to\infty} \left|\mathbf{P}[A \cap B_n] - \mathbf{P}[A]\mathbf{P}[B_n]\right| = 0$. For transitive graphs, the two definitions are easily seen to be equivalent. It is also easy to see, by copying the proof of Lemma 12.4, that mixing implies ergodicity. Simple examples for ergodic but non-mixing percolations are periodic percolations. A non-periodic example comes from a key example of ergodic theory, irrational rotations:

▷ **Exercise 12.5.** *Consider the following site percolation on $\mathbb{Z}$: let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ fixed, $U \sim \mathsf{Unif}[0,1]$ random, and let $\omega_n := 1$ if $U + n\alpha \pmod 1$ is in $[0, 1/2)$, and $\omega_n := 0$ otherwise. Show that this is $\mathbb{Z}$-invariant, ergodic, but not mixing.*

Most results and conjectures below will concern percolation on transitive graphs, but let us point out that there are always natural modifications (with almost identical proofs) for quasi-transitive graphs, i.e., when $V(G)$ has finitely many orbits under $\mathrm{Aut}(G)$.

Here is a basic application of ergodicity and insertion tolerance: {l.01infty}

**Lemma 12.5.** *In any ergodic insertion tolerant invariant percolation process on any infinite transitive graph, the number of infinite clusters is an almost sure constant, namely 0, 1, or $\infty$.*

*Proof.* For any $k \in \{0, 1, 2, \dots, \infty\}$, the event $\{$the number of infinite clusters is $k\}$ is translation invariant, hence it has probability 0 or 1 by ergodicity. I.e., the number of infinite clusters is an almost sure constant. Now, assume that this constant is $1 < k < \infty$. By the measurability of the number of clusters, for any $c < 1$ there exists an integer $r$ such that the probability that the ball $B_r(o)$ intersects at least two infinite clusters is at least $c$. But then, by insertion tolerance, we can change everything in $B_r(o)$ to open, resulting in an event with positive probability, on which the number of infinite clusters is at most $k-1$: a contradiction. □

The number of infinite clusters in Bernoulli percolation is a basic and interesting topic. We will prove later in this section that $p_c(\mathbb{T}_{k+1}, \mathrm{site}) = p_c(\mathbb{T}_{k+1}, \mathrm{bond}) = 1/k$ for the $k+1$-regular tree, and it is pretty obvious that, for all $p \in (p_c(\mathbb{T}_k), 1)$, there are a.s. infinitely many infinite clusters: on the other side of each closed edge or site neighboring an open cluster, there is a new $k$-ary tree that almost surely has infinite clusters again. On the other hand, having infinitely many infinite clusters turns out to be impossible in amenable transitive graphs such as $\mathbb{Z}^d$ (see Section 5.1 for the definition). In rough terms, there is not enough space for many infinite clusters to coexist without getting glued together (made possible by insertion tolerance). This was first proved for Bernoulli percolation on $\mathbb{Z}^d$ in [AiKN87], but the "right" proof is the following one: {t.BurtonKeane

**Theorem 12.6** (Burton-Keane [BurtK89])**.** *For any insertion and deletion tolerant ergodic invariant percolation on any amenable transitive graph, the number of infinite clusters is almost surely 0 or 1.*

*Proof.* Given a percolation configuration $\omega \subseteq E(G)$, a **furcation point** is a vertex in an infinite cluster $\mathscr{C}$ whose removal from $\mathscr{C}$ would break it into at least three infinite connected components.

Now, if there were infinitely many infinite clusters, then, using insertion and deletion tolerance, they could be joined, somewhat similarly to the proof of Lemma 12.5 above, to get that a given vertex is a furcation point with positive probability. However, we have to do this joining procedure now a bit more carefully. First of all, we are going to describe a procedure where we insert and delete edges depending on what the configuration is in a given ball $B_r(o)$. But this is fine, since there are only finitely many possibilities for the configuration inside $B_r(o)$, hence there is one with positive probability, and we can just stick to inserting and deleting the edges corresponding to this one configuration, regardless of the actual percolation configuration, and this will work with positive probability. So, here is one such possible procedure. Take $r$ so large that the ball $B_r(o)$ contains at least three infinite clusters with positive probability. Pick three infinite clusters $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3$ that intersect $B_r(o)$. Delete everything inside $B_r(o)$, then one-by-one insert paths $\gamma_i$ of length $r$ from $o$ to $\mathscr{C}_i$, for $i = 1, 2, 3$, such that the union of these three paths is a tree. The branch point of this tree that is the farthest from $o$ is a furcation point. (For bond percolation, this

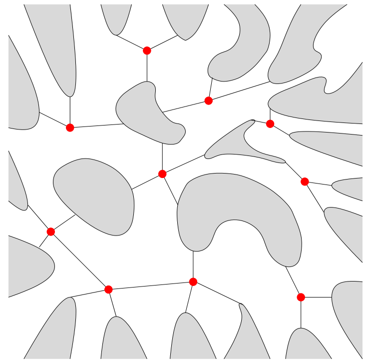


Figure 12.3: Furcation points in an infinite cluster. {f.trifur}

will actually be a trifurcation point, whose deletion results in exactly three infinite clusters, but for site percolation processes, we may get more than three.) So, the fixed ball $B_r(o)$ contains a furcation point with positive probability, and hence the probability that any given vertex of the transitive graph is a furcation point must be positive.

Now that we know that furcation points exist, inside a large Følner set $F_n$ the expected number of furcation points $X_n$ grows linearly with $|F_n|$. On the other hand, deterministically, the inner vertex boundary of $F_n$ has to be at least $X_n+2$. (This requires a little thought. See Figure 12.3 and Exercise 12.6.) Combining these two facts, $|\partial F_n|$ should grow linearly with $|F_n|$, contradicting the definition of a Følner sequence. □

{ex.BKmethod}

▷ **Exercise 12.6.**

**(a)** *Show carefully the claim we used in the Burton-Keane theorem: if $\mathscr{C}_\infty$ denotes the union of all the infinite clusters in some percolation on $G$, and $U \subset V(G)$ is finite, then the size of $\mathscr{C}_\infty \cap \partial_V^{\mathrm{out}} U$ is at least the number of furcation points of $\mathscr{C}_\infty$ in $U$, plus 2.*

**(b)** *Extend the above proof of the Burton-Keane theorem to quasi-transitive graphs.*

Here is a small generalization of the second part of the Burton-Keane proof. Recall that the ends of a graph were defined in Exercise 3.3.

{ex.BK3ends}

▷ **Exercise 12.7.** *In any invariant percolation on any transitive amenable graph, show that there cannot be clusters with at least three ends.*

Without insertion tolerance, the first part of the Burton-Keane proof clearly breaks down. And in fact, the conclusion becomes false:

▷ **Exercise 12.8.** *Give a $\mathbb{Z}^2$-invariant and $\mathbb{Z}^2$-ergodic bond percolation on $\mathbb{Z}^2$ with infinitely many $\infty$ clusters. Then give one that is also invariant under interchanging the two coordinates.*

More challenging are the following exercises:

▷ **Exercise 12.9.** *

**(a)** *Give an $\mathrm{Aut}(\mathbb{Z}^2)$-invariant and $\mathbb{Z}^2$-ergodic bond percolation on $\mathbb{Z}^2$ with exactly two $\infty$ clusters. (Hint: construct a one-ended subtree and consider it together with its dual.)*

**(b)** *Give a deletion-tolerant version of part (a). (Hint: try deleting edges from the previous construction randomly with tiny probabilities.)*

▷ **Exercise 12.10.*** *Is there an ergodic deletion-tolerant $\mathbb{Z}^2$-invariant bond percolation on $\mathbb{Z}^2$ with infinitely many infinite clusters? (Hint: use Exercise 12.7 for the dual percolation.)*

We will now discuss another key example: percolation on a regular tree. This turns out to be a special case of a classical object: the cluster of a fixed vertex in $\mathsf{Ber}(p)$ percolation on a $k+1$-regular tree $\mathbb{T}_{k+1}$ is very close to a **Galton-Watson process** with offspring distribution $\xi = \mathsf{Binom}(k,p)$. A usual GW process is a random tree where we start with a root in the zeroth generation, then each individual in the $n$th generation gives birth to an independent number of offspring with distribution $\xi$, together giving the $(n+1)$th generation. For $\mathsf{Ber}(p)$ percolation on $\mathbb{T}_{k+1}$, the root has a special offspring distribution $\mathsf{Binom}(k+1,p)$. However, this difference does not affect questions like the existence of infinite clusters (which you should be able to verify).

So, to find $p_c(\mathbb{T}_{k+1})$, it is enough to understand when a GW process survives with positive probability. A standard method for this is the following. Consider the probability generating function of the offspring distribution,

$$f(s) := \mathbf{E}[\, s^{\xi}\,] = \sum_{k\geq 0} \mathbf{P}[\, \xi = k\,] s^k. \tag{12.3}$$ {e.GWpgf}

Notice that if $Z_n$ is the size of the $n$th generation, then $\mathbf{E}[\, s^{Z_n}\,] = f^{\circ n}(s)$, therefore we have

$$q := \mathbf{P}[\,\text{extinction}\,] = \lim_{n\to\infty} \mathbf{P}[\, Z_n = 0\,] = \lim_{n\to\infty} f^{\circ n}(0)\,. \tag{12.4}$$ {e.GWq}

Assuming that $\mathbf{P}[\, \xi = 1\,] \neq 1$, the function $f(s)$ is strictly increasing and strictly convex on $[0,1]$, so (12.4) easily implies that $q$ is the smallest root of $s = f(s)$ in $[0,1]$. (Just draw a figure of how $f(s)$ may look like and what the iteration $f^{\circ n}(0)$ does!)

{ex.GWpgf}

▷ **Exercise 12.11.** *Using the above considerations, show that if $\mathbf{E}[\, \xi\,] \leq 1$ but $\mathbf{P}[\, \xi = 1\,] \neq 1$, then the GW process almost surely dies out. Deduce that $p_c(\mathbb{T}_{k+1}) = 1/k$ and $\theta(p_c) = 0$.*

A quite different strategy is the following:

{ex.GWMG}

▷ **Exercise 12.12.***

- **(a)** *Consider a GW process with offspring distribution $\xi$, $\mathbf{E}\xi = \mu$. Let $Z_n$ be the size of the nth level, with $Z_0 = 1$, the root. Show that $Z_n/\mu^n$ is a martingale, and using this, assuming $\mathbf{P}[\, \xi = 1\,] \neq 1$, that $\mu \leq 1$ implies that the GW process dies out almost surely.*
- **(b)** *On the other hand, if $\mu > 1$ and $\mathbf{E}[\, \xi^2\,] < \infty$, first show that $\mathbf{E}\big[\, Z_n^2\,\big] \leq C(\mathbf{E}Z_n)^2$. (Hint: use the conditional variance formula $\mathbf{D}^2[Z_n] = \mathbf{E}\big[\, \mathbf{D}^2[Z_n \mid Z_{n-1}]\,\big] + \mathbf{D}^2\big[\mathbf{E}[\, Z_n \mid Z_{n-1}\,]\big]$.)*
  *Then, using the* **Second Moment Method**, *namely, if $X \geq 0$ a.s., then $\mathbf{P}[\, X > 0\,] \geq (\mathbf{E}X)^2/\mathbf{E}[\, X^2\,]$ (prove this or see (12.14) in Section 12.3), deduce that the GW process survives with positive probability.*
- **(c)** *Extend the above to the case $\mathbf{E}\xi = \infty$ or $\mathbf{D}\xi = \infty$ by a truncation $\xi \mathbf{1}_{\xi<K}$ for $K$ large enough.*

Let us describe yet another strategy, which is robust enough to be used in different finite random graph models (see, e.g., [vdHof13]). Consider the following exploration process of any rooted tree, with the children of each vertex being ordered. During the process, vertices will be active, inactive, or neutral, and active vertices will be ordered. In the 0th step, start with the root as the only active vertex; all other vertices are neutral. In the $(i+1)$th step, examine the children of the first vertex $v$ in the active list after the $i$th step, put these children at the beginning of the active list, and turn $v$ inactive. If the tree is finite, the process will end up with all vertices being inactive; if the tree is infinite, then the process will run forever, with the vertices of the first infinite ray all being put in the active list. See Figure 12.4.

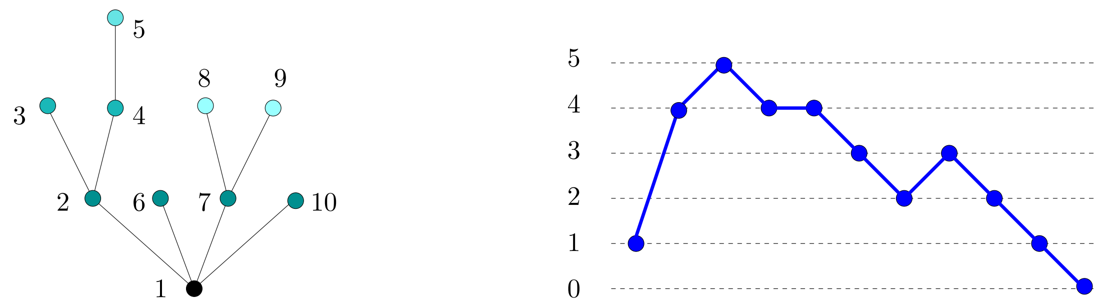



Figure 12.4: On the left, the shades of vertex colours show the order in which the sets of children are put into the active list. The labels on the vertices show the order in which they turn inactive, which is in fact the Depth First Search order. On the right, the steps of the walk are indexed by the vertices as they turn inactive, while the height shows the current number of active vertices. {f.GWexplore}

The above exploration process is well-suited to construct a subcritical GW tree fully or a supercritical GW tree partially. Indeed, let $\{X_i : i \geq 1\}$ be a sequence of iid variables from the offspring distribution $\xi$, which will be the number of children being put into the active list at each step, let $S_0 = 1$, and then $S_{i+1} = S_i + X_{i+1} - 1$ is the size of the active list after the $(i+1)^{\text{th}}$ step. This is a random walk with iid increments distributed as $\xi - 1$. If $\mathbf{E}\xi < 1$, then the walk has a negative drift, hence $S_n = 0$ will eventually happen almost surely. If $\mathbf{E}\xi = 1$ but $\mathbf{P}[\,\xi = 1\,] \neq 1$, then the walk is recurrent, by Exercise 12.13 below. If $\mathbf{E}\xi > 1$, then the walk has a positive drift, hence with positive probability will go to infinity without ever reaching zero: there exists $K > 0$ such that $\mathbf{P}\big[\min_{i\geq 0} S_i > -K\big] > c_1 > 0$, while we get above $K$ before dying with some probability $c_2 > 0$, hence we never die with probability at least $c_1 c_2 > 0$. In summary, we have reached the same conclusion as in Exercises 12.11 and 12.12. {ex.ChuFu}

▷ **Exercise 12.13** (Discrete Chung-Fuchs theorem)**.**

**(a)** *Let $S_n = X_1 + \cdots + X_n$ be a random walk with i.i.d. jumps in $\mathbb{Z}^d$. Show that, for any $m \in \mathbb{Z}_+$,*

$$\sum_{n=0}^{\infty} \mathbf{P}\big[\,\|S_n\|_\infty \leq m\,\big] \leq (2m+1)^d \sum_{n=0}^{\infty} \mathbf{P}[\,S_n = \underline{0}\,]\,.$$

*(Hint: for any $v \in \mathbb{Z}^d$ with $\|v\|_\infty \leq m$, the event $\{S_n = v\}$ can be decomposed as $\bigcup_{\ell=0}^{n} \{S_n = v,\ T_v = \ell\}$, according to the first hitting time of $v$.)*

**(b)** *Assume now that $d = 1$, and that $S_n$ satisfies the following Weak Law of Large Numbers: $S_n/n \xrightarrow{p} 0$. (For instance, $\mathbf{E}X_i = 0$ is enough; see [Dur10, Theorem 2.2.9].) Notice that, for any $m \in \mathbb{Z}_+$ and $A > 0$ arbitrarily large, part (a) implies*

$$\sum_{n=0}^{\infty} \mathbf{P}[\,S_n = 0\,] \geq \frac{1}{2m+1} \sum_{n=0}^{\infty} \mathbf{P}\big[\,|S_n| \leq m\,\big] \geq \frac{1}{2m+1} \sum_{n=0}^{\lfloor Am \rfloor} \mathbf{P}\big[\,|S_n| \leq n/A\,\big]\,.$$

*Deduce from this and the WLLN that the expected number of returns to 0 is infinite. Conclude that the walk is recurrent.*

The "real" Chung-Fuchs theorem says that even for arbitrary real-valued random walks on $\mathbb{R}$, the Weak Law of Large Numbers $S_n/n \xrightarrow{p} 0$ is sufficient for neighbourhood-recurrence; that is, $\mathbf{P}\big[\,\|S_n\|_\infty < \epsilon \text{ i.o.}\,\big] = 1$ for any $\epsilon > 0$; see [Dur10, Section 4.2]. The proof follows the strategy of the previous exercise, with the extra difficulty that the small errors in several "close" returns may build up or cancel each other out, and thus it is not obvious that the equivalences

$$\sum_{n=0}^{\infty} \mathbf{P}\big[\, \|S_n\|_\infty < \epsilon \,\big] = \infty \iff \mathbf{P}\big[\, \exists\, n : \|S_n\|_\infty < \epsilon \,\big] = 1 \iff \mathbf{P}\big[\, \|S_n\|_\infty < \epsilon \text{ i.o.} \,\big] = 1 \tag{12.5}$$ {e.returns}

hold. Nevertheless, this is indeed the case; see [Dur10, Theorem 4.2.6].

{ex.CauchyRW}

▷ **Exercise 12.14** (Chung-Fuchs is not sharp)**.** *Recall the standard Cauchy density:* $\frac{1}{\pi(1+x^2)}$ *for* $x \in \mathbb{R}$. *A natural way to obtain it is to take the* $x$ *coordinate of the intersection of a line at a uniform random orientation from the origin of the* $(x, y)$ *plane with the line* $y = 1$.

- **(a)** *Let* $X$ *and* $Y$ *be independent standard normal variables. Show that* $X/Y$ *has Cauchy distribution.*
- **(b)** *Prove that the harmonic measure on the line* $y = 1$ *for 2-dimensional Brownian motion* $(X_t, Y_t)$ *started at the origin is given by the Cauchy distribution. (Hint: we want the distribution of* $X_\tau$, *where* $\tau = \min\{t : Y_t = 1\}$. *Show first, using the reflection principle for Brownian motion, that* $\sqrt{\tau}$ *has the same distribution as* $1/|Y_1|$.*)*
- **(c)** *From part (b) and the scale invariance of 2-dim BM, deduce that a sum of i.i.d. Cauchy variables,* $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$, *has the distribution of* $n\xi_1$. *Conclude that the Cauchy distribution does not satisfy* $S_n/n \xrightarrow{p} 0$.
- **(d)** *Show that for any interval* $(-\epsilon, \epsilon)$, *the expected number of returns of the Cauchy walk* $S_n$ *is infinite. By (12.5), this implies that* $S_n$ *is recurrent.*

The following result is a striking example of a probabilistic duality between sub-critical and super-critical GW trees. It will be exploited in our proof of the Erdős-Rényi giant cluster phase transition, Theorem 12.23. For two-dimensional percolation, using that there is also a planar duality, there is a close analogy to this (just like in Figure 12.1), but for higher dimensions and other Cayley graphs, only a vague intuition remains.

▷ **Exercise 12.15** (Galton-Watson duality)**.*** *Either by computing generating functions directly, or by using a Doob transform argument (see Lemma 6.13), show the following duality of super- and sub-critical GW trees. Consider a supercritical* $\mathsf{GW}_\xi$ *tree, with generating function* $f(z) = \mathbf{E}[\, z^\xi \,]$ *and extinction probability* $q = f(q)$. {ex.GWduality}

- **(a)** *Condition* $\mathsf{GW}_\xi$ *on non-extinction, and take the subtree of those vertices that have an infinite line of descent. Show that this is a GW tree with offspring distribution* $\xi^*$, *where*

$$\mathbf{P}[\, \xi^* = k \,] = \sum_{j=k}^{\infty} \binom{j}{k} (1-q)^{k-1} q^{j-k} \mathbf{P}[\, \xi = j \,] \,.$$

  *Deduce that the generating function* $f^*(z) = \mathbf{E}[\, z^{\xi^*} \,]$ *is obtained by taking the part of* $f(z)$ *in the* $[q, 1]^2$ *square and rescaling it to the square* $[0, 1]^2$. *Note that* $\mathbf{P}[\, \xi^* = 0 \,] = 0$ *and* $\mathbf{E}\xi^* = \mathbf{E}\xi$.
- **(b)** *Condition* $\mathsf{GW}_\xi$ *on extinction. Show that we get a subcritical GW tree, with offspring distribution* $\tilde{\xi}$, *whose generating function* $\tilde{f}(z)$ *is obtained by taking the part of* $f(z)$ *in the* $[0, q]^2$ *square and rescaling it to the square* $[0, 1]^2$. *Note that* $\mathbf{E}\tilde{\xi} = f'(q) < 1$.

  *In particular, if we condition the GW tree with offspring distribution* $\mathsf{Poisson}(\lambda)$ *on extinction, where* $\lambda > 1$, *then we get a GW tree with offspring distribution* $\mathsf{Poisson}(\mu)$ *with* $\mu < 1$, *where* $\lambda e^{-\lambda} = \mu e^{-\mu}$.

For the study of percolation on general locally finite rooted trees $T$, Russ Lyons defined an "average **branching number**" [Lyo90]:

$$\mathsf{br}(T) := \sup\left\{\lambda \geq 1 : \inf_{\Pi} \sum_{e\in\Pi} \lambda^{-|e|} > 0\right\}, \tag{12.6}$$ {e.br}

where the infimum is taken over all cutsets $\Pi \subset E(T)$ separating the root $o \in V(T)$ from infinity, and $|e|$ denotes the distance of the edge $e$ from $o$. The following exercises help digest what this notion measures:

▷ **Exercise 12.16.** *Let $T$ be a locally finite infinite tree with root $o$.*

**(a)** *Show that $\mathsf{br}(T)$ does not depend on the choice of the root $o$.*

**(b)** *Show that the $d+1$-regular tree has $\mathsf{br}(\mathbb{T}_{d+1}) = d$.*

**(c)** *Define the lower growth rate of $T$ by $\underline{\mathsf{gr}}(T) := \liminf_n |T_n|^{1/n}$, where $T_n$ is the set of vertices at distance exactly $n$ from $o$. Show that $\mathsf{br}(T) \leq \underline{\mathsf{gr}}(T)$.*

▷ **Exercise 12.17.** *Find the branching number of the following two trees (see Figure 12.5):*

**(a)** *The quasi-transitive tree with degree 3 and degree 2 vertices alternating.*

**(b)** *The so-called 3-1-tree, which has $2^n$ vertices on each level $n$, with the left $2^{n-1}$ vertices each having one child, the right $2^{n-1}$ vertices each having three children; the root has two children.*

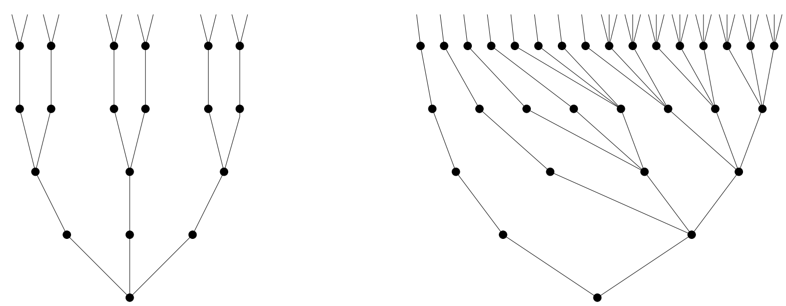


Figure 12.5: A quasi-transitive tree and the 3-1 tree. {f.twotrees}

A clear motivation for definition (12.6) is given by the following interpretation. Let us denote the set of non-backtracking infinite rays starting from $o$ by $\partial T$, the boundary of the tree, equipped with the metric $d(\xi,\eta) := e^{-|\xi\wedge\eta|}$, where $\xi \wedge \eta$ is the last common vertex of the two rays, and $|\xi \wedge \eta|$ is its distance from $o$. Then, basically by definition,

$$e^{\dim_H(\partial T,d)} = \mathsf{br}(T) \qquad \text{and} \qquad e^{\underline{\dim}_M(\partial T,d)} = \underline{\mathsf{gr}}(T)\,,$$

where $\dim_H$ is Hausdorff dimension and $\underline{\dim}_M$ is lower Minkowski dimension. Since Hausdorff dimension has, over the past hundred years, proved a better notion than Minkowski dimension, the branching number ought to be a better way of measuring average branching than growth.

A first sign of the usefulness of branching number is that Lyons proved (via a refinement of the second moment method that we saw in Exercise 12.12) that $p_c(T) = 1/\mathsf{br}(T)$. This easily implies that $\mathsf{br}(\mathsf{GW}_\xi) = \mathbf{E}\xi$ a.s. on nonextinction, which is another "proof" that this is a good definition of average branching. Indeed, if we perform $\mathsf{Ber}(p)$ percolation on $\mathsf{GW}_\xi$, then the component of the root is simply another GW tree, with mean offspring $p\mathbf{E}\xi$. Hence $p_c(\mathsf{GW}_\xi) = 1/\mathbf{E}\xi$ a.s. on nonextinction, giving the formula $\mathsf{br}(\mathsf{GW}_\xi) = \mathbf{E}\xi$.

The branching number turns out to govern the behavior of most stochastic processes on trees. For instance, if we take $\lambda$-biased **homesick random walk**, where the edge going towards the starting point $o$ has weight $\lambda$ compared to the outgoing edges that have weight 1, the walk is recurrent for $\lambda > \mathsf{br}(T)$ and transient for $\lambda < \mathsf{br}(T)$.

▷ **Exercise 12.18.** *Prove the last statement on transience and recurrence using flows and cutsets in electric networks (see Theorem 6.9 and Proposition 6.10).*

Lyons has found even closer connections between percolation and random walks on general trees; see [Lyo92], and, of course, [LyPer16]. For some more examples of the branching number playing a role in processes on trees, see [Vir02] regarding the speed of random walks, [EvKPS00] regarding the Ising model, and [BalPP06] regarding bootstrap percolation.

## 12.2 Percolation on infinite groups: $p_c$, $p_u$, unimodularity, and general invariant percolations

{ss.percinfty}

Now, back to the question of $p_c \in (0,1)$, the lower bound of (12.1) generalizes easily: if $G$ has maximal degree $d$, then $p_c(G) \geq 1/(d-1)$. See Exercise 12.19 for a more general statement. However, the upper bound relied on planar duality, hence it is certainly less robust. The straightforward generalization that does hold is Exercise 12.20 (a).

{ex.pcfactor}

▷ **Exercise 12.19.** *Assume that $\pi : G' \twoheadrightarrow G$ is a topological covering between infinite graphs, or in other words, $G$ is a factor graph of $G'$. Show that $p_c(G') \leq p_c(G)$.*

{ex.cutsets}

▷ **Exercise 12.20.**

**(a)** *Show that if in a graph $G$ the number of minimal edge-cutsets (a subset of edges whose removal disconnects a given vertex from infinity, minimal w.r.t. containment) of size $n$ is at most $\exp(Cn)$ for some $C < \infty$, then $p_c(G) \leq 1 - \epsilon(C) < 1$.*

**(b)** *Fix $o \in V(G)$ in a graph with maximal degree $\Delta$. Prove that the number of connected sets $o \in S \subset V(G)$ of size $n$ is at most $\Delta(\Delta-1)^{2n-3}$. (Hint: any $S$ has a spanning tree, and one can go around a tree visiting each edge twice.) Conclude that $\mathbb{Z}^d$, $d \geq 2$, has an exponential bound on the number of minimal cutsets. In particular, $p_c(\mathbb{Z}^d) < 1$, although we already knew that from $\mathbb{Z}^2 \subseteq \mathbb{Z}^d$.*

▷ **Exercise 12.21.** ** *Let $\lambda(G) := \limsup_{n\to\infty} |\{S \subset V(G) : o \in S \text{ connected}, |S| = n\}|^{1/n}$ denote the exponential growth rate of the number of "lattice animals". We saw in part (b) of the previous exercise that $\lambda(G) \leq (\Delta-1)^2$ for any graph of maximal degree $\Delta$. What is the smallest possible upper bound here? Kesten's book [Kes82] has a beautiful argument proving $\lambda(G) \leq (\Delta-1)e$: for site percolation at $p = 1/(\Delta-1)$, write the probability that the cluster of $o$ is finite using lattice animals and their outer vertex boundaries.*

Now, here is a fundamental and natural conjecture on the non-triviality of $p_c$:

{c.pc<1}

**Conjecture 12.7** (**Non-triviality of the percolation phase transition** [BenS96c]). *If $G$ is a non-one-dimensional graph, i.e., it satisfies an isoperimetric inequality $IP_{1+\epsilon}$ for some $\epsilon > 0$ (defined in Section 5.1), then $p_c < 1$.*

This has been verified in many cases, including all known groups. (As a little philosophical exercise, the Reader is invited to ponder what "all known groups" may mean.) We start with Cayley graphs of finitely presented groups.

Let $G(V,E)$ be a bounded degree graph, and let $\partial G$ be the set of its ends, see Section 3.1. Let $\mathsf{CutCon}(G)$ be the **cutset-connectivity** of $G$: the smallest $t \in \mathbb{Z}_+$ such that any minimal edge-cutset $\Pi$ between any two elements of $V(G) \cup \partial G$ is $t$-connected in the sense that in any non-trivial partition of it into two subsets, $\Pi = \Pi_1 \cup \Pi_2$, there are $e_i \in \Pi_i$ whose distance in $G$ is at most $t$. In words, it is the largest "gap" that a minimal edge-cutset can have between its elements. For instance, $\mathsf{CutCon}(\mathbb{Z}^2) = 1$, $\mathsf{CutCon}(\text{hexagonal lattice}) = 2$, and $\mathsf{CutCon}(\mathbb{T}_d) = 1$ for all $d$, despite the fact that cutsets separating a vertex from infinity (i.e., from the set of all ends of $\mathbb{T}_d$) do not have bounded connectivity. We are going to prove that if $\Gamma$ is a finitely presented group, then any finitely generated Cayley graph $G(\Gamma, S)$ has $\mathsf{CutCon}(G) < \infty$; the first proof [BabB99] used cohomology groups, but there is a few-line linear algebra proof by Ádám Timár [Tim07], which we now present. As we will see, this is closely related to having an exponential bound on the number of minimal cutsets. Timár also proved that $\mathsf{CutCon}(G) < \infty$ and having an exponential bound are both quasi-isometry invariants.

**Proposition 12.8.** *Note that each cycle in a graph $G$ can be viewed as a configuration of edges, i.e., an element of $\{0,1\}^{E(G)}$, or even as an element of the vector space $\mathbb{F}_2^{E(G)}$. The* **cycle space** *of $G$ over $\mathbb{F}_2$ is then the linear subspace spanned by all the cycles.* {pr.CutCon}

*Assume that the cycles of length at most $t$ generate the entire cycle space of $G$. (This is obviously the case if $G$ is the Cayley graph of a finitely presented group.) Then $\mathsf{CutCon}(G) \leq t/2$.*

*Proof.* Let $x, y \in V(G) \cup \partial G$ and $\Pi = \Pi_1 \cup \Pi_2$ a minimal edge cutset separating them, with a nontrivial partition. The minimality of $\Pi$ implies that each edge $e \in \Pi$ has the property that one of the endpoints of $e$ is connected to $x$, the other endpoint to $y$, by paths that are disjoint from $\Pi$. Therefore, if $x$ (or $y$) is an end, we can find a vertex $x'$ ($y'$) such that there is a path between $x$ and $x'$ ($y$ and $y'$) in $G \setminus \Pi$. Otherwise, simply set $x' := x$ ($y' := y$).

Again by the minimality of $\Pi$, there is a path $P_i$ between $x'$ and $y'$ that avoids $\Pi_{3-i}$, for $i = 1, 2$. Now look at $P_1 + P_2 \in \mathbb{F}_2^{E(G)}$. This is clearly in the cycle space, so, we can write $P_1 + P_2 = \sum_{c \in K} c$ for some finite set $K$ of cycles of length at most $t$. Let $K_1 \subseteq K$ be the subset of cycles that are disjoint from $\Pi_2$, and write

$$\theta := P_1 + \sum_{c \in K_1} c = P_2 + \sum_{c \in K \setminus K_1} c \,.$$

We see from the first sum that this $\theta \subset E(G)$ is disjoint from $\Pi_2$. But the only odd degree vertices in it are $x'$ and $y'$, so it must contain a path from $x'$ to $y'$. That must intersect $\Pi$, but is disjoint from $\Pi_2$, hence it intersects $\Pi_1$. But in the second sum, $P_2$ is disjoint from $\Pi_1$, so there must be some cycle in $K \setminus K_1$ that intersects $\Pi_1$. It also intersects $\Pi_2$, and its length is at most $t$, which proves the claim. □

We saw in Exercise 3.4 that an infinite group can have 1, 2, or continuum many ends. Firstly, having two ends is equivalent to being a finite extension of $\mathbb{Z}$, which implies, of course, being one-dimensional and having $p_c = 1$. Secondly, if $\Gamma$ is a finitely presented group with one end, then the proposition says that minimal cutsets between a vertex $o$ and infinity have bounded connectivity. This implies that once we fix an edge in a minimal cutset of size $n$, there are only exponentially many possibilities for the cutset, by Exercise 12.20 (b). So, if we show that there is a set $S_n$ of edges with size at most exponential in $n$ such that each such cutset must intersect it, then we get an exponential upper bound on the number of such cutsets, and Exercise 12.20 (a) gives $p_c(G) < 1$. We claim that the ball $B_{An}(o)$ is a suitable choice for $S_n$ for $A > 0$ large enough. If we assume that $\Gamma$ has at least quadratic volume growth, then for any finite set $K \subset V(G)$ the Coulhon-Saloff-Coste isoperimetric inequality, Theorem 5.11, says that $|\partial_E K| \geq c\sqrt{|K|}$, since $\rho(|K|) \leq C\sqrt{|K|}$. Therefore, if the component of $o$ in $G \setminus \Pi$ contains $B_{An}(o)$, then

$|\Pi| \geq c\sqrt{|B_{An}(o)|} \geq c' An$. So, by choosing $A$ large enough, a cutset of size $n$ around $o$ must intersect $B_{An}(o)$, as desired. And $\Gamma$ must indeed have volume growth at least quadratic: if it has polynomial growth, then by Gromov's Theorem 10.1 it is almost nilpotent, hence it has an integer growth rate that cannot be 1, because then $\Gamma$ would be a finite extension of $\mathbb{Z}$ and it would have two ends.

We are left with the case that $\Gamma$ has continuum many ends. But then it is non-amenable by Exercise 5.4. And, for any nonamenable graph $G$, we have $p_c(G) \leq 1/(h+1) < 1$, even without finitely presentedness or even transitivity, as proved by [BenS96c] (or see [LyPer16, Theorem 6.24]):

{pr.nonamenexp

**Proposition 12.9.** *For bond percolation on any graph $G$ with edge Cheeger constant $h > 0$, we have $p_c(G) \leq 1/(h+1) < 1$. Moreover, for any $p > 1/(h+1)$, we have $\mathbf{P}_p[\, n < |\mathscr{C}(o)| < \infty \,] < \exp(-cn)$ for some $c = c(p) > 0$.*

*Proof.* Fix an arbitrary ordering of the edges, $E(G) = \{e_1, e_2, \dots\}$. Explore the cluster of a fixed vertex $o$ by taking the first $e_i$ with an endpoint in $o$, examining its state, extending the cluster of $o$ by this edge if its open, then taking the first unexamined $e_i$ with one endpoint in the current cluster of $o$ and the other endpoint outside, and so on. If the full $\mathscr{C}(o)$ is finite, then this process stops after exploring an open spanning tree of the cluster plus its closed boundary and possibly other closed edges between vertices in the spanning tree. If $\mathscr{C}(o) = n$, then we have found $n-1$ open edges and at least $|\partial_E \mathscr{C}(o)| \geq hn$ closed edges in this process of examining i.i.d. $\mathsf{Ber}(p)$ variables. But if $p > 1/(h+1)$, then this is exponentially unlikely, by a standard large deviation estimate, say, Proposition 1.8. Furthermore, with positive probability there is no such $n$, just as a biased random walk on $\mathbb{Z}$ might never get back to the origin. $\square$

It is not very hard to show (but needs some "stochastic domination" techniques we have not discussed yet) that $p_c(G(\Gamma, S)) < 1$ is independent of the generating set. Moreover, it is invariant under quasi-isometries.

It is also known [LyPer16, Theorem 7.24] that any group of exponential growth has $p_c < 1$; see Exercise 12.23 below for an example. On the other hand, recall that groups of polynomial growth are all almost nilpotent and hence finitely presented, see Exercise 4.3 and Section 10.1. This means that among groups, Conjecture 12.7 remains open only for groups of intermediate growth. However, all known examples of such groups, see Section 15.1, have Cayley graphs containing a copy of $\mathbb{Z}^2$, hence the conjecture holds also for them [MuPa01].

The general Conjecture 12.7 is also known to hold for planar graphs of polynomial growth that have an embedding into the plane without vertex accumulation points [Koz07].

{ex.LLCG}

▷ **Exercise 12.22.** *Recall from Exercise 5.6 that the lamplighter group $\Gamma = \mathbb{Z}_2 \wr \mathbb{Z}$ with generating set $S = \{\mathsf{R}, \mathsf{Rs}, \mathsf{L}, \mathsf{sL}\}$ is the Diestel-Leader graph $\mathsf{DL}(2,2)$. Show that $\mathsf{CutCon}(\mathsf{DL}(2,2)) = \infty$.*

{ex.LLpc}

▷ **Exercise 12.23.** *Show that the lamplighter group with generating set $S = \{\mathsf{L}, \mathsf{R}, \mathsf{s}\}$ has $p_c(G(\Gamma, S)) < 1$, by finding a subgraph in it isomorphic to the so-called Fibonacci tree $F$: a directed universal cover of the directed graph with vertices $\{1, 2\}$ and edges $\{(12), (21), (22)\}$. (There are two directed covers, with root either 1 or 2.) Find the exponential volume growth $\lim_n (\log |B_n|)/n$ of $F$, and just quote the result (due to Russ Lyons) that for periodic trees like $F$, this being positive implies $p_c(F) < 1$.*

Now, the next natural question is what happens exactly at $p_c$. For the case of regular trees, in Exercises 12.11 and 12.12, and in the random walk exploration approach of Figure 12.4, we have seen several different reasons why critical Galton-Watson processes die out. We will see in Section 12.4 that critical percolation on nice planar lattices, such as $\mathbb{Z}^2$, also dies out. A main conjecture in percolation theory is that the critical behaviour $\theta(p_c) = 0$ should hold in general:

{c.thetapc}

**Conjecture 12.10** (**Continuity of the percolation phase transition** [BenS96c])**.** *On any transitive graph with $p_c < 1$, $\theta(p_c) = 0$.*

*We will show in Corollary 12.14 that $\theta(p)$ is right-continuous with the only possible discontinuity being at $p = p_c$, hence the conjecture is equivalent to $\theta(p)$ being continuous everywhere.*

That transitivity is needed can be seen from the case of general trees:

▷ **Exercise 12.24.** *Consider a spherically symmetric tree $T$ where each vertex on the $n^{th}$ level $T_n$ has $d_n \in \{k, k+1\}$ children, such that $\lim_{n\to\infty} |T_n|^{1/n} = k$, but $\sum_{n=0}^{\infty} k^n/|T_n| < \infty$. Using the second moment method, show that $p_c = 1/k$ and $\theta(p_c) > 0$.*

There are several reasons to believe Conjecture 12.10. As we have partly seen and will partly discuss below, it is known to hold in the extremal cases: $\mathbb{Z}^2$ and other planar lattices on one hand, and regular trees and non-amenable Cayley graphs on the other. It is also known on $\mathbb{Z}^d$ with $d \geq 19$. One of the most famous problems of statistical physics is the case of $\mathbb{Z}^3$ — and here one can test the conjecture by computer simulations. Also, simple models of statistical physics tend to have a continuous phase transition. (In physics language, the **phase transition is of second order**, i.e., key observables like $\theta(p)$ are continuous but non-differentiable at $p_c$.) However, it should be noted that the $\mathsf{FK}(p,q)$ random cluster models (discussed in Section 13.1), a dependent bond percolation model whose $q = 1$ case is just Bernoulli bond percolation, conjecturally have a **first order** (i.e., discontinuous) phase transition on the lattice $\mathbb{Z}^2$ for $q > 4$, at $p_c(q) = \frac{\sqrt{q}}{1+\sqrt{q}}$. Consequently, there are probably no serious philosophical reasons why the percolation phase transition must be continuous.

My main intuition why the conjecture should hold is the following. A general phenomenon, which will be discussed in more detail in Section 12.5, is that infinite clusters usually inherit the rough geometric properties of the underlying transitive graph. Of course, this should be interpreted properly: e.g., infinite clusters at $p < 1$ never satisfy any $IP_\psi$ with $\limsup_{x\to\infty} \psi(x) = \infty$, since, due to randomness, there will be arbitrary large bad pieces; nevertheless, some weaker properties, e.g., satisfying a so-called anchored isoperimetric inequality $IP^*_\psi$ should already be inherited (proved, for instance, on $\mathbb{Z}^d$ for all $p > p_c$, and on any finitely presented Cayley graph for $p$ close enough to 1, in [Pet08]). Now, if a transitive graph $G$ has an infinite cluster at some $p < 1$, then it also satisfies $IP_2$, since transitive graphs of polynomial growth are all quasi-isometric to nilpotent groups and hence have integer volume growth and isoperimetry (see Section 5.3 and the end of Section 10.1). Assuming the above idea of inheriting geometric properties, any infinite cluster $\mathscr{C}$ should also have some sort of an at least 2-dimensional structure, which, in line with Conjecture 12.7, should ensure $p_c(\mathscr{C}) < 1$. But percolation on a percolation cluster is just percolation at a smaller value, hence this would mean that the interval of $p$ values with an infinite cluster is open from the right, and therefore we have $\theta(p_c(G)) = 0$.

A small issue with turning this argument into an actual proof is that an anchored isoperimetric inequality $IP^*_d$ with $d \geq 2$, mentioned above, does not actually imply $p_c < 1$, hence the survival of something stronger would be needed. A much bigger issue is that it is hard to imagine how the survival of these geometric properties at some $p$ could be proved without such a detailed knowledge of percolation at that value that would imply the non-existence of infinite clusters anyway. For instance, on $\mathbb{Z}^d$, the situation is as follows. It is known that the half-space $\mathbb{Z}^d_+ = \mathbb{Z}^{d-1} \times \mathbb{Z}_+$ satisfies $p_c(\mathbb{Z}^d_+) = p_c(\mathbb{Z}^d)$ and that $\theta(p_c) = 0$ holds there [BarGN91]. Therefore, if there is an infinite cluster at $p_c$, which must be unique by the Burton-Keane Theorem 12.6, then it will look quite strange, something like a giant swirl: any coordinate hyperplane will cut it into finite pieces. This sounds crazy, but nobody has succeeded in the past decades in ruling out this scenario.

The attentive reader may ask: how does the above intuitive "inherited geometry" argument break for those $\mathsf{FK}(p,q)$ models with a discontinuous phase transition? See the end of Section 13.1 for an answer that reveals that the $\mathsf{FK}(p,q)$ model should be blamed for this failure: for $q > 4$, the density of bonds itself is discontinuous in the parameter $p$. Hence, even though at $p_c(q)$ one expects to have an infinite cluster $\mathscr{C}$ that resembles $\mathbb{Z}^2$ well, and thus $\mathsf{FK}(1-\epsilon,q)$ on $\mathscr{C}$ should have an infinite cluster for $\epsilon > 0$ small enough, there exists no $p < p_c(q)$ such that $\mathsf{FK}(p,q)$ on $\mathbb{Z}^2$ would stochastically dominate $\mathsf{FK}(1-\epsilon,q)$ on $\mathscr{C}$. Consequently, this failure for the FK model does not weaken the credibility of the argument for the case of Bernoulli percolation.

Let us also mention that for quite a few years it was not known even on $\mathbb{Z}^d$ whether

$$p_T := \sup\{p : \mathbf{E}_p|\mathscr{C}_o| < \infty\} \qquad = \text{ or } < \qquad p_H := p_c = \sup\{p : \theta(p) = 0\}\,, \tag{12.7}$$ {e.pTpH}

where $T$ and $H$ are in the honor of Temperley and Hammersley. First it was proved in [AiN84, Lemma 3.1] that $\mathbf{E}_{p_H}|\mathscr{C}_o| = \infty$, even in the stronger form that $\lim_{p\nearrow p_H}\mathbf{E}_p|\mathscr{C}_o| = \infty$. Then $p_T = p_H$ was established independently by Menshikov [Men86] and Aizenman-Barksy [AiB87], and it is since then that simply $p_c$ has been used universally. In fact, they did not only prove finite expectation of the cluster size below $p_c$, but also the exponential decay

$$\mathbf{P}_p[\, o \longleftrightarrow \partial B_n(o)\,] \le e^{-n\sigma(p)} \qquad \text{for all } p < p_c\,. \tag{12.8}$$ {e.expdec}

The arguments of [AiN84, AiB87] work with little modifications for all transitive graphs; see [Koz11, v1 on arXiv] or [AntV08]. This approach is based on certain partial differential inequalities; we will give a flavor of it in the proof of Theorem 12.29 in Section 12.4. For a full treatment of all the above results on $\mathbb{Z}^d$, see [Gri99, Chapters 5 and 6].

▷ **Exercise 12.25.** *Consider the d-ary canopy tree of Figure 14.1: infinitely many leaves on level 0, grouped into d-tuples, each tuple having a parent on level $-1$, which are grouped again in d-tuples, and so on, along infinitely many levels. Show that $p_T = 1/\sqrt{d}$ while $p_c = 1$.*

▷ **Exercise 12.26.** *Prove using subadditivity that $\sigma(p) := \lim_{n\to\infty} \frac{-1}{n}\log \mathbf{P}_p[\, o \longleftrightarrow \partial B_n(o)\,]$ exists in any transitive graph.*

Let us now see in more detail for what graphs Conjecture 12.10 has actually been proved. Besides regular trees, it is quite classical also on nice planar lattices, where even the critical value is known in some cases, e.g., $p_c(\mathbb{Z}^2, \text{bond}) = p_c(\mathsf{TG}, \text{site}) = 1/2$, where $\mathsf{TG}$ is the triangular grid — these are Kesten's theorems from 1980. Percolation, and more generally, statistical mechanics at criticality in the plane is indeed a miraculous world, exhibiting **conformal invariance**. This field has seen amazing progress in the past few years, due to the work of Schramm, Smirnov, and others; see Section 12.4 for a bit more details. As an appetizer: although the value of $p_c$ for percolation is a lattice-dependent local quantity (see Conjecture 14.18), critical percolation itself should be **universal**: "viewed from far", it should look the same and be conformally invariant on any planar lattice, even though criticality happens at different densities. For instance, **critical exponents** such as $\theta(p) = (p-p_c)^{5/36+o(1)}$ as $p \searrow p_c$ should always hold, though the existence and values of such exponents are proved only for site percolation on the triangular lattice.

Conjecture 12.10 is also known for $\mathbb{Z}^d$ with $d \ge 19$, using a perturbative Fourier-type expansion method called the Hara-Slade **lace expansion**. Again, this method is good enough to calculate, e.g., $\theta(p) =$

$(p-p_c)^{1+o(1)}$ as $p \searrow p_c$. This, together with other critical exponents, are conjecturally shared by many-many transitive graphs, namely all **mean-field** graphs; this should include Euclidean lattices for $d > 6$, all non-amenable groups, and probably most groups "in between". This is known only in few cases, like "highly non-amenable" graphs (including regular trees, of course); again see Section 12.4 for more information.

The following important general theorem settles Conjecture 12.10 for all non-amenable groups. On the other hand, the cases of $\mathbb{Z}^d$, with $3 \le d \le 18$, and of all non-Abelian amenable groups remain wide open.

{t.BLPSpc}

**Theorem 12.11** (Benjamini-Lyons-Peres-Schramm [BLPS99a, BLPS99b])**.** *For any non-amenable Cayley (or more generally, unimodular transitive) graph, percolation at $p_c$ dies out.*

We will give a rough sketch of a proof of this theorem, but first we need to define when a transitive graph $G$ is called **unimodular**. Let $\Gamma$ be the automorphism group of $G$, and $\Gamma_x$ be the stabilizer of the vertex $x$. Then the condition is that $|\Gamma_x y| = |\Gamma_y x|$ for all $(x,y) \in E(G)$. So, not only the graph looks the same from all vertices, but it does not have $+$ and $-$ directions in which it looks different on a quantitative level. (But different directions are still possible in a finer sense: see Exercise 12.27). A simple example of a non-unimodular transitive graph is the **grandparent graph**: take a 3-regular tree, pick an end of it, and add an edge from each vertex to its grandparent towards the fixed end. Another class of examples are the Diestel-Leader graphs $\mathsf{DL}(k,\ell)$ with $k \neq \ell$, defined just before Exercise 5.6. A more complicated example can be found in [Tim06b].

The importance of unimodularity is the Mass Transport Principle, discovered by Olle Häggström: $G$ is unimodular iff for any random function $f(x,y,\omega)$, where $x,y \in V(G)$ and $\omega \in \Omega$ is the randomness, that is diagonally invariant, i.e., $f\big(\gamma(x),\gamma(y),\omega^\gamma\big) = f(x,y,\omega)$ for any $\gamma \in \mathrm{Aut}(G)$, with the definition of $\omega^\gamma$ given before Lemma 12.4, we have

$$\sum_{y \in V} \mathbf{E} f(x,y,\omega) = \sum_{y \in V} \mathbf{E} f(y,x,\omega). \qquad (12.9)$$ {e.MTP}

We think of $f(x,y,\omega)$ as the the mass sent from $x$ to $y$ when the situation is given by $\omega$, e.g., the percolation configuration. Then the MTP means the conservation of mass on average.

Given a non-unimodular transitive graph, one can easily construct a deterministic mass transport rule that does not satisfy (12.9): for instance, in the grandparent graph, if every vertex sends mass 1 to each grandchild, then the outgoing mass is 4, but the incoming mass is only 1. In the other direction, given a unimodular graph, a simple resummation argument gives (12.9). As the simplest case, Cayley graphs do satisfy the Mass Transport Principle (and hence are unimodular): using $F(x,y) := \mathbf{E} f(x,y,\omega)$, we have $\sum_{x \in G} F(o,x) = \sum_{x\in G} F(x^{-1},o) = \sum_{y \in G} F(y,o)$, where, in the first equality we used that multiplying from the left by a group element is a graph-automorphism, and in the second equality we used that $x \mapsto x^{-1}$ is a self-bijection of $\Gamma$. See [LyPer16, Sections 8.1, 8.2] for more details on MTP and unimodularity.

{ex.+-}

$\triangleright$ **Exercise 12.27.**

**(a)** *Give an example of a unimodular transitive graph $G$ such that there exist neighbours $x,y \in V(G)$ such that there is no graph-automorphism interchanging them.*

**(b)** * *Can you give an example with a Cayley graph?*

In some sense, MTP is a weak form of averaging. Exercise 12.31 below is a good example of this. And it is indeed weaker than usual averaging:

$\triangleright$ **Exercise 12.28** (Soardi-Woess 1990)**.** *Show that amenable transitive graphs are unimodular.*

A typical way of using the MTP is to show that whenever there exist some invariantly defined special points of infinite clusters in an invariant percolation process, then there must be infinitely many in any infinite cluster. For instance:

{ex.trifurcati

▷ **Exercise 12.29.** *Recall the notion of a trifurcation point from the proof of the Burton-Keane Theorem 12.6.*

**(a)** *In any invariant percolation process on any transitive graph $G$, show that the number of trifurcation points is either almost surely 0 or $\infty$. (If the process is also ergodic, then either a.s. 0, or a.s. $\infty$.)*

**(b)** *In an invariant percolation process on a unimodular transitive graph $G$, show that almost surely the number of trifurcation points in each infinite cluster is 0 or $\infty$.*

**(c)** *Give an invariant percolation on a non-unimodular transitive graph with infinitely many trifurcation points a.s., but only finitely many in each infinite cluster.*

The previous exercise about trifurcation points and Exercise 3.4 about ends of groups have the following joint generalization:

{ex.perc3ends}

▷ **Exercise 12.30.**

**(a)** *In an invariant percolation process on a unimodular transitive graph $G$, show that almost surely the number of ends of each infinite cluster is 1 or 2 or continuum.*

**(b)** *Give an invariant percolation on a non-unimodular transitive graph that has infinite clusters with more than two but finitely many ends.*

A more quantitative use of the MTP is the following:

{ex.marginonam

▷ **Exercise 12.31.** *Let $\omega$ be any invariant bond percolation on a transitive unimodular graph $G$. Let $\alpha_K$ be the average degree inside a finite subgraph $K \subset G$, and let $\alpha(G)$ be the supremum of $\alpha_K$ over all finite $K$. Then clearly $\alpha(G) + h_E(G) = \deg_G(o)$, where $h_E$ is the Cheeger constant $\inf_S |\partial_E S|/|S|$. Show that if $\mathbf{E}[\deg_\omega(o)] > \alpha(G)$, then $\omega$ has an infinite cluster with positive probability.*

In words, since $\alpha(G)$ is the supremum of the average degrees in finite subgraphs, it is not surprising that a mean degree larger than $\alpha(G)$ implies the existence of an infinite cluster. The MTP is needed to pass from spatial averages to means w.r.t. invariant measures.

{ex.tree2/3mar

▷ **Exercise 12.32.** *The bound of Exercise 12.31 is tight: show that for the set of invariant bond percolations on the 3-regular tree $\mathbb{T}_3$ without an infinite cluster, the supremum of edge-marginals is 2/3. (Hint: the complement of the unique invariant perfect matching has density 2/3 and consists of $\mathbb{Z}$ components.)*

Exercise 12.31 is, of course, vacuous if $G$ is amenable. And non-amenability is in fact essential for the conclusion that a large edge-marginal implies the existence of an infinite cluster, as can be seen from the following two exercises:

▷ **Exercise 12.33.** *For any $\epsilon > 0$, give an example of an invariant bond percolation process $\omega$ on $\mathbb{Z}^d$ with only finite clusters a.s., but $\mathbf{E}[\deg_\omega(o)] > 2d - \epsilon$.*

{ex.marginamen

▷ **Exercise 12.34.** *Generalize the previous exercise to all amenable Cayley graphs. (Hint: close the boundaries of a set of randomly translated Følner sets, with a "density" high enough to ensure having only finite clusters, but low enough to ensure high edge marginals. The statement holds not only for Cayley graphs, but also for all amenable transitive graphs; however, for any two vertices in a Cayley graph, there is a unique natural automorphism moving one to the other, which makes the proof a bit easier.)*

Note that Exercises 12.31 and 12.34 together give a percolation characterization of amenability, similar to Kesten's random walk characterization, Theorem 7.3:

{t.marginal}

**Theorem 12.12** (Percolation characterization of amenability [BLPS99a])**.** *A unimodular transitive graph is amenable iff for any $\epsilon > 0$ there is an invariant bond percolation $\omega$ with finite clusters only and edge-marginal $\mathbf{P}[\, e \in \omega \,] > 1 - \epsilon$.*

A related characterization can be found in Exercise 13.24. On the other hand, maybe surprisingly, a very similar condition characterizes not non-amenability but a stronger property, Kazhdan's (T): see Theorem 12.19 below.

*Sketch of proof of Theorem 12.11.* We will base our sketch on [BLPS99b]. First of all, by the ergodicity of Bernoulli percolation (Lemma 12.4) and by Lemma 12.5, we need to rule out the case of a unique and the case of infinitely many infinite clusters.

Let $\omega \subseteq E(G)$ be the percolation configuration at $p_c$, and assume it has a unique infinite cluster $\mathscr{C}_\infty$. For any $\epsilon > 0$, we define a new invariant percolation $\xi_\epsilon$ on $E(G)$, thicker than $\omega$ in some sense and sparser in another. For each $x \in V(G)$, there is a random finite set $\{x_i^*\}$ of vertices that are the closest points of $\mathscr{C}_\infty$ to $x$. Choose one of them uniformly at random, denoted by $x^*$. Let $\gamma_\epsilon$ be an independent $\mathsf{Ber}(\epsilon)$ bond percolation. Then let the edge $(x, y) \in E(G)$ be open in $\xi_\epsilon$ if $\mathrm{dist}(x, \mathscr{C}_\infty) < 1/\epsilon$, $\mathrm{dist}(y, \mathscr{C}_\infty) < 1/\epsilon$, and $x^*$ and $y^*$ are connected in $\omega \setminus \gamma_\epsilon$. It is easy to see that $\lim_{\epsilon \to 0} \mathbf{P}[\, (x, y) \in \xi_\epsilon \,] = 1$. Hence, by Exercise 12.31, for some small enough $\epsilon > 0$, there is an infinite cluster in $\xi_\epsilon$ with positive probability. However, it is also easy to check that such an infinite cluster implies the existence of an infinite cluster in $\omega \setminus \gamma_\epsilon$. But this is just $\mathsf{Ber}(p_c - \epsilon)$ percolation, so it has no infinite clusters a.s. — a contradiction.

Assume now that there are infinitely many infinite clusters in $\omega$. Using insertion tolerance, they can be glued to get infinitely many trifurcation points as in Exercise 12.29. Moreover, the same application of MTP shows that, almost surely, if a trifurcation point is removed, then each of the resulting infinite clusters still has infinitely many trifurcation points. Let $\mathcal{V} \subset V(G)$ denote the set of trifurcation points; see Figure 12.6 (a). There is an obvious graph structure $\mathcal{G}$ on $\mathcal{V}$ as vertices, with an edge between two trifurcation points if there is a path connecting them in $\omega$ that does not go through any other trifurcation point; see Figure 12.6 (b). This graph, although not at all a subgraph of $G$, represents the structure of the infinite clusters of $\omega$ well, which is a critical percolation structure; on the other hand, each component of $\mathcal{G}$ is kind of tree-like, with a lot of branching. These two properties appear to point in different directions, so we can hope to derive a contradiction. For the actual proof, we will exhibit a nonamenable spanning forest $\mathscr{F}$ inside each component of $\mathcal{G}$ in an invariant way.

For each $v \in \mathcal{V}$, let $\{\mathscr{C}_i(v) : i = 1, \ldots, j_v\}$ be the set of infinite clusters of $G \setminus \{v\}$ neighbouring $v$, where $j_v \geq 3$. For each $1 \leq i \leq j_v$, let $\{w_i^\ell : 1 \leq \ell \leq k_i\}$ be the set of $\mathcal{G}$-neighbours of $v$ in $\mathscr{C}_i(v)$. We now use some extra randomness: assign an i.i.d. $\mathsf{Unif}[0, 1]$ label to each $v \in \mathcal{V}$, and for each $v$ and $1 \leq i \leq j_v$, draw an edge from $v$ to that element of $\{w_i^\ell : 1 \leq \ell \leq k_i\}$ that has the smallest label. See Figure 12.6 (c). We then forget the orientations of the edges to get $\mathscr{F}$. It is not very hard to show that $\mathscr{F}$ has no cycles; the Reader is invited to write a proof or look it up in [BLPS99b].

Now again let $\gamma_\epsilon$ be an independent $\mathsf{Ber}(\epsilon)$ bond percolation. Then $\omega \setminus \gamma_\epsilon$ has only finite clusters. Let $\mathscr{F}_\epsilon$ be the following subgraph of $\mathscr{F}$: if $(x, y) \in E(\mathscr{F})$, then $(x, y) \in \mathscr{F}_\epsilon$ if $x$ and $y$ are in the same cluster of $\omega \setminus \gamma_\epsilon$. Clearly, $\mathscr{F}_\epsilon$ is an invariant bond percolation process on $V(G)$, with edges usually not in $E(G)$. It has only finite clusters, while $\lim_{\epsilon \to \infty} \mathbf{P}[\, (x, y) \in \mathscr{F} \setminus \mathscr{F}_\epsilon \,] = 0$. But each tree of $\mathscr{F}$ is a non-amenable tree with minimum degree at least 3, suggesting that this cannot happen. We cannot just use Exercise 12.31,

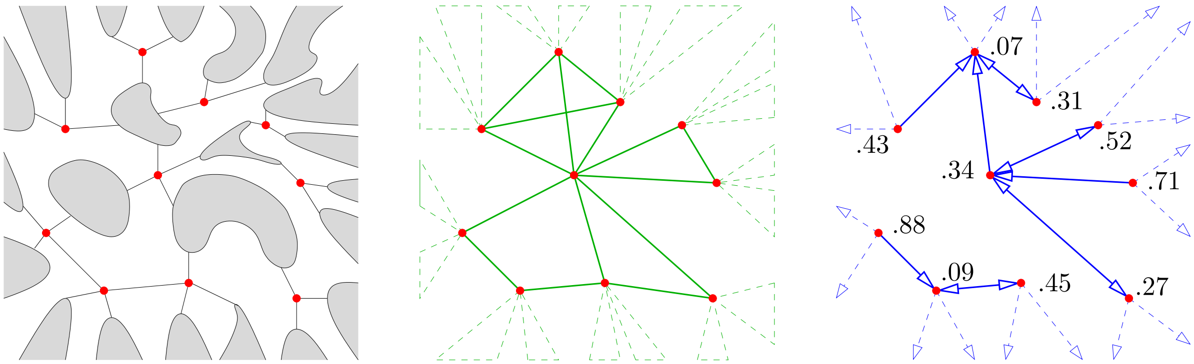



Figure 12.6: Constructing the graph $\mathcal{G}$ and forest $\mathscr{F}$ of trifurcation points in an infinite cluster. {f.trifurtree}

since $\mathscr{F}$ itself is not a transitive unimodular graph, but a similar MTP argument on the entire $V(G)$ can be set up to find the contradiction. We omit the details. □

The above proof breaks down for non-unimodular transitive graphs. Nevertheless, Conjecture 12.10 has been established for most such graphs in the union of [Tim06b] and [PerPS06].

We have seen around the Burton-Keane Theorem 12.6 that the number of infinite clusters in Bernoulli percolation is a relevant question. For the non-amenable transitive case, where having infinitely many infinite clusters is a possibility, one can define a second critical point, $p_u(G) := \inf\{p : \mathbf{P}_p[\exists! \infty \text{ cluster}] > 0\}$. The following theorem implies that for all $p \in (p_u, 1]$ there is uniqueness a.s., and for all $p \in (p_c, p_u)$ there is non-uniqueness. Note that there is no monotonicity that would make this obvious: as we raise $p$, on the one hand, infinite clusters can merge, reducing their number, but on the other hand, new infinite clusters could also appear by finite clusters merging. The following result says that the second case does not occur: {t.relentless}

**Theorem 12.13** (Above $p_c$, new infinite clusters do not appear [HäPS99])**.** *Consider the standard monotone coupling of* $\mathsf{Ber}(p)$ *percolations* $\{\omega_p : p \in [0,1]\}$ *using* $\mathsf{Unif}[0,1]$ *labels, on any infinite transitive graph. Then, a.s. in this standard coupling, if $p$ is such that $\omega_p$ has an infinite cluster a.s., and $p < p'$, then each infinite cluster of $\omega_{p'}$ contains an infinite cluster of $\omega_p$.*

The original proof, which we do not discuss here, uses Invasion Percolation; see Section 13.4. There is also a proof using the indistinguishability of infinite clusters; see Theorem 12.18 below and the paragraphs afterwards.

Before turning to the study of $p_u$, we give a nice application of the theorem in a different direction. This proof was shown to me by Gady Kozma, but it is also very similar to [LyPer16, Exercise 7.33]. {c.thetacont}

**Corollary 12.14.** *On any transitive graph, $\theta(p)$ is right-continuous at every $p \in [0,1]$ and continuous at every point except possibly at $p_c$.*

*Proof.* Right-continuity follows immediately from $\theta(p)$ being a decreasing limit (as $n \to \infty$) of the monotone increasing continuous functions $p \mapsto \mathbf{P}_p[\, o \longleftrightarrow \partial B_n(o) \,]$. Alternatively, here is a more probabilistic argument. In the standard coupling, let $C_p$ be the event that $\mathscr{C}_o$ is infinite in $\omega_p$; it is enough to show that $C_p = \bigcap_{q: p<q} C_q$. The direction $\subseteq$ is obvious, while $\supseteq$ follows from noticing that if $|\mathscr{C}_o| < \infty$ in $\omega_p$, then $\mathscr{C}_o$ has a finite boundary, hence we can raise $p$ a bit without changing $\mathscr{C}_o$ at all.

To prove left-continuity at every $p > p_c$, take $q \in (p_c, p)$, and let $R_q^p$ be the distance of $o$ from the closest infinite cluster of $\omega_q$, with distance measured within $\omega_p$. Theorem 12.13 ensures that $R_q^p$ is finite if and only if $o$ is in an infinite cluster of $\omega_p$. In other words,

$$\theta(p) - \theta(q) = \mathbf{P}[\, 0 < R_q^p < \infty \,]\,. \tag{12.10}$$ {e.thetaR}

Furthermore, $R_q^p$ is clearly a decreasing function of $q$.

Now, given any $\epsilon > 0$, if $r = r(q, \epsilon)$ is large enough, then $\mathbf{P}[\, r < R_q^p < \infty \,] < \epsilon$. Given this $r$, if $\delta_r > 0$ is small enough, then $\mathbf{P}\big[\, B_r(o) \cap \omega_p \setminus \omega_{p-\delta_r} = \emptyset \,\big] > 1 - \epsilon$. For $\delta := \min\{\delta_r, p - q\}$, using the monotonicity of $R_q^p$, we have

$$\begin{aligned}
\mathbf{P}\big[\, 0 < R_{p-\delta}^p < \infty \,\big] &= \mathbf{P}\big[\, 0 < R_{p-\delta}^p \le r \,\big] + \mathbf{P}\big[\, r < R_{p-\delta}^p < \infty \,\big] \\
&\le \mathbf{P}\big[\, \omega_p \setminus \omega_{p-\delta} \cap B_r(o) \ne \emptyset \,\big] + \mathbf{P}[\, r < R_q^p < \infty \,] \\
&\le \epsilon + \epsilon\,.
\end{aligned}$$

By (12.10), this shows left-continuity.

Let us note that we did not use the full power of Theorem 12.13 in this proof: instead of having it almost surely simultaneously for all pairs $p < p'$, it is enough if for any pair $p < p'$ we have it almost surely, since we used it only for $q$, $p - \delta$, and $p$. $\square$

The immediate questions regarding $p_u$ are whether the intervals $(p_c, p_u)$ and $(p_u, 1)$ are always non-empty when there are no simple obstacles forbidding it, and whether there are infinitely many or only one infinite cluster at $p_u$ itself.

{c.pcpu}

**Conjecture 12.15** (**The non-uniqueness phase** [BenS96c]). *For transitive graphs, $p_c < p_u$ iff $G$ is non-amenable.*

{c.pu}

**Conjecture 12.16** (**The uniqueness phase** [BenS96c]). *For transitive non-amenable graphs with one end, $p_u < 1$.*

▷ **Exercise 12.35.** *Show that in a transitive graph with infinitely many ends, $p_u = 1$.*

Before discussing what is known about the non-triviality of these intervals, here is an important characterization of uniqueness:

{t.uniconn}

**Theorem 12.17** ([LySch99]). *If $\omega$ is an ergodic insertion-tolerant invariant percolation on a unimodular transitive graph $G$ satisfying $\inf_{x,y \in V(G)} \mathbf{P}\big[\, x \overset{\omega}{\longleftrightarrow} y \,\big] > 0$, then $\omega$ has a unique infinite cluster.*

*Conversely, if an ergodic invariant $\omega$ satisfying the FKG inequality has a unique infinite cluster, then $\inf_{x,y \in V(G)} \mathbf{P}\big[\, x \overset{\omega}{\longleftrightarrow} y \,\big] > 0$.*

The second part is obvious: if $\mathscr{C}_\infty$ is the unique infinite cluster of $\omega$, then $\{x \in \mathscr{C}_\infty\}$ is an increasing event with a positive probability $p > 0$ that is independent of $x$, hence $\mathbf{P}[\, x \overset{\omega}{\longleftrightarrow} y \,] \ge \mathbf{P}[\, x, y \in \mathscr{C}_\infty \,] \ge p^2$.

▷ **Exercise 12.36.** *Give an example of a* $\mathsf{Ber}(p)$ *percolation on a Cayley graph $G$ that has non-uniqueness, but there is a sequence $x_n \in V(G)$ with* $\mathrm{dist}(x_0, x_n) \to \infty$ *and* $\inf_n \mathbf{P}_p[\, x_0 \longleftrightarrow x_n \,] > 0$.

▷ **Exercise 12.37.*** *Give an example of an ergodic uniformly insertion tolerant invariant percolation on $\mathbb{Z}^2$ with a unique infinite cluster but* $\inf_{x,y \in \mathbb{Z}^2} \mathbf{P}\big[\, x \overset{\omega}{\longleftrightarrow} y \,\big] = 0$. *(Hint: you can use the ideas of [HäM09].)*

▷ **Exercise 12.38.** *Using Theorem 12.17, prove that $p_u(\mathbb{T}_d \times \mathbb{Z}) \le 1/2$ for any $d \ge 2$.*

The proof of the first part of Theorem 12.17 in [LySch99] uses a fundamental result from the same paper:

{t.indist}

**Theorem 12.18** (Cluster indistinguishability [LySch99]). *If $\omega$ is an ergodic insertion-tolerant invariant percolation on a unimodular transitive graph $G$, and $\mathcal{A}$ is a Borel-measurable translation-invariant set of subgraphs of $G$, then either all infinite clusters of $\omega$ are in $\mathcal{A}$ a.s., or none.*

A rough intuitive explanation of how Theorem 12.18 implies Theorem 12.17 is the following. First of all, we want to define a "cluster density" for each infinite cluster $\mathscr{C}$. In an amenable transitive graph, averaging along Følner sets would be fine, but we cannot do that in general; instead, let us run a simple random walk from an arbitrary fixed starting point $o$, and look at the relative frequency of visits to $\mathscr{C}$ (as a limit over the first $n$ steps). Using a subadditive limit argument, one can show that this limit exists for each $\mathscr{C}$ and is an almost sure constant, denoted by $\alpha_o(\mathscr{C})$. Now, if we start the walk from a different vertex $o'$, then with positive probability we get to $o$ in a finite number of steps, and then the limit frequency will forget this starting segment, hence $\alpha_{o'}(\mathscr{C}) = \alpha_o(\mathscr{C})$ with positive probability. Since these are almost sure constants, we get that $\alpha_o(\mathscr{C})$ does not depend on $o$. Thus, it is an invariant quantity, and by cluster indistinguishability, it must be the same $\alpha \in [0,1]$ for each cluster. However, if $\inf_{x,y} \mathbf{P}\big[\, x \stackrel{\omega}{\longleftrightarrow} y \,\big] = \rho > 0$, then the random walk frequency $\alpha$ is clearly at least $\rho$ (obvious in expectation, and frequency is an a.s. constant). Since the sum of these frequencies is at most 1, there must be only finitely many infinite clusters. Then, by insertion tolerance and ergodicity, there must be a unique one.

As promised earlier, Theorem 12.18 also easily implies Theorem 12.13 (above $p_c$, new infinite clusters do not appear), at least in its weak form that for any pair $p < p'$ it holds almost surely, as pointed out in [LySch99]. Indeed, note that, for $p < p'$ and conditioned on $\omega_{p'}$, the configuration $\omega_p$ is just percolation on $\omega_{p'}$ with density $p/p'$. Assuming $\theta(p) > 0$, there must be an infinite cluster of $\omega_{p'}$ with $\theta(p/p') > 0$. But then, by indistinguishability, every infinite $\omega_{p'}$-cluster must have $\theta(p/p') > 0$, and hence, by Kolmogorov's 0-1 law, every $\omega_{p'}$-cluster must have an infinite $\omega_p$-cluster almost surely.

Regarding $p_c < p_u$, it was shown by Pak and Smirnova-Nagniebeda [PaSN00] that every non-amenable group has a generating set satisfying this. Of course, this property should be independent of the generating set taken (probably also a quasi-isometry invariant). Here is a brief explanation of what kind of Cayley graphs make $p_c < p_u$ easier. As usually, we will consider bond percolation.

In a transitive graph $G$, let $a_n$ be the number of simple loops of length $n$ starting (and ending) at a given vertex $o$, and let $\gamma(G) := \limsup_n a_n^{1/n}$. The smaller this is, the more treelike the graph is, hence there is hope that $p_u$ will be larger. Indeed, as proved by Schramm, see [LyPer16, Theorem 7.30],

$$\frac{1}{\gamma(G)} \le p_u(G)\,, \tag{12.11}$$ {e.gammapu}

for any transitive graph $G$. The proof is a nice counting argument, which we outline briefly. Taking $p < p_+ < 1/\gamma$, it is enough to prove by the easy direction of Theorem 12.17 that, for $o$ and $x$ far away from each other, $\mathbf{P}_p[\, o \longleftrightarrow x \,]$ must be small. If $o \longleftrightarrow x$ at level $p$ (in the standard coupling), then either there are many cut-edges between them already at level $p_+$, or there are many $p_+$-open simple loops. In the first scenario, keeping many (say, $k$) cut-edges open even at level $p$ is exponentially costly: the probability is $(p/p_+)^k$, conditionally on the connection at $p_+$. But the second scenario is unlikely because there are not enough simple loops in $G$: if $u_k(r)$ denotes the probability that there is some $x \notin B_r(o)$ such that $o$ and $x$ are $p_+$-connected with at most $k$ cut-edges, then $u_0(r) \to 0$ as $r \to \infty$ because of $p_+ < 1/\gamma$, and the same can be shown for each $u_k(r)$ by induction on $k$, by noticing that $u_{k+1}(r) \le u_0(s) + |B_s(o)|\, u_k(r-s)$. Then, by choosing $k$ then $r$ large enough, both contributions to $\mathbf{P}_p[\, o \longleftrightarrow x \,]$ will be small, proving (12.11).

Now, if $G$ is $d$-regular, then $a_n/d^n \leq p_n(o,o)$ for SRW on $G$, hence

$$\frac{1}{d_G\,\rho(G)} \leq \frac{1}{\gamma(G)}\,, \tag{12.12}$$ {e.gammarho}

where $\rho$ is the spectral radius of the SRW. On the other hand, by Proposition 12.9,

$$p_c(G) \leq \frac{1}{h_E(G)+1} < \frac{1}{h_E(G)} = \frac{1}{d_G\,\iota_E(G)}\,, \tag{12.13}$$ {e.pcrho}

where $h_E$ is the edge Cheeger constant defined using the ratios $|\partial_E S|/|S|$, while $\iota_E$ is the edge Cheeger constant of the Markov chain (the SRW), i.e., it uses the ratios $C(\partial_E S)/\pi(S)$, as in Section 7.2. After comparing the three displayed inequalities (12.11, 12.12, 12.13), the aim becomes to find a Cayley graph $G$ for which $\iota_E$ is close to 1 and $\rho$ is close to 0. Fortunately, these two aims are really the same, by the quantitive bound $\iota_E^2/2 \leq 1-\rho \leq \iota_E$ from Kesten's Theorem 7.3. And making $\rho \to 0$ is easy: take any finite generating set $S$, for which we have some $\rho(G(\Gamma,S)) = \rho_0 < 1$, then take $G(\Gamma,S^k)$ for some large $k$, where $S^k = S\cdots S$ is the multiset of all possible $k$-wise products (i.e., we keep the multiplicities with which group elements occur as $k$-wise products). The transition matrix for SRW on the resulting multigraph is just the $k^{\text{th}}$ power of the original transition matrix, hence $\rho(G(\Gamma,S^k)) = \rho_0^k \to 0$ as $k \to \infty$, and we are done.

It is not known if this $\rho \to 0$ can be achieved with generating sets without multiplicities. For instance, would the ball $B_k^S(1)$ work, given by any finite generating set $S$? The answer is "yes" for groups having a free subgroup $F_2$, or having the so-called **Rapid Decay property**, which includes all Gromov-hyperbolic groups. See [PaSN00] for the (easy) proofs of these statements.

The following exercise shows that if we do not insist on adding edges in a group-invariant way (i.e., by increasing the generating set), but still want to add them only "locally", then we indeed can push $h_E$ close to the degree (or in other words, can push $\iota_E$ close to 1) without using multiple edges. Even the outer vertex Cheeger constant $h_V := \inf |\partial_V^{\text{out}} S|/|S|$ can be close to the degree, which is of course stronger, since $|\partial_V^{\text{out}} S| \leq |\partial_E S| \leq (d-1)|S|$ in a $d$-regular graph.

▷ **Exercise 12.39.** *Show that for any $d$-regular non-amenable graph $G$ and any $\epsilon > 0$, there exists $K < \infty$ such that we can add edges connecting vertices at distance at most $K$, such that the new graph $G^*$ will be $d^*$-regular, no multiple edges, and $\iota_V(G^*) := h_V(G^*)/d^*$ will be larger than $1-\epsilon$ for the outer vertex Cheeger constant. (Hint: use the wobbling paradoxical decomposition from Exercise 5.9. The Mass Transport Principle shows that this proof cannot work in a group-invariant way.)*

▷ **Exercise 12.40.** ***

**(a)** *Is it true that $\iota_E(G(\Gamma,B_k^S))/|B_k^S| \to 1$ as $k \to \infty$ for any nonamenable group $\Gamma$ and the ball of radius $k$ in any finite generating set $S$?*

**(b)** *Is it true that $\iota_V(G(\Gamma,B_k^S))/|B_k^S| \not\to 1$ for any group $\Gamma$ and any finite generating set $S$?*

Possibly the best attempt so far at proving $p_c < p_u$ in general is an unpublished argument of Oded Schramm, which we discuss in Section 12.5 below, see Theorem 12.36.

There is an interpretation of $p_c < p_u$ in terms of the Free and Wired Minimal Spanning Forests, see Theorem 13.13 below.

Conjecture 12.16 on $p_u < 1$ is known under some additional assumptions (besides being nonamenable and having one end): $\mathsf{CutCon}(G) < \infty$ (for instance, being a finitely presented Cayley graph) or being a Kazhdan group are sufficient, see [BabB99, Tim07] or [LyPer16, Section 7.6], and [LySch99], respectively. Here is how being Kazhdan plays a role:

{t.Kazhdanclos

**Theorem 12.19** ([GlW97]). *A f.g. infinite group $\Gamma$ is Kazhdan iff the measure $\mu_{\text{half}}$ on $2^\Gamma$ that gives probability half to the emptyset and probability half to all of $\Gamma$ is not in the weak* closure of the $\Gamma$-invariant ergodic probability measures on $2^\Gamma$.*

*In other words, for a transitive d-regular graph $G(V,E)$ and any $o \in V$, let*

$$\delta^{\text{erg}}(G) := \sup \left\{ \frac{\mathbf{E}_\mu \big| \{(o,x) \in E : \sigma(x) = \sigma(o)\} \big|}{d_G} : \begin{array}{l} \textit{ergodic invariant measures } \mu \textit{ on } \sigma \in \{\pm 1\}^V \\ \textit{with } \mathbf{E}_\mu \sigma(o) = 0 \end{array} \right\}.$$

*Then, a f.g. group $\Gamma$ is Kazhdan iff any (or one) of its Cayley graphs $G$ has $\delta^{\text{erg}}(G) < 1$.*

There are some natural variants of $\delta^{\text{erg}}(G)$: instead of all ergodic measures, we can take only the tail-trivial ones (i.e., all tail events have probability 0 or 1), or only those that are factors of an i.i.d. $\mathsf{Unif}[0,1]$ process on $V$ or $E$ (i.e., we get $\sigma$ as a measurable function $f$ of the i.i.d. process $\omega$, where $f$ commutes with the action of $\Gamma$ on the configurations $\sigma$ and $\omega$). The corresponding suprema are denoted by $\delta^{\text{tt}}$ and $\delta^{\text{fiid}}$. Clearly, $\delta^{\text{erg}}(G) \geq \delta^{\text{tt}}(G) \geq \delta^{\text{fiid}}(G)$ **(wait, not that clearly...)**, but are they really different? An unpublished result of Benjy Weiss and Russ Lyons is that, for Cayley graphs, $\delta^{\text{fiid}}(G) < 1$ is equivalent to nonamenability. (One direction is easy: see Exercise 12.41 (a) below.) On the other hand, it is only conjectured that $\delta^{\text{tt}}(G) < 1$ is again equivalent to nonamenability. In general, it is a hard task to find out what tail trivial processes are factors of some i.i.d. process. See [LyNaz11] and Section 14.2 for a bit more on these issues.

{ex.agree}

▷ **Exercise 12.41.**

**(a)** *Show that $\delta^{\text{fiid}}(G) = 1$ for any amenable transitive graph $G$. (Hint: have i.i.d. coin flips in large Følner neighbourhoods vote on the $\sigma$-value of each vertex.)*

**(b)** *Show that $\delta^{\text{erg}}(\mathbb{T}_3) = 1$. (Hint: free groups are not Kazhdan e.g. because they surject onto $\mathbb{Z}$.)*

▷ **Exercise 12.42.** ***

**(a)** *Find the value of $\delta^{\text{fiid}}(\mathbb{T}_3)$.*

**(b)** *Show that $\delta^{\text{tt}}(\mathbb{T}_3) < 1$.*

We come back now to the question of $p_u < 1$:

{c.Kazhdanpu}

**Corollary 12.20** ([LySch99]). *If $G$ is a Cayley graph of an infinite Kazhdan group $\Gamma$, then $p_u(G) < 1$. Moreover, at $p_u$ there is non-uniqueness.*

*Sketch of proof of $p_u < 1$.* Assume $p_u(G) = 1$, and let $\omega_p$ be $\mathsf{Ber}(p)$ percolation at some $p < 1$. Let $\eta_p$ be the invariant site percolation where the vertex set of each cluster of $\omega_p$ is completely deleted with probability 1/2, independently of other clusters. By Theorem 12.17, we have $\inf_{x,y} \mathbf{P}_p[\, x \longleftrightarrow y \,] = 0$. It is not surprising that this implies that $\eta_p$ is ergodic. On the other hand, as $p \to 1$, it is clear that $\eta_p$ converges to $\mu_{\mathsf{half}}$ in the weak* topology, so Theorem 12.19 says that $\Gamma$ could not be Kazhdan. □

Let us also sketch a non-probabilistic proof that uses directly Definition 7.12 of being Kazhdan, instead of the characterization Theorem 12.19. It is due to [IKT09], partly following a suggestion made in [LySch99].

*Sketch of another proof of $p_u < 1$.* Let $\Omega = \{0,1\}^\Gamma$, with the product Bernoulli measure $\mu_p$ with density $p$. Given the right Cayley graph $G(\Gamma, S)$ for a finite generating set $S$, and a site percolation configuration $\omega \in \Omega$, let $\mathscr{C}_\omega$ be the set of its clusters, $\mathscr{C}_\omega(g)$ be the cluster of $g$, and $\ell^2(\mathscr{C}_\omega)$ be the Hilbert space of square summable real-valued sequences defined on $\mathscr{C}_\omega$. Consider now the direct integral of these Hilbert spaces over all $\omega$,

$$\mathcal{H} := \int_\Omega \ell^2(\mathscr{C}_\omega)\, d\mu_p(\omega)\,, \qquad (\varphi,\psi)_{\mu_p} := \int_\Omega \sum_{C\in\mathscr{C}_\omega} \varphi(\omega,C)\psi(\omega,C)\, d\mu_p(\omega)\,, \text{ for } \varphi,\psi\in\mathcal{H}\,.$$

$\Gamma$ has a natural unitary (in fact, orthogonal) representation $\rho$ on $\mathcal{H}$, translating vectors by $\rho_g(\varphi)(\omega,C) := \varphi(\omega^g, g^{-1}C)$ for $C\in\mathscr{C}_\omega$, where $\omega^g(h) := \omega(gh)$; note here that $g^{-1}C\in\mathscr{C}_{\omega^g}$, since $g^{-1}x \overset{\omega^g}{\longleftrightarrow} g^{-1}y$ iff $x \overset{\omega}{\longleftrightarrow} y$. Now, if $\iota\in\mathcal{H}$ is the vector that is 1 at $\mathscr{C}_\omega(1)$ and 0 at the other clusters of $\omega$, for each $\omega$, then $\|\iota\|_{\mu_p}=1$, and

$$\mathbf{P}_p[\, g\longleftrightarrow h\,] = \mathbf{P}_p[\,\mathscr{C}(g)=\mathscr{C}(h)\,] = (\rho_g(\iota),\rho_h(\iota))_{\mu_p} = (\iota,\rho_{gh^{-1}}(\iota))_{\mu_p}\,.$$

This implies, by the way, that $\mathbf{P}_p[\, g\longleftrightarrow h\,]$ is a positive definite function.

Take an $\epsilon>0$ smaller than the Kazhdan constant $\kappa(\Gamma,S)>0$. If $p$ is close enough to 1, then $\|\iota-\rho_s(\iota)\|_{\mu_p} = 2-2(\iota,\rho_s(\iota)) = 2\mathbf{P}_p[\,1\not\longleftrightarrow s\,]<\epsilon$ for all $s\in S$. Hence, by the Kazhdan property, there is an invariant vector $\xi\in\mathcal{H}$.

It is easy to see that a vector $\xi$ is invariant under the representation $\rho$ iff $\xi(\omega,C)=\xi(\omega^g,g^{-1}C)$ for all $g\in\Gamma$. On the other hand, for each $\omega$, we have $\xi(\omega,\cdot)\in\ell^2(\mathscr{C}_\omega)$, hence its maximum is attained at finitely many clusters. Taking these clusters, we get an invariant choice of finitely many clusters. By a simple application of the Mass Transport Principle, all of these clusters must be infinite. If there are infinitely many infinite clusters in $\mu_p$, then, by the cluster indistinguishability Theorem 12.18, there is no invariant way to choose finitely many of them, hence we must have a unique infinite cluster instead. □

▷ **Exercise 12.43.** *Fill in the missing details in either of the above proof sketches for $p_u<1$ for Kazhdan groups.*

▷ **Exercise 12.44** (Todor Tsankov)**.** *** *Prove $p_c<p_u$ for Kazhdan groups by finding an appropriate representation.*

## 12.3 Percolation on finite graphs. Threshold phenomema

{ss.percfin}

The best-known example is the Erdős-Rényi random graph model $G(n,p)$, which is just $\mathsf{Ber}(p)$ percolation on the complete graph $K_n$. We will give a very brief introduction; see [JaŁR00] for more on random graphs, [AloS00] for probabilistic combinatorics in general, and [KalS06] for a nice survey of influences and threshold phenomena that we will define in a second. [Gri10] contains a bit of everything, similarly to the present notes.

A **graph property** $\mathcal{A}$ over some vertex set $V$ is a subset of $\{0,1\}^{V\times V}$ that is invariant under the diagonal action of the permutation group $\mathrm{Sym}(V)$ (by diagonal action, we mean that $\omega(v,w)^\pi := \omega(\pi(v),\pi(w))$ for any $\pi\in\mathrm{Sym}(V)$), or in other words, that does not care about the labelling of the vertices. Examples are "containing a triangle", "being connected", "being 3-colourable", and so on. Such properties are most often monotone increasing or decreasing (as defined just before Theorem 12.3). It was noticed by Erdős and Rényi [ErdR60] that, in the $G(n,p)$ model, monotone graph properties have a relatively sharp threshold: there is a short interval of $p$ values in which they become extremely likely or unlikely. (Recall from the first page of Section 12.1 that the standard coupling implies that, for monotone properties $\mathcal{A}$, the function $p\mapsto\mathbf{P}_p[\,\mathcal{A}\,]$ is monotone.) Here is a simple example:

Let $X$ be the number of triangles contained in $G(n,p)$ as a subgraph. Clearly, $\mathbf{E}_pX=\binom{n}{3}p^3$, hence, if $p=p(n)=o(1/n)$, then $\mathbf{P}_p[\,X\geq 1\,]\leq\mathbf{E}_pX\to 0$. What can we say if $p(n)n\to\infty$? We have $\mathbf{E}_pX\to\infty$, but, in order to conclude that $\mathbf{P}_p[\, X \geq 1\, ] \to 1$, we also need that $X$ is somewhat concentrated. This is the easiest to do via the **Second Moment Method**: for any random variable $X \geq 0$, applying Cauchy-Schwarz to $\mathbf{E}[\, X\, ] = \mathbf{E}[\, \mathbf{1}_{X>0}\, X\, ]$ gives

$$\mathbf{P}[\, X > 0\, ] \geq \frac{(\mathbf{E}X)^2}{\mathbf{E}[\, X^2\, ]}\,. \tag{12.14}$$ {e.2MM}

More generally, for $t \in (0,1)$, we have $\mathbf{E}[\, X\, ](1-t) \leq \mathbf{E}[\, \mathbf{1}_{X>t\mathbf{E}X}\, X\, ]$, and Cauchy-Schwarz gives

$$\mathbf{P}[\, X > t\, \mathbf{E}X\, ] \geq (1-t)^2 \frac{(\mathbf{E}X)^2}{\mathbf{E}[\, X^2\, ]}\,. \tag{12.15}$$ {e.2MM2}

This is called sometimes the **Paley-Zygmund inequality**. Note that it is better than Chebyshev's inequality in that it gives a non-empty result even when $\mathrm{Var}X > (\mathbf{E}X)^2$.

Now, back to the number of triangles, if $I_\Delta$ is the indicator variable for the event that the triangle $\Delta \subset E(K_n)$ is open, then $I_\Delta$ and $I_{\Delta'}$ are independent if $\Delta$ and $\Delta'$ do not share an edge, and

$$\begin{aligned}
\mathrm{Var}_p[X] = \sum_{\substack{\Delta,\Delta' \subset E(K_n) \\ \text{triangles}}} \mathrm{Cov}_p[I_\Delta, I_{\Delta'}] &= \sum_{\Delta} \mathrm{Var}_p[I_\Delta] + \sum_{|\Delta \cap \Delta'|=1} \mathrm{Cov}_p[I_\Delta, I_{\Delta'}] \\
&= \binom{n}{3} p^3(1-p^3) + 2\binom{n}{2}\binom{n-2}{2}(p^5 - p^6) \\
&\sim \frac{n^3}{6}p^3 + \frac{n^4}{2}p^5\,.
\end{aligned}$$

Thus, for $p \sim \lambda/n$ with $\lambda \in (0,\infty)$, we have $\mathbf{E}_p[\, X^2\, ] \leq C(\lambda)(\mathbf{E}_p X)^2$. Moreover, $\lim_{\lambda\to\infty} C(\lambda) = 1$. Therefore, the first moment estimate and (12.14) yield that

$$0 < a(\lambda) \leq \mathbf{P}_p[\, X \geq 1\, ] \leq b(\lambda), \qquad \text{with } \lim_{\lambda\to 0} b(\lambda) = 0 \text{ and } \lim_{\lambda\to\infty} a(\lambda) = 1\,. \tag{12.16}$$ {e.triangles}

The following exercise gives an example of subgraph containment where the Second Moment Method *fails*. See [AloS00, Sections 4.4 and 10.1] for a thorough discussion of this phenomenon and what can one do about it.

▷ **Exercise 12.45.** *Let $H$ be the following graph with 5 vertices and 7 edges: a complete graph $K_4$ with an extra edge from one of the four vertices to a fifth vertex. Show that if $5/7 > \alpha > 4/6$, and $p = n^{-\alpha}$, then the expected number of copies of $H$ in $G(n,p)$ goes to infinity, but nevertheless the probability that there is at least one copy goes to 0.*

Here are now the exact general definitions for threshold functions (if it looks complicated, it is only because it covers all cases in one formula):

{d.threshold}

**Definition 12.21.** *Consider the* $\mathsf{Ber}(p)$ *product measure on the base sets $[n] = \{1,\dots,n\}$, and let $\mathcal{A}_n \subseteq \{0,1\}^{[n]}$ be a sequence of increasing events (not the empty and not the full). For $t \in [0,1]$, let $p_t^{\mathcal{A}}(n)$ be the $p$ for which $\mathbf{P}_p[\, \mathcal{A}_n\, ] = t$, and call $p^{\mathcal{A}}(n) := p^{\mathcal{A}}_{1/2}(n)$ the* **critical probability** *for $\mathcal{A}$. (These exist since $p \mapsto \mathbf{P}_p[\, \mathcal{A}_n\, ]$ is strictly increasing and continuous, equalling 0 and 1 at $p=0$ and $p=1$, respectively.) The sequence $\mathcal{A} = \mathcal{A}_n$ is said to have a* **threshold** *if*

$$\mathbf{P}_{p(n)}[\, \mathcal{A}_n\, ] \to \begin{cases} 1 & \text{if } \frac{p(n)}{p^{\mathcal{A}}(n)} \vee \frac{1-p^{\mathcal{A}}(n)}{1-p(n)} \to \infty\,, \\ 0 & \text{if } \frac{p(n)}{p^{\mathcal{A}}(n)} \wedge \frac{1-p^{\mathcal{A}}(n)}{1-p(n)} \to 0\,. \end{cases}$$

*Furthermore, the threshold is* **sharp** *if for any $\epsilon > 0$, we have*

$$\frac{\left|p^{\mathcal{A}}_{1-\epsilon}(n) - p^{\mathcal{A}}_{\epsilon}(n)\right|}{p^{\mathcal{A}}(n) \wedge (1 - p^{\mathcal{A}}(n))} \to 0 \text{ as } n \to \infty\,,$$

*and it is* **coarse** *if there are $\epsilon, c > 0$ such that the above ratio is larger than $c$ for all $n$.*

*Similar definitions can be made for decreasing events.*

A sequence $\mathcal{A}_n$ of events will often be defined only for some subsequence $n_k \to \infty$: for instance, the base set $[n_k]$ may stand for the vertex or the edge set of some graph $G_k(V_k, E_k)$.

The following basic result is due to Bollobás and Thomason [BolT87] (with a different proof, using isoperimetric considerations — we will see soon what thresholds have to do with isoperimetry):

{ex.threshold}

▷ **Exercise 12.46.** *Prove that for any sequence monotone events $\mathcal{A} = \mathcal{A}_n$ and any $\epsilon$ there is $C_\epsilon < \infty$ such that $\left|p^{\mathcal{A}}_{1-\epsilon}(n) - p^{\mathcal{A}}_{\epsilon}(n)\right| < C_\epsilon\, p^{\mathcal{A}}_{\epsilon}(n) \wedge (1 - p^{\mathcal{A}}_{1-\epsilon}(n))$. Conclude that every sequence of monotone events has a threshold. (Hint: take many independent copies of a low density percolation to get success with good probability at a larger density.)*

An example that has a threshold but a very coarse one is dictatorship by the $i$th variable: $\mathcal{D}_i := \{\omega : i \in \omega\}$, for some fixed $i \in [n]$. For this event, $\mathbf{P}_p[\,\mathcal{D}_i\,] = p$. A less trivial example of a coarse threshold is the event of containing a triangle, by (12.16). We will describe an example of a sharp threshold in Theorem 12.23

Can we tell without exact calculations how sharp the threshold of a given sequence of events is? A short threshold interval means that $\mathbf{P}_p[\,\mathcal{A}\,]$ changes rapidly with $p$, hence the following result, called the **Margulis-Russo formula**, is fundamental in the study of threshold phenomena: for any event $\mathcal{A}$,

$$\begin{aligned}\frac{d}{dp}\mathbf{P}_p[\,\mathcal{A}\,] &= \sum_{i\in[n]} \bar{I}^{\mathcal{A}}_p(i)\,, \text{ which is}\\ &= \sum_{i\in[n]} \mathbf{P}_p[\,i \text{ is pivotal for } \mathcal{A}\,] \text{ for } \mathcal{A} \text{ increasing,}\end{aligned} \tag{12.17}$$ {e.Russo}

where $\bar{I}^{\mathcal{A}}_p(i) := \mathbf{P}_p[\,\mathcal{A} \mid i \in \omega\,] - \mathbf{P}_p[\,\mathcal{A} \mid i \notin \omega\,]$ is the **signed influence** of the variable $i$ on $\mathcal{A}$, while **pivotal** means that changing the variable in a given configuration $\omega$ changes the outcome of the event. The ordinary (unsigned) **influence** is just $I^{\mathcal{A}}_p(i) := \mathbf{P}_p[\,i \text{ is pivotal for } \mathcal{A}\,]$, equalling $\bar{I}^{\mathcal{A}}_p(i)$ for $\mathcal{A}$ increasing.

The proof of (12.17) is simple: write $\mathbf{P}_p[\,\mathcal{A}\,] = \sum_\omega \mathbf{1}_{\mathcal{A}}(\omega)\mathbf{P}_p[\,\omega\,]$, compute the derivative for each term,

$$\frac{d}{dp}\mathbf{P}_p[\,\mathcal{A}\,] = \sum_\omega \mathbf{1}_{\mathcal{A}}(\omega)\mathbf{P}_p[\,\omega\,]\left[|\omega|\frac{1}{p} - (n - |\omega|)\frac{1}{1-p}\right], \tag{12.18}$$ {e.ddpA}

then, by monitoring that a given configuration $\omega \in \mathcal{A}$ appears for what $\eta$ configurations as $\eta \cup \{i\} = \omega$, notice that

$$\sum_{i=1}^n \mathbf{P}_p[\,\mathcal{A} \mid i \in \omega\,] = \sum_\omega \mathbf{1}_{\mathcal{A}}(\omega)\mathbf{P}_p[\,\omega\,]\left[|\omega| + |\omega|\frac{1-p}{p}\right],$$

and similarly,

$$\sum_{i=1}^n \mathbf{P}_p[\,\mathcal{A} \mid i \notin \omega\,] = \sum_\omega \mathbf{1}_{\mathcal{A}}(\omega)\mathbf{P}_p[\,\omega\,]\left[(n - |\omega|) + (n - |\omega|)\frac{p}{1-p}\right],$$

and the difference of the last two equations is indeed equal to (12.18).

Here is a more intuitive proof that we explain, for simplicity, for increasing $\mathcal{A}$. In the standard coupling of the $\mathsf{Ber}(p)$ measures for $p \in [0,1]$, if we gradually raise the density from $p$ to $p+\epsilon$, then the increase in the probability of the event is exactly the probability that there is a newly opened variable that is pivotal at that moment. Probabilities are harder to compute than expectations, hence look at the expected number of these pivotal openings, which is

$$\int_p^{p+\epsilon} \mathbf{E}_q[\,\text{number of pivotals for } \mathcal{A}\,]\, dq\,.$$

For very small $\epsilon$ (depending even on $n$, the number of bits), the probability that there is more than one bit that has opened is very small. This implies that (1) the expectation in the integral is basically constant on $[p, p+\epsilon]$, hence the integral is close to $\epsilon\,\mathbf{E}_p[\,\text{number of pivotals for } \mathcal{A}\,]$, and (2) this expected number of pivotal openings is close to the probability of having any, which is exactly what we are after. Thus, dividing by $\epsilon$ and taking $\lim_{\epsilon\to 0}$ gives (12.17).

{ex.ctsRusso}

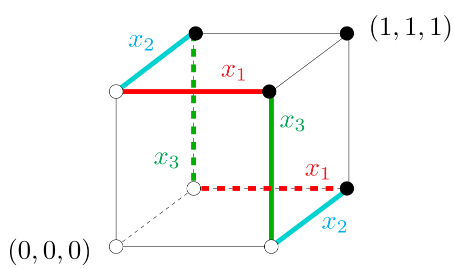
▷ **Exercise 12.47.** *Let $f : \{0,1\}^{[n]} \longrightarrow \mathbb{R}$ be any monotone function (not necessarily Boolean), and consider the derivatives $\nabla_i f(\omega) := f(\omega \cup \{i\}) - f(\omega \setminus \{i\})$. Show that*

$$\frac{d}{dp}\mathbf{E}_p[\,f\,] = \sum_{i\in[n]} \mathbf{E}_p\big[\,|\nabla_i f|\,\big]\,.$$

So, by (12.17), one can prove a sharp (or coarse) threshold for some increasing $\mathcal{A} = \mathcal{A}_n$ by showing that the **total influence** $I_p^{\mathcal{A}} := \sum_i I_p^{\mathcal{A}}(i) = \mathbf{E}_p[\,\text{number of pivotals for } \mathcal{A}\,]$ is large (or small) for all $p$ around the critical density $p^{\mathcal{A}}(n)$.

▷ **Exercise 12.48.** *In the random graph $G(n,p)$ with $p = \lambda/n$, for $\mathcal{A}_n = \{$containing a triangle$\}$, show directly that the expected number of pivotal edges is $\asymp n$ (with factors depending on $\lambda$), and hence, by Russo's formula (12.17), the threshold window is of size $p_{1-\epsilon}^{\mathcal{A}}(n) - p_{\epsilon}^{\mathcal{A}}(n) \asymp 1/n$, as we also saw from (12.16).*

Note that $I_p^{\mathcal{A}}$ is the size of the edge boundary of $\mathcal{A}$ in $\{0,1\}^{[n]}$, measured using $\mathbf{P}_p$:




$$I_p^{\mathcal{A}} = \textstyle\sum_{(x,y)\in\partial_E\mathcal{A}} \big(\mathbf{P}_p[x] + \mathbf{P}_p[y]\big)\,.$$

Figure 12.7: For majority on 3 variables, $\mathcal{A} = \mathsf{Maj}_3 := \big\{x \in \{0,1\}^3 : \sum_{i=1}^3 x_i \geq 2\big\}$, each variable is pivotal in four configurations (at the endpoints of two edges). {f.InfIsop}

Therefore, bounding the total influence from below is the same as proving isoperimetric inequalities in the hypercube. For instance, for the uniform measure $p = 1/2$, the relationship between total influence and edge boundary is $I_{1/2}^{\mathcal{A}} = |\partial_E \mathcal{A}|/2^{n-1}$, and the classical edge-isoperimetric inequality for the hypercube, Exercise 5.12, becomes

$$I_{1/2}^{\mathcal{A}} \geq 2\,\mathbf{P}_{1/2}[\,\mathcal{A}\,]\,\log_2 \frac{1}{\mathbf{P}_{1/2}[\,\mathcal{A}\,]}\,. \tag{12.19}$$ {e.InfIsop}

An easier inequality is the following Poincaré inequality (connecting isoperimetry and variance, just as in Section 8.1):

$$I_{1/2}^{\mathcal{A}} \geq 4\,\mathbf{P}_{1/2}[\,\mathcal{A}\,]\,(1 - \mathbf{P}_{1/2}[\,\mathcal{A}\,])\,. \tag{12.20}$$ {e.InfPoin}

Also, compare these inequalities with Exercise 6.13 that uses a slightly larger "boundary".

{ex.InfIsop}

▷ **Exercise 12.49.**

**(a)** *Prove the identity* $I_{1/2}^{\mathcal{A}} = |\partial_E \mathcal{A}|/2^{n-1}$.

**(b)** *Show that, among all monotone events* $\mathcal{A}$ *on* $[n]$*, the total influence* $I_{1/2}^{\mathcal{A}}$ *is maximized by the majority* $\mathsf{Maj}_n$*, and find the value. (Therefore,* $\mathsf{Maj}_n$ *has the sharpest possible threshold at* $p = 1/2$*. For general* $p$*, but still bounded away from 0 and 1, the optimum remains similar: see (12.28).)*

{ex.InfPoin}

▷ **Exercise 12.50.** *Prove the Poincaré inequality (12.20).*

*Hint: Define a map from the set of pairs* $(\omega, \omega') \in \mathcal{A} \times \mathcal{A}^c$ *into* $\partial_E \mathcal{A}$*. Alternatively, use discrete Fourier analysis, defined very briefly as follows:*

*For any function* $f : \{-1,1\}^n \longrightarrow \mathbb{R}$ *of* $n$ *bits, define the* **Fourier-Walsh coefficients** $\widehat{f}(S) := \mathbf{E}_{1/2}\big[\, f(\omega)\, \chi_S(\omega) \,\big]$*, where* $\big\{\chi_S(\omega) := \prod_{i \in S} \omega(i) \,:\, S \subseteq [n]\big\}$ *is an orthonormal basis w.r.t.* $(f,g) := \mathbf{E}_{1/2}[\, fg \,]$*. (In a slightly different language, these are the characters of the group* $\mathbb{Z}_2^n$*.) Determine the variance* $\mathrm{Var} f$*, and, for Boolean functions* $f = \mathbf{1}_{\mathcal{A}}$*, the total influence* $I_{1/2}^{\mathcal{A}}$ *in terms of the squared coefficients* $\widehat{f}(S)^2$*.*

For balanced events, i.e., when $\mathbf{P}_{1/2}[\,\mathcal{A}\,] = 1/2$, the stronger (12.19) gives $I_{1/2}(\mathcal{A}) \geq 1$, which is sharp, as shown by the dictator, $\mathcal{D}_i := \{\omega : i \in \omega\}$ as an event, or $D_i(\omega) := \omega(i)$ as a Boolean function, mentioned above as an example of a very coarse threshold.

More generally, we have seen in Exercise 5.12 that the sets with small edge boundary depend on as few coordinates as possible. What can be said for $p \neq 1/2$? What properties have sharp thresholds? There is an ultimate answer by Friedgut and Bourgain [FriB99]: an increasing property $\mathcal{A}$ has a coarse threshold at $p^{\mathcal{A}}$ iff it is *local*, i.e., it can be $\epsilon$-approximated at $p^{\mathcal{A}}$ by an increasing event $\mathcal{B}$ whose minimal elements have size bounded by some $B(\epsilon)$. Typical examples are the events of containing a triangle or some other fixed subgraph, either anywhere in the graph, or on a fixed subset of the bits as in a dictator event. Sharp thresholds correspond to *global* properties, such as connectivity, and $k$-colorability for $k \geq 3$. (Although $\mathcal{A}_k := \{\text{non-}k\text{-colorable}\}$ can be enforced by having a $(k+1)$-clique, the critical probability for the latter event is much higher, hence this does not mean locality. Could there exist any local event that has a smaller critical probability, but still has a noticeable effect on $\mathcal{A}_k$? It turns out that even conditioning on a specific $k$-clique to be present at the critical probability of $\mathcal{A}_k$ does not significantly increase the probability of $\mathcal{A}_k$, which implies that $\mathcal{A}_k$ cannot be approximated by bounded events.) The exact results are slightly different in the case of graph and hypergraph properties (Friedgut) and general events (Bourgain), and we omit them. Note that although it is easy to show locality of events that are "obviously" local, it might be much harder to prove that something is global and hence has a sharp threshold. Therefore, it is still useful to have more robust conditions for the quantification of how sharp a threshold is. The following is a key theorem, which is a generalization of the $p = 1/2$ case proved in [KahKL88], strengthening (12.20); see also [Tal94]:

{t.BKKKL}

**Theorem 12.22** ([BouKKKL92])**.** *For* $\mathsf{Ber}(p)$ *product measure on* $[n]$*, and any nontrivial event* $\mathcal{A} \subset \{0,1\}^{[n]}$*, we have*

$$I_p^{\mathcal{A}} \geq c\, \mathbf{P}_p[\,\mathcal{A}\,]\, (1 - \mathbf{P}_p[\,\mathcal{A}\,])\, \log \frac{1}{2\, m_p^{\mathcal{A}}}\,, \tag{12.21}$$ {e.totalInf}

*where* $m_p^{\mathcal{A}} := \max_i I_p^{\mathcal{A}}(i)$*. Furthermore,*

$$m_p^{\mathcal{A}} \geq c\, \mathbf{P}_p[\,\mathcal{A}\,]\, (1 - \mathbf{P}_p[\,\mathcal{A}\,])\, \frac{\log n}{n}\,. \tag{12.22}$$ {e.maxInf}

*In both inequalities,* $c > 0$ *is an absolute constant.*

For instance, if we can prove for some sequence of monotone events $\mathcal{A} = \mathcal{A}_n$ that $m_p^{\mathcal{A}} \to 0$ as $n \to \infty$, uniformly in a large enough interval of $p$ values around $p^{\mathcal{A}}(n)$, then (12.21) and the Margulis-Russo formula (12.17) show that the threshold interval is small: $\left|p_{1-\epsilon}^{\mathcal{A}}(n) - p_{\epsilon}^{\mathcal{A}}(n)\right| \to 0$ for any $\epsilon > 0$. Note that this condition is going in the direction of excluding locality: a bounded set of small-influence bits usually do not have a noticeable influence even together. Furthermore, if there is a transitive group on $[n]$ under which $\mathcal{A}_n$ is invariant (such as a graph property), then all the individual influences are the same, so $I_p^{\mathcal{A}} = n\, m_p^{\mathcal{A}}$, and (12.22) implies that the threshold interval is at most $C_\epsilon / \log n$. We will see some applications of these ideas in Sections 12.4 and 13.3. The proofs of these influence results, including the Friedgut-Bourgain theorem, use Fourier analysis on the hypercube $\mathbb{Z}_2^n$, as defined briefly in Exercise 12.50.

Note that we have already encountered the idea that small individual influences imply a sharp threshold: the Azuma-Hoeffding large deviation inequality for martingales of bounded increments (Proposition 1.8), used in the concentration results of (6.8) and Exercises 6.12 and 6.13. The connections between influences, the concentration of measure phenomenon and isoperimetric inequalities are explored in depth in [Tal94, Tal96, Led01].

We will now discuss a sharp threshold, concerning the property most relevant from the viewpoint of percolation. On a finite graph, instead of infinite clusters (that contain a given vertex with positive probability), we talk about **giant clusters**, i.e., clusters that occupy a positive fraction of the vertices of $G_n = (V_n, E_n)$. For the Erdős-Rényi model $G(n,p)$, the threshold for the appearance of a giant cluster is around $1/n$:

{t.giant}

**Theorem 12.23** ([ErdR60, Łucz90, Ald97]). *Let $\mathscr{C}_i$ be the $i^{th}$ largest cluster in the random graph $G(n,p)$ with $p \sim (1+\epsilon)/n$, possibly $\epsilon = \epsilon_n$.*

**(i)** *For $n^{-1/3} \ll -\epsilon_n \ll 1$, for any $j \geq 1$ fixed,*

$$\frac{|\mathscr{C}_j|}{2\epsilon_n^{-2}\log(\epsilon_n^3 n)} \xrightarrow{\mathbf{P}} 1\,.$$

*In particular, if $\epsilon_n < \epsilon$ for some fixed $\epsilon < 0$, then the size of the largest cluster is $O(\log n)$.*

**(ii)** *For $\epsilon_n = \lambda n^{-1/3}$, where $\lambda \in \mathbb{R}$ is fixed, the random vector $(|\mathscr{C}_1|, |\mathscr{C}_2|, \dots)/n^{2/3}$ has a non-degenerate limiting distribution $(\xi_1(\lambda), \xi_2(\lambda), \dots)$, which can be described as follows. Consider Brownian motion with drift $\lambda - t$ at time $t$, or in other words, $W_t := B_t + \lambda t - t^2/2$. Now $\xi_i(\lambda)$ is the $i^{th}$ longest excursion of $W_t - \min_{s\in[0,t]} W_s$ away from zero.*

**(iii)** *For $n^{-1/3} \ll \epsilon_n \ll 1$, the largest cluster is*

$$\frac{|\mathscr{C}_1|}{2\epsilon_n n} \xrightarrow{\mathbf{P}} 1\,,$$

*while, for all $j \geq 2$,*

$$\frac{|\mathscr{C}_j|}{2\epsilon_n^{-2}\log(\epsilon_n^3 n)} \xrightarrow{\mathbf{P}} 1\,.$$

*In particular, for $\epsilon_n > \epsilon > 0$, there is a giant cluster, with all other components having size $O(\log n)$.*

*Proof for fixed $\epsilon$.* We sketch a simple proof that uses the exploration random walk that we also used earlier for Galton-Watson trees, a relative of the Depth First Search; see Figure 12.4. A similar approach, with explicit bounds and also handling $\epsilon = \epsilon_n$ varying in the critical window, can be found in [NaPe10]. An even simpler approach than ours, which uses Depth First Search directly, but yields much less, is in [KrSu13].

So, recall the exploration process of Figure 12.4 for Galton-Watson trees, with its list of active vertices $S_0 = 1$ and $S_{i+1} = S_i + X_{i+1} - 1$. For the same process in $G(n,p)$, the difference is that the sequence $\{X_i : i \geq 1\}$, the number of vertices put in the active list in each step, is not an iid sequence, but $X_{i+1} \sim \mathsf{Binom}(n - S_i - i, p)$. So, conditionally on $S_i$, the increment $X_{i+1}$ is still independent of $X_1, \dots, X_i$, and as long as $i = o(n)$ (which implies $S_i = o(n)$ as well, with high probability), the distribution of $X_i$ is still close to $\mathsf{Poisson}(1+\epsilon)$. More precisely, note that the number $U_i$ of unexamined vertices after the $i^{\text{th}}$ step is $U_i \sim \mathsf{Binom}(n-1, (1-p)^i)$, hence

$$X_{i+1} \sim \mathsf{Binom}(n-1, (1-p)^i p)\,, \tag{12.23}$$ {e.exploreXi}

which of course is stochastically dominated by $\mathsf{Binom}(n-1,p)$. Similarly, since $S_i + U_i + i = n$,

$$S_i \sim \mathsf{Binom}(n-1, 1-(1-p)^i) - i + 1\,. \tag{12.24}$$ {e.exploreSi}

**(i)** For $\epsilon < 0$, we get an upper bound on the size of the cluster of a given vertex from being stochastically dominated by a GW tree with offspring distribution $\mathsf{Binom}(n-1,p)$, which has a mean strictly less than 1. The probability of having total volume larger than $n$ is exponentially small, hence the largest cluster in $G(n,p)$ will have a volume $O(\log n)$ by a union bound (and it is easy to see that we indeed get a lot of clusters of this size).

**(ii)** For $\epsilon = 0$, the GW exploration walk has increments of mean zero and finite variance, hence is recurrent. The probability that it stays positive for time at least $k$ is comparable to the probability that it gets to height $\sqrt{k}$, which is $\asymp 1/\sqrt{k}$; see Exercise 6.18. Therefore, by stochastic domination, letting $N_k := |\{v : \text{the cluster of } v \text{ has size } > k\}|$, we have $\mathbf{E}_{G(n,1/n)}[\,N_k\,] \leq O(n/\sqrt{k})$. In particular, for any sequence $k = k_n \gg n^{2/3}$, we get that

$$n^{2/3}\,\mathbf{P}_{G(n,1/n)}[\,N_k \geq 1\,] \ll \mathbf{E}_{G(n,1/n)}[\,N_k\,] \ll n^{2/3}\,, \tag{12.25}$$ {e.DeepTrick}

which implies that $|\mathscr{C}_1| = O(n^{2/3})$ with probability tending to 1.

We can easily get a matching lower bound: if the exploration walk of a given vertex in $G(n,1/n)$ stays positive for time $o(n^{2/3})$ only, then we can start a new run of the exploration walk in the unexplored part of the graph, with increments still close to $\mathsf{Poisson}(1)$, then again and again. It turns out that we can have about $n^{1/3}$ runs before we have explored so many vertices that the increments $X_i$ start becoming visibly smaller, and hence it is very likely that an excursion of duration at least $\delta n^{2/3}$, which happens with probability around $n^{-1/3}/\sqrt{\delta}$, will sooner or later appear among our tries. In more detail: if we denote by $Y_1, \dots, Y_{n^{1/3}}$ the i.i.d. lengths of excursions away from 0 for a random walk $Z_0 = 0, Z_1, Z_2, \dots$ on $\mathbb{Z}$, with i.i.d. $\mathsf{Poisson}(1) - 1$ jumps, which stochastically dominate the lengths of our first $n^{1/3}$ tries in the exploration process, then, for any small $\kappa > 0$,

$$\mathbf{P}[\,Y_1 + \cdots + Y_{n^{1/3}} > Kn^{2/3}\,] = \mathbf{P}\big[\,\#\{1 \leq i \leq Kn^{2/3} : Z_i = 0\} < n^{1/3}\,\big] < \kappa \tag{12.26}$$ {e.RWzeros}

if $K$ is large enough, independently of $n$. See Exercise 12.51 below for a proof. So, we can assume, with probability at least $1-\kappa$, that during the first $n^{1/3}$ tries in the exploration process, the number of unexplored vertices is at least $n - Kn^{2/3}$, and therefore, the downwards drift in the exploration random walk is at most $Kn^{-1/3}$ in each step. An excursion in the i.i.d. walk stays alive for at least $\delta n^{2/3}$ steps with probability around $n^{-1/3}/\sqrt{\delta}$, and conditioning on this event, it is positive with a margin at least $\sqrt{\delta} n^{1/3}$ with a uniformly positive probability (since the conditioned walk stochastically dominates the unconditioned i.i.d. walk). In the actual exploration walk, the drift collected in $\delta n^{2/3}$ steps is at most

$\delta n^{2/3} K n^{-1/3} = \delta K n^{1/3}$, which is less than the margin $\sqrt{\delta} n^{1/3}$ if $\delta$ is small enough. That is, the probability that even the actual exploration walk stays alive is still at least $cn^{-1/3}/\sqrt{\delta}$, with an absolute constant $c > 0$. (To make this argument precise, one can couple the exploration walk with an i.i.d. walk with negative drift, and use an exponential large deviations bound on the latter.) Hence, the overall probability that all of the $n^{1/3}$ runs fail to reach length $\delta n^{2/3}$ is at most $\kappa + (1 - c\,n^{-1/3}/\sqrt{\delta})^{n^{1/3}}$, which is small if first $\kappa$ then $\delta$ are chosen to be small, hence our claim is proved.

**(iii)** For $\epsilon > 0$, the probability of survival of a $\mathsf{Poisson}(1+\epsilon)$ GW tree is $(2 + o(1))\epsilon$ (see Exercise 12.61 in Section 12.4). By stochastic domination and GW duality (Exercise 12.15), with probability at least $1 - (2 + o(1))\epsilon$, the cluster of a fixed vertex has a size that is bounded from above by a random variable that is independent of $n$. How large will the cluster get if the exploration process survives this initial phase and gets unbounded? We claim that, for any $\alpha < 2$, $\epsilon > 0$, and any sequence $k_n \to \infty$,

$$\mathbf{P}\big[\, S_i > 0 \text{ for all } i = k_n + 1, \dots, \alpha\epsilon n \;\big|\; S_{k_n} > 0 \,\big] = 1 - o(1)\,. \tag{12.27}$$ {e.StayingAliv

Indeed, look at (12.24): using $1-(1-p)^i \geq \frac{(1+\epsilon)i}{n} - \frac{(1+\epsilon)^2 i^2}{2n^2}$ (which follows, e.g., from inclusion-exclusion), one can easily check that there exist $\delta = \delta_{\alpha,\epsilon} > 0$ and $I_\epsilon \in \mathbb{N}$ such that, for all $I_\epsilon \leq i \leq \alpha\epsilon n$, we have $1-(1-p)^i > i(1+\delta)/n$, and hence $\mathbf{P}[\, S_i > 0\,] \geq 1 - \exp(-c_{\alpha,\epsilon} i)$ holds by the large deviations bound (1.11). Then, a union bound over $i = k_n+1, \dots, \alpha\epsilon n$ implies (12.27). On the other hand, for any $\beta > 2$ and $\epsilon > 0$, there exists $\delta = \delta_{\alpha,\epsilon} > 0$ such that $1 - (1-p)^{\beta\epsilon n} < \beta\epsilon - \delta$, hence (12.24) now implies that $\mathbf{P}[\, S_{\beta\epsilon n} > 0\,]$ is exponentially small in $n$. That is, the exploration process does not stop until $(2 - o(1))\epsilon n$ vertices are accumulated in the cluster, but then will die soon, reaching only $(2 + o(1))\epsilon n$ vertices.

Summarizing, the cluster is small (independently of $n$) with probability at least $1 - (2 + o(1))\epsilon$, of size $(2 + o(1))\epsilon n$ with probability $(2 + o(1))\epsilon n$, and something else with at most the remaining probability $o(\epsilon)$. Therefore, when we explore the clusters one-by-one, the first few clusters will be small (for an approximately $\mathsf{Geom}(2\epsilon)$ random number of times), and then we get a cluster of size $\sim (2 + o(1))\epsilon n$. At this point, we are left with $n' \sim (1 - 2\epsilon)n$ vertices, while $p = \frac{1+\epsilon}{n} \sim \frac{(1+\epsilon)(1-2\epsilon)}{n'}$, which is strictly subcritical, hence the largest cluster in the rest of the graph will be of size $O(\log n)$. $\square$

Given the above critical behaviour, it is easy to explain intuitively why the critical window is $\epsilon \asymp n^{-1/3}$. At $p = 1/n$, the number of closed edges between any two of the large clusters of size $\asymp n^{2/3}$ is $\asymp n^{4/3}$. Thus, we need to raise $p$ by $\asymp n^{-4/3}$ to see these large clusters merge, and after raising $p$ by a large constant times $n^{-4/3}$, almost all of these clusters have joined together, giving birth to the giant cluster. Exercise 12.65 in Section 12.4 below gives one way to make this intuition precise. Furthermore, see [DinKLP11] for an anatomy of the young giant.

{ex.RWzeros}

▷ **Exercise 12.51.** *The goal of this exercise would be to prove the bound (12.26) above on the zeros of an i.i.d. random walk on the integers. However, I'm still searching for the simplest approach. Meanwhile:*

**(a)** *Using $\mathbf{P}[\, Z_{2k} = 0\,] \asymp 1/\sqrt{k}$, which follows from the local CLT, as in [Dur10, Theorem 3.5.2], together with the First and Second Moment Methods, prove the weaker result that there exist* some $\kappa > 0$ *such that $\mathbf{P}\big[\, \#\{1 \leq i \leq n : Z_i = 0\} > \kappa\sqrt{n}\,\big] > \kappa$, for all $n$.*

**(b)** *For SRW on $\mathbb{Z}$, calculate explicitly the probability $\mathbf{P}[\, Z_{2k} = 0\,]$, then use Stirling's formula to show that the number of zeroes up to time $n$, divided by $\sqrt{n}$ has a limit distribution: the absolute value of a standard normal variable!*

**(c)*** *Using part (a), fill in the details in the strategy outlined in [Pet09] to prove that*

$$\mathbf{P}\big[\, \#\{1 \leq i \leq n : Z_i = 0\} < \kappa\sqrt{n}\,\big] \asymp \kappa\,,$$

*for all $\kappa \in (0,1)$, uniformly in $n$.*

▷ **Exercise 12.52.** *To make sure you understand the proof above for the case $\epsilon > 0$, explain how it is possible that the exploration process runs for $\sim 2\epsilon n$ steps despite the fact that already after cutting off $\gamma\epsilon n$ vertices, for any $\gamma > 1$, the remaining graph is $G((1-\gamma\epsilon)n, \frac{1+\epsilon}{n})$, which is subcritical.*

Beyond the birth of the giant cluster, another natural phase transition is connectedness, which happens at $(1+o(1))\frac{\ln n}{n}$, with a sharp transition. (In fact, Friedgut's above characterization of coarse thresholds implies that coarse thresholds can occur only at critical values of the form $p_c(n) \asymp n^{-a/b}$, $a/b \in \mathbb{Q}$, not at $\frac{\ln n}{n}$.) The following exercise gives the intuition how this value arises.

▷ **Exercise 12.53.** *Let $X_\lambda(n)$ be the number of isolated vertices in the random graph $G(n, \frac{\lambda \ln n}{n})$.*

**(a)** *Show that $\mathbf{E}X_\lambda(n) \sim n^{1-\lambda}$ as $n \to \infty$. Deduce that, for $\lambda > 1$ fixed, with probability tending to 1 there exist no isolated vertices. For $\lambda < 1$ fixed, using the 2nd Moment Method, show that there exist isolated vertices with probability tending to 1.*

**(b)** *Show that if $\alpha > 1-\lambda > 0$, then the probability that there exists a union of components that has total size between $n^\alpha$ and $n - n^\alpha$ is going to 0. This is an indication that isolated vertices are the main obstacles to connectivity.*

There is much beyond percolation on the complete graph $K_n$. The first question is the analogue of $p_c < 1$ for non-1-dimensional graphs:

{c.finitepc}

**Conjecture 12.24** (Benjamini). *Let $G_n = (V_n, E_n)$ be a sequence of connected finite transitive graphs with $|V_n| \to \infty$ and diameter $\mathrm{diam}(G_n) = o(|V_n|/\log|V_n|)$. Then there is $a, \epsilon > 0$ such that*

$$\mathbf{P}_{1-\epsilon}[\textit{there is a connected component of size at least } a|V_n|] > \epsilon$$

*for all large enough $n$.*

▷ **Exercise 12.54.** *Show by example that the $o(|V_n|/\log|V_n|)$ assumption is sharp.*

The next question is the uniqueness of giant clusters.

{c.finiteuniqu

**Conjecture 12.25** ([AloBS04]). *If $G_n$ is a sequence of connected finite transitive graphs with $|V_n| \to \infty$, then for any $a > 0$ and $\epsilon > 0$,*

$$\sup_{p<1-\epsilon} \mathbf{P}_p[\textit{there is more than one connected component of size at least } a|V_n|] \to 0$$

*as $n \to \infty$, where $\mathbf{P}_p$ denotes the probability with respect to $\mathsf{Ber}(p)$ percolation.*

▷ **Exercise 12.55.** *Show by example that the $1-\epsilon$ cutoff is needed.*

Besides the Erdős-Rényi $G(n,p)$, another classical example where these conjectures are known to hold is the hypercube $\{0,1\}^n$, again with critical value $(1+o(1))/n$ [AjKSz82]. Furthermore, they are known for expanders, even without transitivity [AloBS04]. The proof of Conjecture 12.25 for this case proceeds by first showing a simple general upper bound on the average influence, somewhat complementing (12.22): for any increasing event $\mathcal{A} \subset \{0,1\}^{[n]}$, if $p \in (\epsilon, 1-\epsilon)$, then $\exists\, \alpha = \alpha(\epsilon)$ such that, for a uniformly chosen random $i \in [n]$,

$$\mathbf{P}_p[\, i \text{ is pivotal for } \mathcal{A}\,] \le \frac{\alpha}{\sqrt{n}}\,; \qquad (12.28)$$ {e.pmaxInf}

in other words, the total influence is always $I_p^{\mathcal{A}} = O(\sqrt{n})$. (This generalizes Exercise 12.49 (b) from the case $p = 1/2$.) This bound is applied to proving that there cannot be many edges whose insertion would connect two large clusters. On the other hand, two macroscopic clusters in an expander would necessarily produce such pivotal edges, hence there is uniqueness.

Despite the conjectured uniqueness of the giant cluster, there is a possible analogue of the non-uniqueness phase of the non-amenable case: in the intermediate regime, the identity embedding of the giant cluster into the original graph should have large metric distortion. (Locally we see many large clusters, only later they hook up, due to the finiteness of the graph itself.) For the hypercube, the conjectured second critical value is around $1/\sqrt{n}$ [AngB07].

## 12.4 Critical percolation: the plane, trees, scaling limits, critical exponents, mean field theory

{ss.critperc}

Statistical mechanics systems with phase transitions are typically the most interesting at criticality, and Bernoulli percolation is a key example for the study of critical phenomena. Critical percolation is best understood in the plane and on tree-like (so-called mean field) graphs, where the latter is understood broadly and vaguely: e.g., $\mathbb{Z}^d$ for high $d$ is locally tree-like enough and the global structure is simple enough so that critical percolation can be understood quite well, while Gromov-hyperbolic groups are very much tree-like globally, but presently this is not enough for a good understanding. This section will concentrate on critical planar percolation and the case of regular trees, with a brief discussion on more general ideas and the mean field theory.

The planar self-duality of the lattice $\mathbb{Z}^2$ was apparent in our proof of the upper bound in $1/3 \le p_c(\mathbb{Z}^2) \le 2/3$. A more striking (but still very simple) consequence of this self-duality is the following:

{l.half}

**Lemma 12.26.** *The events of having an open left-right crossing in an $n \times (n+1)$ rectangle in $\mathsf{Ber}(1/2)$ bond percolation $\mathbb{Z}^2$, and also in an $n \times n$ rhombus in $\mathsf{Ber}(1/2)$ site percolation $\mathsf{TG}$ both have a probability exactly 1/2, regardless of $n$.*

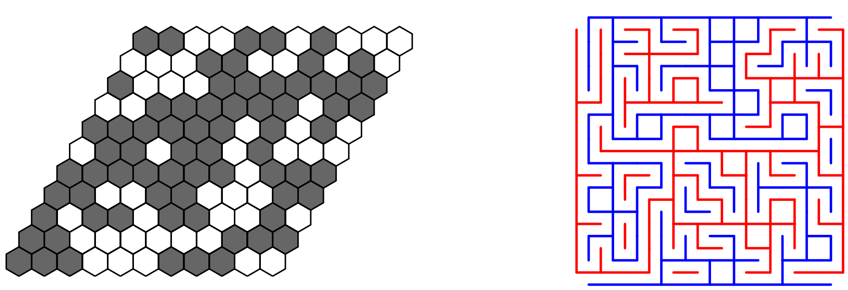


Figure 12.8: The self-duality of percolation on $\mathsf{TG}$ and $\mathbb{Z}^2$. {f.duality}

For the proof, we will need some basic deterministic planar topological results. First of all, opening/closing the sites of $\mathsf{TG}$ is the same as colouring the faces of the hexagonal lattice white/black, so, for better visualization, we will use the latter, as in the left hand picture of Figure 12.8. Now, it is intuitively quite clear that, in any two-colouring, either there is a left-right white crossing, or a top-bottom black crossing in the rhombus, exactly one of the two possibilities. (In other words, a hex game will always end with a winner.) The fact that there cannot be both types of crossings is a discrete version of Jordan's curve theorem, and the fact that at least one of the crossings must occur is a discrete version of Brouwer's fixed point theorem in two dimensions: any continuous map from the closed disk to itself must have a fixed point. However, to actually prove the discrete versions (even assuming the topology theorems), some combinatorial hacking is inevitable, as shown, e.g., by colouring the faces of $\mathbb{Z}^2$: if being neighbours requires a common edge, then it can happen that neither crossing is present, while if it requires only a common corner, then both crossings might be present at the same time.

I am not going to do the discrete hacking in full detail, but let me mention two approaches. The most elegant one I have seen is an inductive proof via Shannon's and Schensted's game of Y, see [PerW10, Section 1.2.3]. A more natural approach is to use the **exploration interface**. Add an extra row of hexagons (an outer boundary) to each of the four sides, the left and right rows coloured white, the top and bottom rows coloured black; the colours of the four corner hexagons will not matter. Now start a path on the edges of the hexagonal lattice in the lower right corner, with black hexagons on the right, white ones on the left. One can show that this path exploring the percolation configuration cannot get stuck, and will end either at the upper left or the lower right corner. Again, one can show that in the first case the right boundary of the set of explored hexagons will form a black path from the top to the bottom side of the rhombus, while, in the second case, the left boundary will form a white path from left to right, and we are done.

For bond percolation on $\mathbb{Z}^2$, a similar statement holds. Given a percolation configuration in the $n\times(n+1)$ rectangle (the red bonds on the right hand picture of Figure 12.8), consider the dual $(n+1)\times n$ rectangle on the dual lattice, with the dual configuration (the blue bonds): a dual edge is open iff the corresponding primal edge was closed. Again, either there is a left-right crossing on the primal graph, or a top-bottom crossing on the dual graph, exactly one of the two possibilities. The exploration interface now goes with primal bonds on its left and dual bonds on its right, with the left and right sides of the rectangle fixed to be present in the primal configuration, and the top and bottom sides fixed to be in the dual configuration.

▷ **Exercise 12.56.** *Assuming the fact that at least one type of crossing is present in any two-colouring of the $n \times n$ rhombus, prove Brouwer's fixed point theorem in two dimensions.*

*Proof of Lemma 12.26.* By the above discussion, if $A$ is the event of a left-right primal crossing in $\mathsf{Ber}(p)$ percolation, then the complement $A^c$ is the event of a top-bottom dual crossing (on $\mathsf{TG}$, primal/dual simply mean open/closed). But, by colour-flipping and planar symmetry, we have $\mathbf{P}_p[\,A^c\,] = \mathbf{P}_{1-p}[\,A\,]$. This says that $\mathbf{P}_p[\,A\,] + \mathbf{P}_{1-p}[\,A\,] = 1$, hence, for $p = 1-p = 1/2$, we have $\mathbf{P}_{1/2}[\,A\,] = 1/2$, as desired. □

For a physicist, the fact that at $p = 1/2$ there is a non-trivial event with a non-trivial probability that is independent of the size of the system suggests that this should be the critical density $p_c$. This intuition becomes slightly more grounded once we know that the special domains considered in Lemma 12.26, where percolation took place, are actually not that special:

{pr.RSW}

**Proposition 12.27** (Russo-Seymour-Welsh estimates)**.** *Consider $\mathsf{Ber}(1/2)$ site percolation on $\mathsf{TG}_\eta$, the triangular grid with mesh size $\eta$, represented as a black-and-white colouring of the hexagonal lattice, so that connections are understood as white paths. We will use the notation $\mathbf{P} = \mathbf{P}^\eta_{1/2}$.*

**(i)** *Let $D \subset \mathbb{C}$ be homeomorphic to $[0,1]^2$, with piecewise smooth boundary, and let $a, b, c, d \in \partial D$ be the images of the corners of $[0,1]^2$. Then*

$$0 < c_0 < \mathbf{P}\big[\,ab \longleftrightarrow cd \text{ in percolation on } \mathsf{TG}_\eta \text{ inside } D\,\big] < c_1 < 1 \tag{12.29}$$ {e.RSWquad}

*for some $c_i(D,a,b,c,d)$ and all $0 < \eta < \eta_0(D,a,b,c,d)$ small enough.*

**(ii)** *Let $A \subset \mathbb{C}$ be homeomorphic to an annulus, with piecewise smooth inner and outer boundary pieces $\partial_1 A$ and $\partial_2 A$. Then*

$$0 < c_0 < \mathbf{P}\big[\, \partial_1 A \longleftrightarrow \partial_2 A \text{ in percolation on } \mathsf{TG}_\eta \text{ inside } A \,\big] < c_1 < 1 \tag{12.30}$$ {e.RSWannu}

*for some $c_i(D,a,b,c,d)$ and all $0 < \eta < \eta_0(A)$ small enough.*

*Similar statements hold for $\mathsf{Ber}(1/2)$ bond percolation on $\mathbb{Z}^2$, with any reasonable definition of "open crossing of a domain".*

*Proof.* The key special case is the existence of some $s > r$ such that crossing an $r \times s$ rectangle in the harder (length $s$) direction has a uniformly positive probability, depending only on $r/s$. In fact, we are going to show the recursive inequality

$$\mathbf{P}\big[\, \mathsf{LR}(r,2s) \,\big] \geq \mathbf{P}\big[\, \mathsf{LR}(r,s) \,\big]^2/4\,, \tag{12.31}$$ {e.RSWkey}

where $\mathsf{LR}(r,s)$ is the left-to-right crossing event in the rectangle $r \times s$ in the horizontal (length $s$) direction, and $r, s$ are arbitrary, except that are chosen relative to the mesh $\eta$ in a way that the vertical midline of the $r \times 2s$ rectangle is an axis of symmetry of the "lattice rectangle" (i.e., the union of lattice hexagons intersecting the rectangle), and both halves "basically" agree with the $r \times s$ lattice rectangle (the hexagons cut into two by the midline are considered to be part of both halves). See Figure 12.9.a.

Then, since we see from Lemma 12.26 that $\mathsf{LR}(\lfloor \sqrt{3} n \rfloor \eta, n\eta) \geq 1/2$ for any $n \in \mathbb{Z}_+$, repeated applications of (12.31) yield the positivity claim for all $r, s$. After this, one can easily convince themselves that gluing open connections in overlapping rectangles via the Harris-FKG inequality implies all the claims of the proposition. We leave the details to Exercise 12.57.

Before proving (12.31), let us note that

$$\mathbf{P}_p[\, \mathsf{LR}(r,(1+2\epsilon)r) \,] \geq \mathbf{P}_p[\, \mathsf{LR}(r,(1+\epsilon)r) \,]^3 \tag{12.32}$$ {e.RSWeasy}

holds trivially for any $p \in [0,1]$ and $\epsilon > 0$, since left-to-right crossings in the leftmost and rightmost $r \times (1+\epsilon)r$ subrectangles of the $r \times (1+2\epsilon)r$ rectangle, plus an up-down crossing in the middle $r \times r$ subsquare, together ensure $\mathsf{LR}(r,(1+2\epsilon)r)$. Compared to (12.32), the main point of (12.31) is that it handles the case $s < r < 2s$.

The following very simple proof of (12.31) is due to Stas Smirnov. (Anecdote: when Oded Schramm received the proof from Smirnov, in the form of a one-page fax containing basically just (12.31) and Figure 12.9, he considered posting the fax to the `arXiv` under Smirnov's name, to make sure that everyone gets to know this beautiful proof.)

Let $D$ be the $r \times 2s$ lattice rectangle, with left and right halves $A$ and $B$. Fixing the outer boundary black along the top side of $A$ and white along its left side, start an exploration interface in the upper left corner until it hits one of the two other sides of $A$. The event of hitting the midline between $A$ and $B$ is equivalent to $\mathsf{LR}(A)$: the right (white) boundary of the stopped interface will be the uppermost left-right crossing of $A$ (provided that a crossing exists). Condition now on this event and also on the right (white) boundary $\gamma$ of the interface. See again Figure 12.9.a. Note that the configuration in the part of $A$ below $\gamma$ and in the entire $B$ is independent of $\gamma$.

Now let the reflection of $\gamma$ across the midline between $A$ and $B$ be $\tilde{\gamma}$; if $\gamma$ ended at a hexagon cut into two by the midline, then this hexagon will be in $\gamma$ and not in $\tilde{\gamma}$. See Figure 12.9.b. In the subdomain of $D$ below $\gamma \cup \tilde{\gamma}$, we can repeat the argument of Lemma 12.26, with boundary arcs $\gamma$, $\tilde{\gamma}$, $\beta$, $\alpha$ in a clockwise order, where $\alpha$ is the bottom side of $A$ together with the part of the left side below $\gamma$, and similarly for $\beta$

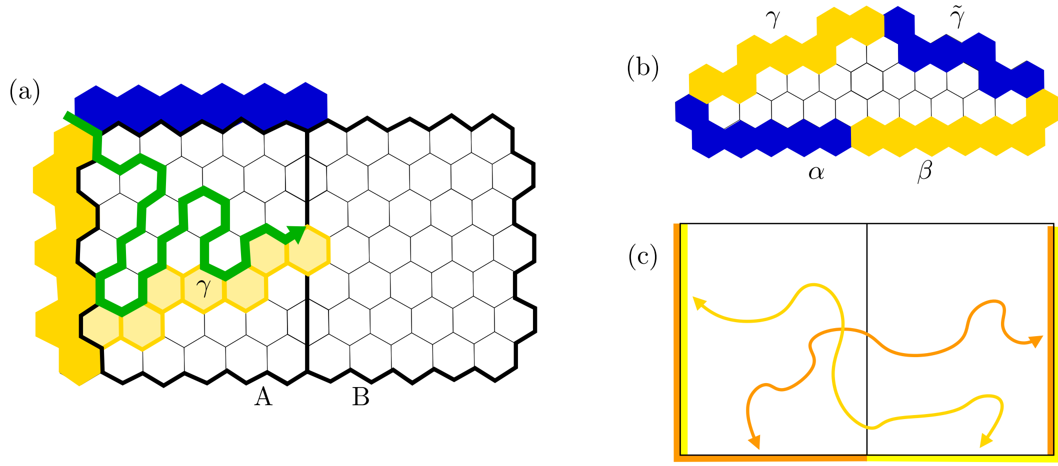



Figure 12.9: Smirnov's proof of RSW. Open/white denoted by yellow, closed/black denoted by blue. {f.RSW}

in $B$. (The possible middle hexagon on the bottom side, just below the midline, will be in $\beta$ and not in $\alpha$.) Namely, there is either a white crossing between $\gamma$ and $\beta$ or a black crossing between $\tilde\gamma$ and $\alpha$, exactly one of the two possibilities. By (almost-)symmetry, we have $\mathbf{P}[\,\gamma \longleftrightarrow \beta\,] \geq 1/2$. If the white crossing between $\gamma$ and $\beta$ occurs, then, together with the white $\gamma$, we get a white crossing from the left side of $A$ to the bottom or right side of $B$. So, denoting this event by $\mathsf{LR}(|A, B\lrcorner)$, we get, after averaging the conditional probability lower bound $1/2$ over all possible choices of $\gamma$, that $\mathbf{P}\big[\,\mathsf{LR}(|A, B\lrcorner)\,\big] \geq \mathbf{P}\big[\,\mathsf{LR}(A)\,\big] \cdot 1/2$. Similarly, we have $\mathbf{P}\big[\,\mathsf{LR}(\llcorner A, B|)\,\big] \geq \mathbf{P}\big[\,\mathsf{LR}(B)\,\big] \cdot 1/2$. Since $\beta$ and its reflection intersect only in at most one hexagon, the intersection of the events $\mathsf{LR}(|A, B\lrcorner)$ and $\mathsf{LR}(\llcorner A, B|)$ implies $\mathsf{LR}(D)$, see Figure 12.9.c. So, by the FKG-Harris inequality,

$$\mathbf{P}\big[\,\mathsf{LR}(D)\,\big] \geq \mathbf{P}\big[\,\mathsf{LR}(|A, B\lrcorner) \cap \mathsf{LR}(\llcorner A, B|)\,\big] \geq \mathbf{P}\big[\,\mathsf{LR}(A)\,\big]\mathbf{P}\big[\,\mathsf{LR}(B)\,\big]/4,$$

which is just (12.31). The next exercise finishes the proof. $\square$

{ex.RSWfin}

▷ **Exercise 12.57.** *Using (12.31), complete the proofs of (i) and (ii) of the proposition, and explain what to change so that the proof works for bond percolation on $\mathbb{Z}^2$, too.*

There are several other proofs of Proposition 12.27, which is important, since they generalize in different ways. The above proof used the symmetry between primal and dual percolation, and hence $p_c = 1 - p_c$, in a crucial way. Nevertheless, the result is also known, e.g., for critical site percolation on $\mathbb{Z}^2$; in particular, $\theta(p_c) = 0$ is known there. One symmetry is still needed there: there is no difference between horizontal and vertical primal crossings, since the lattice has that non-trivial symmetry. One of the most general such results is [GriMan11]. It is also important that on such nice lattices, a joint generalization of (12.31) and (12.32) holds, even in a stronger form [Gri99, Lemma 11.73]: for all $p$,

$$\mathbf{P}_p\big[\,\mathsf{LR}(r, \ell r)\,\big] \geq f_\ell\big(\mathbf{P}_p\big[\,\mathsf{LR}(r)\,\big]\big)\,, \text{ with } f_\ell(x) \geq c(\ell)\, x \text{ and } \lim_{x \to 1} f_\ell(x) = 1\,. \tag{12.33}$$ {e.RSWpkey}

Some people (including Grimmett) call Proposition 12.27 the "box-crossing lemma", and (12.31) or (12.33) the "RSW-inequality".

As we mentioned above, these RSW estimates are the sign of criticality. For instance, they imply the following polynomial decay, which means that $p = 1/2$ is neither very subcritical nor supercritical:

{ex.1arm}

▷ **Exercise 12.58.** *Show that for $p = 1/2$ site percolation on $\mathsf{TG}$ or bond percolation on $\mathbb{Z}^2$, for the* **one-arm probability** $\alpha_1(r,R) := \mathbf{P}[\,\partial B_r(o) \longleftrightarrow \partial B_R(o)\,]$ *and* $\alpha_1(n) := \mathbf{P}[\,0 \longleftrightarrow \partial B_n(o)\,]$, *we have the following* **quasi-multiplicativity** *and* **polynomial decay***:*

**(a)** *there exists $c > 0$ such that, for any radii $\rho \le r/2 < 2r < R$,*

$$c\,\alpha_1(\rho,r)\,\alpha_1(r,R) \le \alpha_1(\rho,R) \le \alpha_1(\rho,r)\,\alpha_1(r,R)\,; \qquad (12.34)$$ {e.qmulti}

**(b)** *there exist constants $c_i, \alpha_i$ such that*

$$c_1\,n^{-\alpha_1} \le \alpha_1(n) \le c_2\,n^{-\alpha_2}\,. \qquad (12.35)$$ {e.1armpoly}

However, more work was needed to fully establish the natural conjecture:

{t.HarrisKesten}

**Theorem 12.28** (Harris-Kesten theorem [Har60],[Kes80])**.** $p_c(\mathbb{Z}^2, \text{bond}) = p_c(\mathsf{TG}, \text{site}) = 1/2$. *Moreover,* $\theta(p_c) = 0$.

*Sketch of proof.* The simpler direction is $p_c \ge 1/2$: the RSW estimates in annuli, part (ii) of Proposition 12.27, imply that the probability of connected to a large distance goes to zero (12.35), hence $\theta(1/2) = 0$. However, Harris did not have the RSW estimates available, hence followed a different route. If there is an infinite white cluster on $\mathsf{TG}$ at $p = 1/2$ a.s., then, by symmetry, there is also an infinite black cluster. We know from Theorem 12.6 that there is a unique white and a unique black infinite cluster. Given the insertion and deletion tolerance and the FKG property, it is pretty hard to imagine that a unique infinite white cluster can exist and still leave space for the infinite black cluster, so this really should not happen. The simplest realization of this idea is due to Yu Zhang (1988, unpublished), as follows.

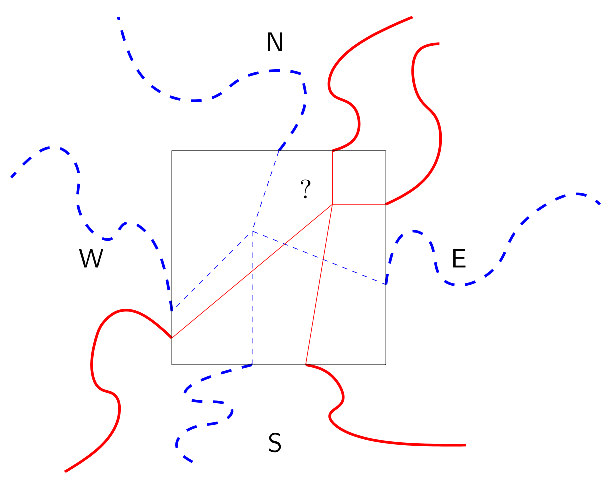



Figure 12.10: Zhang's argument for the impossibility of $\theta(1/2) > 0$. {f.Zhang}

Let $\mathcal{C}_i(n)$ for $i \in \{\mathsf{N},\mathsf{E},\mathsf{S},\mathsf{W}\}$ be the event that the North, East, South, West side of the $n \times n$ box $B(n)$ is connected to infinity within $\mathbb{R}^2 \setminus B(n)$, respectively. Also, let $\mathcal{D}_i(n)$ be the same events in the dual percolation. Assuming $\theta(1/2) > 0$, the box $B(n)$ will intersect the infinite cluster for $n$ large enough, hence, using the FKG-inequality, we have

$$\mathbf{P}\big[\,\mathcal{C}_i(n)^c\,\big]^4 \;\le\; \mathbf{P}\Big[\bigcap_{i\in\{\mathsf{N},\mathsf{E},\mathsf{S},\mathsf{W}\}} \mathcal{C}_i(n)^c\Big] \to 0 \qquad \text{as } n \to \infty\,,$$

and the same for $\mathcal{D}_i(n)$. Therefore,

$$\mathbf{P}\Big[\bigcap_{i\in\{\mathsf{N},\mathsf{E},\mathsf{S},\mathsf{W}\}} \mathcal{C}_i(n)\cap\mathcal{D}_i(n)\Big] > 0$$

for $n$ large enough. However, this event, together with the uniqueness of both the primal and the dual infinite cluster, is impossible due to planar topology. See Figure 12.10.

For the direction $p_c \le 1/2$, the key idea is that at $p = 1/2$ we already have that large boxes are crossed with positive probability (the RSW lemma), hence, with a bit of experience with the sharp threshold phenomena in critical systems, discussed in Section 12.3, it seems plausible that by raising the density to any $1/2+\epsilon$, the probability of crossings in very large boxes will be very close to 1. Indeed, this can be proved in several ways. Kesten, in order to use the Margulis-Russo formula (12.17), first of all proved that $\mathbf{E}_{1/2}[\,|\mathsf{Piv}(n)|\,] \ge c\log n$, where $\mathsf{Piv}(n)$ is the set of pivotal bits for the left-right crossing of the $n\times n$ square. This is not surprising: it is easy to prove using RSW and two exploration paths that there is a pivotal point with positive probability, at distance at least $n/4$, say, from the boundary of the square, and then one can use RSW in dyadic annuli around that first pivotal point to produce logarithmically many pivotals with high conditional probability. See Figure 12.11 and Exercise 12.59.

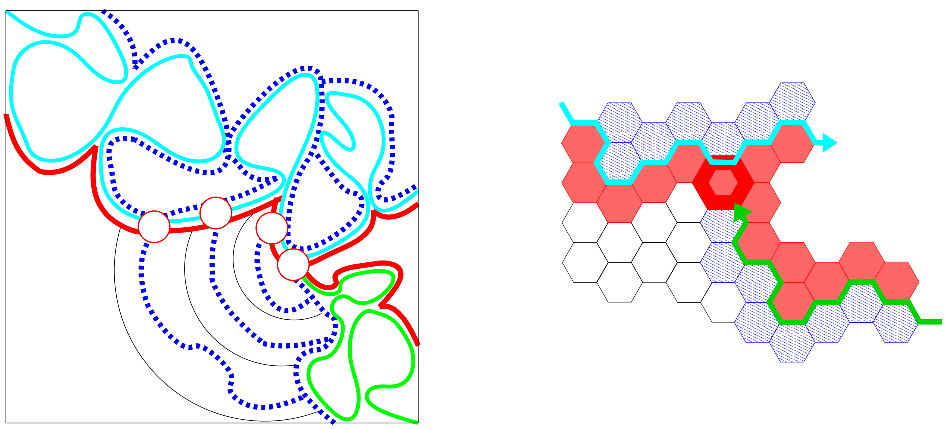


Figure 12.11: Producing $c\log n$ pivotals in an $n$-box with positive probability. {f.logpiv}

{ex.logpiv}

▷ **Exercise 12.59.**

**(a)** *Show that there is at least one pivotal for* $\mathsf{LR}(n)$ *with a uniformly positive probability. (Hint: using an exploration path, find the topmost crossing if it exists, then make sure that there are no further disjoint crossings by using another exploration.)*

**(b)** *Modify the previous part to show that there are pivotals and all of them are at least distance* $n/4$ *from the boundary with positive probability. (Hint: guide the exploration paths using crossings of the right colors.)*

**(c)** *Produce logarithmically many pivotals with positive probability. (Hint: look at Figure 12.11.)*

In fact, the same proof, just inside the $n\times 2n$ rectangle instead of a square, shows that assuming only that $p$ satisfies $1-\epsilon > \mathbf{P}_p[\,\mathsf{LR}(n,2n)\,] > \epsilon$, and using the easy RSW estimate (12.32), we still get $\mathbf{E}_p[\,|\mathsf{Piv}(n,2n)|\,] \ge c(\epsilon)\log n$. Therefore, by the Margulis-Russo formula, $\frac{d}{dp}\mathbf{P}_p[\,\mathsf{LR}(n,2n)\,] \ge c(\epsilon)\log n$ for all $p$ with $1-\epsilon > \mathbf{P}_p[\,\mathsf{LR}(n,2n)\,] > \epsilon$, which means that the size of this $\epsilon$-threshold interval for

$\mathsf{LR}(n,2n)$ must be at most $C(\epsilon)/\log n$. On the other hand, the hard RSW estimate (12.31) tells us that $\mathbf{P}_{1/2}[\,\mathsf{LR}(n,2n)\,] > c > 0$ with a uniform $c > 0$. Therefore, for all $\epsilon, \delta > 0$, if $n > n_{\epsilon,\delta}$, then we have

$$\mathbf{P}_{1/2+\delta}[\,\mathsf{LR}(n,2n)\,] > 1 - \epsilon\,. \tag{12.36}$$ {e.epsdelta}

A slightly different route to arrive at the same conclusion, with less percolation-specific arguments but with more sharp threshold technology, noticed by [BolR06], is to use the BKKKL Theorem 12.22 instead of the Margulis-Russo formula. Then we need to bound only the maximal influence $m_p^{\mathsf{LR}(n)}$ that any bit might have on the event $\mathsf{LR}(n)$. Well, if a bit is pivotal, then it has the alternating 4-arm event to the sides of the $n \times n$ square. Wherever the bit is located, at least one of these primal and one of these dual arms is of length $n/2$, hence

$$m_p^{\mathsf{LR}(n)} \;\le\; \mathbf{P}_p\big[\,0 \longleftrightarrow \partial B(n/2)\,\big] \wedge \mathbf{P}_{1-p}\big[\,0 \longleftrightarrow \partial B(n/2)\,\big] \;\le\; c_2\, n^{-\alpha_2} \qquad \text{for all } p\,,$$

by Exercise 12.58. Then part (12.21) of Theorem 12.22 shows that the $\epsilon$-threshold interval for $\mathsf{LR}(n)$ is at most $C(\epsilon)/\log n$, and then we get (12.36) again.

Now, the highly probable large crossings given by (12.36) can be combined using the FKG property and a simple **renormalization** idea to produce long connections. Namely, take a tiling of the infinite lattice by $n \times n$ boxes, giving a grid $\mathsf{G}_n$ isomorphic to $\mathbb{Z}^2$ (with the $n$-boxes as vertices and with the obvious neighbouring relation), and define the following dependent bond percolation process on $\mathsf{G}_n$: declare the edge between two boxes open if the $n \times 2n$ or $2n \times n$ rectangle given by the union of these two boxes is crossed in the long direction and each box is crossed in the orthogonal direction. See Figure 12.12.

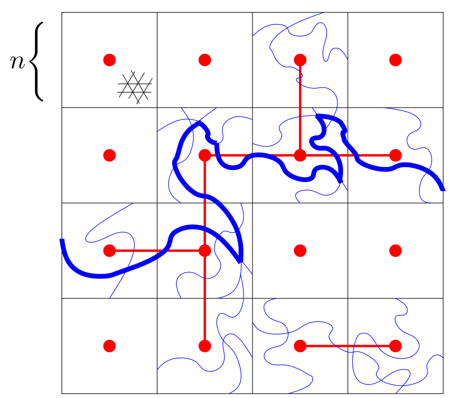



Figure 12.12: Supercritical renormalization: bond percolation on the grid $\mathsf{G}_n$ of $n$-boxes. {f.renorm}

The probability of each edge of $\mathsf{G}_n$ to be open is larger than $1 - \delta$ if $n$ is large enough, and if two edges do not share an endpoint, then their states are independent of each other. Therefore, a small variation on the Peierls contour argument used in the proof of (12.1) applies: by taking every other edge along a dual circuit of length $n$, the probability that this contour is completely closed (i.e., dual-open) is at most $\delta^{n/2}$. If $\delta < 1/9$, then this exponentially small probability beats the exponential number $n3^n$ of dual circuits, and we get the existence of an infinite cluster in the renormalized percolation process. This implies the existence of an infinite cluster also in the original percolation, and finishes the proof of $\theta(1/2 + \epsilon) > 0$. $\square$

At criticality, there are large finite clusters (for instance, by Exercise 12.58), but there is no infinite one. It is tempting to try to define a "just-infinite cluster" at $p_c$. Two natural ideas are to hope that the conditional measures $\mathbf{P}_{p_c}[\,\cdot \mid 0 \longleftrightarrow \partial B_n(o)\,]$ have a weak limit as $n \to \infty$, or that the measures $\mathbf{P}_p[\,\cdot \mid 0 \longleftrightarrow \infty\,]$

have a weak limit as $p \searrow p_c$. It was indeed proved in [Kes86] that both limits exist, and they are equal; this measure is called **Kesten's Incipient Infinite Cluster** or the **IIC**. The proof was recently streamlined and generalized in [BasS17]; we suggest that the interested reader reads that short paper. It was later shown by Kesten's PhD student Antal Járai that many other natural definitions also yield the same measure [Jár03]. Also, see [HamPS15] for a way to produce the IIC from dynamical percolation (where bits are switching between open and closed independently, using Poisson clocks): it is the law of the configuration at a "typical time" when the cluster of the origin is infinite. Such exceptional times exist at $p_c$ on $\mathbb{Z}^2$ and the triangular lattice [SchrSt10, GarPS10a], but not in high dimensions and regular trees [HäPSt97].

▷ **Exercise 12.60.**

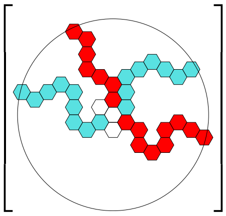
**(a)** *Show that the "conditional FKG-inequality" does not hold: find three increasing events $A, B, C$ in some* $\mathsf{Ber}(p)$ *product measure space such that* $\mathbf{P}_p[\, AB \mid C\,] < \mathbf{P}_p[\, A \mid C\,]\,\mathbf{P}_p[\, B \mid C\,]$.

**(b)** *Show that the conditional FKG-inequality would imply that* $\mathbf{P}_p[\,\cdot \mid 0 \longleftrightarrow \partial B_{n+1}(o)\,]$ *stochastically dominates* $\mathbf{P}_p[\,\cdot \mid 0 \longleftrightarrow \partial B_n(o)\,]$ *restricted to any box* $B_m(0)$ *with* $m < n$. *(However, this monotonicity is not known and might be false, and hence it was proved without relying on it that, for* $p = p_c(\mathbb{Z}^2)$*, these measures have a weak limit as* $n \to \infty$*, the IIC.)*

Here will come a very short intro to the miraculous world of statistical mechanics in the plane at criticality, see [Wer07, Wer04, Schr07]. Very briefly, the main point is that many such models have **conformally invariant scaling limits**: the classical example is that simple random walk on any planar lattice converges after suitable rescaling to the conformally invariant planar Brownian motion (Paul Lévy, 1948). Similar results hold or should hold for the uniform spanning tree, the loop-erased random walk, critical percolation, critical Ising model, the self-avoding walk (the $n \to \infty$ limit of the uniform measures on self-avoiding walks of length $n$), domino tilings (the uniform measure on perfect matchings), the Gaussian Free Field (the "canonical" random height function), and the FK random-cluster models. For instance, for percolation, although the value of $p_c$ is a lattice-dependent local quantity (see Conjecture 14.18), critical percolation itself should be **universal**: "viewed from far", it should look the same and be conformally invariant on any planar lattice, even though criticality happens at different densities. However, the existence and conformal invariance of a critical percolation scaling limit has been proved so far only for site percolation on the triangular lattice, by Stas Smirnov (2001); there is some small combinatorial miracle there that makes the proof work. On the other hand, Oded Schramm noticed in 1999 that using the conformal invariance and the Markov property inherent in such random models, many questions can be translated (via the **Stochastic Löwner Evolution**) to Itô calculus questions driven by a one-dimensional Brownian motion. The methods developed in the last decade are good enough to attack the popular percolation questions regarding **critical exponents** (presently proved for site percolation on the triangular lattice): the one-arm exponent is

$$\alpha_1(n) := \mathbf{P}[\, 0 \longleftrightarrow \partial B_n(0)\,] = n^{-\rho_1 + o(1)}, \ \text{with } \rho_1 = 5/48\,,$$

the alternating four-arm (or pivotality) exponent is

$$\alpha_4(n) := \mathbf{P}\Bigg[$$

$$\Bigg] = n^{-\rho_4 + o(1)}, \ \text{with } \rho_4 = 5/4\,,$$

and the near-critical percolation probability satisfies

$$\theta(p) = (p - p_c)^{\beta + o(1)}, \text{ as } p \searrow p_c, \text{ with } \beta = 5/36 .$$

There is a striking relation between these three exponents:

$$\beta = 5/36 = 5/48 \times \frac{1}{2 - 5/4} = \frac{\rho_1}{2 - \rho_4} . \tag{12.37}$$

{e.KestenScali

This is **Kesten's near-critical scaling relation** [Kes87], which he proved long before the existence and the values of these exponents were verified. Let us explain very briefly the reasons for this relation.

The first step is that

$$\mathbf{E}_{p_c} |\mathsf{Piv}(n)| \asymp n^2 \alpha_4(n) = n^{3/4 + o(1)} \tag{12.38}$$

{e.piv}

holds, whose proof uses two facts: 1) For most vertices in the $n \times n$ square, being pivotal, i.e., having the four alternating arms to the correct boundary pieces, is basically equivalent, up to constant probability factors, to having four alternating arms to a distance comparable to $n$, whose probability is comparable to $\alpha_4(n)$. This is a generalization of the quasi-multiplicativity relation (12.34), and is proved using RSW and FKG technology, but is far from being trivial, since the alternating 4-arm event is not a monotone event. 2) Close to the boundary, the probability of being pivotal starts being different, but it is not actually larger than in the bulk, and there are much fewer such vertices, hence their contribution to the expectation is negligible.

The second step is that (12.38) suggests, via the Margulis-Russo formula (12.17), that the critical window for the left-right crossing event should be around $n^{-3/4}$. This assumes that $\mathbf{E}_p |\mathsf{Piv}(n)|$ remains close to $\mathbf{E}_{p_c} |\mathsf{Piv}(n)|$ in the entire critical window, i.e., that $\alpha_4^p(n)$ satisfies **near-critical stability**. This stability result relies on the observation that the pivotality of a vertex $y$ for the alternating 4-arm event around $x$ is equivalent up to constant probability factors to having the alternating 4-arm event not only around $x$ but also around $y$, and hence $\mathbf{E}_p |\mathsf{Piv}(n)|$ governs not only the change in $\mathbf{P}_p[\,\mathsf{LR}(n)\,]$, but also the change in $\alpha_4^p(n)$. Indeed, taking absolute values in the Margulis-Russo formula (12.17) for non-monotone events, and using that quasi-multiplicativity holds in the entire near-critical window $\big(p^{\epsilon}_{\mathsf{LR}(n)},\, p^{1-\epsilon}_{\mathsf{LR}(n)}\big)$, with quasi-multiplicativity factors depending on $\epsilon$, we obtain

$$\left| \frac{d}{dp} \alpha_4^p(n) \right| \leq C_\epsilon \, n^2 \, \alpha_4^p(n)^2 .$$

Dividing by $\alpha_4^p(n)$ and integrating from $p^\epsilon$ to any $p \in \big(p^\epsilon,\, p^{1-\epsilon}\big)$, we arrive at

$$\big| \log \alpha_4^p(n) - \log \alpha_4^{p^\epsilon}(n) \big| \leq C_\epsilon \big( \mathbf{P}_p[\,\mathsf{LR}(n)\,] - \mathbf{P}_{p^\epsilon}[\,\mathsf{LR}(n)\,] \big) \leq C_\epsilon ,$$

or $e^{-C_\epsilon} \leq \alpha_4^p(n) / \alpha_4^{p^\epsilon}(n) \leq e^{C_\epsilon}$, which is the desired stability result. Moreover, since the change in any $k$-arm probability $\alpha_k^p(n)$ is also governed by $\mathbf{E}_p |\mathsf{Piv}(n)|$, we obtain near-critical stability for them, as well (in particular, for $k = 1$).

This near-critical window of size $n^{-3/4 + o(1)}$ means that, at $p_c + \epsilon$, squares of size larger than $\epsilon^{-4/3 + o(1)}$ are already very well connected, while in squares of size $n \ll \epsilon^{-4/3 + o(1)}$, since $\epsilon \ll n^{-3/4 + o(1)}$, crossing probabilities are still very close to their critical values, hence the configuration looks critical (including the stability of arm-exponents). This characteristic scale $L(\epsilon) = \epsilon^{-4/3 + o(1)}$ is usually called the **correlation length**. Now the point is that

$$\mathbf{P}_{p_c + \epsilon}[\, 0 \longleftrightarrow \infty \,] \asymp \mathbf{P}_{p_c + \epsilon}[\, 0 \longleftrightarrow \partial B_{\epsilon^{-4/3 + o(1)}}(0) \,] = \alpha_1(\epsilon^{-4/3 + o(1)}) = \epsilon^{5/36 + o(1)} , \tag{12.39}$$

{e.KestenScali

giving Kesten's scaling relation (12.37).

Obviously, there were a few leaps of faith in this sketch, but we wanted to draw attention to the fact that near-critical exponents can sometimes be derived from critical ones. A more robust version of this argument, which does not use the Margulis-Russo formula, and can thus be generalized to Poissonian dynamics other than just monotonically changing $p$, and even to non-product measures such as the $\mathsf{FK}(p,q)$ random cluster measures of Section 13.1, is given in [GarPS13b]. Despite this robustness, Kesten's scaling relation turns out to be **false** for the FK random cluster measures [DuCGP13]. The reason is that changing $p$ around $p_c$ in those models turns out to be not at all a Poissonian dynamics; instead, in any **monotone coupling** (analogous to the standard coupling of percolation, which was Poissonian in that case), new edges appear in a fascinating self-organized way that reduces the near-critical window compared to what we would guess from the number of pivotals at criticality. See the end of Section 13.1, around (13.12), for a brief description of how such monotone couplings can be constructed and what they look like.

It is much easier to understand critical and near-critical behaviour of percolation on regular trees.

{ex.treebeta}

▷ **Exercise 12.61.** *Using any of the three methods introduced in Section 12.1 to show the critical mean offspring for Galton-Watson trees, show that on a regular tree $\mathbb{T}_d$, $d \geq 3$, we have $\theta(p) \asymp p - p_c$ as $p \searrow p_c$. In more detail:*

**(a)** *Bound the location of the smaller solution of $f(s) = s$, where $f(s)$ is the generating function of the offspring distribution $\mathsf{Binom}(d, 1/d + \epsilon)$. See Exercise 12.11.*

**(b)** *Refine the first and second moment methods of Exercise 12.12. (Hint for the upper bound: since $\mathbf{E}_{p_c+\epsilon}[\, Z_{1/\epsilon} \,] \asymp 1$, a bound $\theta(p_c+\epsilon) \gg \epsilon$ would imply that $\mathbf{E}_{p_c+\epsilon}[\, Z_{1/\epsilon} \mid Z_{1/\epsilon} > 0 \,] \ll 1/\epsilon$. On the other hand, the recursive nature of the tree and $\theta(p_c + \epsilon) \gg \epsilon$ would imply $\mathbf{E}_{p_c+\epsilon}[\, Z_{1/\epsilon} \mid Z_{1/\epsilon} > 0 \,] \gg 1/\epsilon$.)*

**(c)** *Use the exploration random walk of Figure 12.4 and martingale considerations as in Exercise 6.15. (Actually, since following that exercise, we will want to find an $0 < r \neq 1$ such that $\mathbf{E}[\, r^{\mathsf{Binom}(d,1/d+\epsilon)} \,] = 1$, this solution at the end is quite similar to the one in part (a).)*

The structure of critical GW trees is the easiest to understand in the following special case:

{ex.GWcontour}

▷ **Exercise 12.62.** *Let $T$ be the GW tree with offspring distribution $\xi \sim \mathsf{Geom}(1/2) - 1$. Draw the tree into the plane with root $\rho$, add an extra vertex $\rho'$ and an edge $(\rho, \rho')$, and walk around the tree, starting from $\rho'$, going through each "corner" of the tree once, through each edge twice (once on each side). At each corner visited, consider the graph distance from $\rho'$: let this be process be $\{X_t\}_{t=0}^{2n}$, which is positive everywhere except at $t = 0, 2n$, where $n$ is the number of vertices of the original tree $T$.*

Figure 12.13: The contour walk around a tree. {f.GWcontour}

**(a)** *Using the memoryless property of $\mathsf{Geom}(1/2)$, show that $\{X_t\}$ is SRW on $\mathbb{Z}$.*

**(b)** *Show that $\mathbf{P}[T$ has height $\geq n]$ (i.e., the one-arm probability $\alpha_1(n)$) is $1/n$.*

**(c)** *Show that, conditioning $T$ to have height at least $n$, with high probability the height will be around $n$ and the total volume will be around $n^2$, where "around" means "up to constant factors".*

For general offspring distribution, the task is a bit harder, but far from impossible:

▷ **Exercise 12.63.** * *Prove that, for any critical GW tree with offspring distribution having a variance $\sigma^2 \in (0,\infty)$, the one-arm probability is $\alpha_1(n) := \mathbf{P}[T \text{ has height} \geq n] \asymp 1/n$, with factors depending on $\sigma$. Choose one of the following strategies:* {ex.tree1arm}

**(a)** *Estimate the probability $\mathbf{P}[\, Z_n > 0\,] = f^{\circ n}(0)$ using the generating function $f(s)$ of (12.3).*

**(b)** *Compare the height of the tree with the maximum of the exploration random walk defined around Figure 12.4, and use Exercise 6.18.*

For a general critical GW tree with offspring variance $\sigma^2 \in (0,\infty)$, the contour walk of Figure 12.13 does not have independent increments. Nevertheless, generalizing Exercise 12.62, it was proved by David Aldous that, conditioning the tree to have volume $n$, the contour walk with the usual space-time scaling converges to a Brownian excursion. In other words, the conditioned GW tree, as a metric space, converges in law in the Gromov-Hausdorff topology of metric spaces to the tree defined by a Brownian excursion, called Aldous' **Continuum Random Tree** [Ald91].

▷ **Exercise 12.64.** * *Consider critical percolation on $\mathbb{T}_d$. In Exercise 12.61, we found the off-critical exponent $\beta = 1$, while in Exercise 12.63, we found the one-arm exponent to be $\rho_1 = 1$. Can you relate these two exponents to each other, similarly to Kesten's scaling relation for critical planar percolation, (12.37,12.39)?* {ex.treeKesten

▷ **Exercise 12.65.** * *A task similar to the previous exercise is to prove for the Erdős-Rényi random graph $G(n,p)$ that the critical window where the volumes of the largest clusters are all comparable to $n^{2/3}$ is $p = (1+\lambda n^{-1/3})/n$, $\lambda \in (-\infty,\infty)$. See Theorem 12.23.* {ex.giantwindo

Conjecture 12.10 is open for $\mathbb{Z}^d$ with $3 \leq d \leq 18$, and all non-Abelian amenable groups. For $\mathbb{Z}^d$ with $d \geq 19$, a perturbative Fourier-type expansion method called the Hara-Slade **lace expansion** works, see [Sla06]. Again, this method is good enough to calculate critical exponents, e.g., $\theta(p) \asymp p - p_c$ as $p \searrow p_c$, which are conjecturally shared by many-many transitive graphs, namely all **mean-field** graphs; this should include Euclidean lattices for $d > 6$, all non-amenable groups, and probably most groups "in between". This is very much open, partly due to the problem that lace expansion works best with Fourier analysis, which does not really exist outside $\mathbb{Z}^d$. The method relates Fourier expansions of percolation (or self-avoiding walk, etc.) quantities to those of simple random walk quantities, and if we know everything about Green's function, we can understand how percolation, etc., behaves. (One paper when it is done entirely in "$x$-space", without Fourier, is [HarvdHS03].) The method also identifies that the scaling limit should be the **Integrated Super-Brownian Excursion** on $\mathbb{R}^d$, but does not actually prove the existence of any scaling limit. A readable simple introduction to this scaling limit object is [Sla02].

We have seen that the tree $\mathbb{T}_d$ plays the role of the extremal example among $d$-regular transitive graphs in several different settings: it minimizes the spectral radius, maximizes the speed of SRW, minimizes the value of $p_c$. This is also the case with critical exponents: the mean field exponents are typically the extremal ones. For instance, we have the following result:

{t.thetalin}

**Theorem 12.29.** *For percolation on any infinite transitive graph $G$, assuming that $p_c < 1-\delta$ for some $\delta > 0$, one has*

$$\theta(p) - \theta(p_c) \geq c(\delta)(p - p_c)\,, \text{ for all } p > p_c\,.$$

*In particular, the near-critical exponent $\beta$, whenever exists, is at most* 1.

For the case $\theta(p_c) = 0$, this is a byproduct of the Aizenman-Barsky method of partial differential inequalities [AiB87] that was developed to prove $p_T = p_H$, defined in (12.7); see [Gri99, Theorem 5.48]. It was also proved earlier, but with a small gap, in [ChCh86], using a simpler differential inequality. These references consider percolation on $\mathbb{Z}^d$ only, but the arguments work for any transitive graph with small modifications. Furthermore, Menshikov's proof of $p_T = p_H$ also implies this result, even without the assumption $\theta(p_c) = 0$, but only for transitive graphs of subexponential volume growth; see [Men86] or [Gri99, Theorem 5.8]. The proof we give here is basically due to Gady Kozma (personal communication); it is a simpler, more general, but weaker version of the Aizenman-Barsky approach. Finally, let us mention that the condition $p_c < 1-\delta$ should in fact be vacuous: as we will explain after Schramm's Conjecture 14.18 on the locality of $p_c$, the value 1 seems to be an isolated point in the set of values for $p_c$ for all transitive infinite graphs.

*Proof.* We first of all claim that

$$\mathbf{P}_p[\, n < |\mathscr{C}_o| < \infty \,] \to 0 \text{ as } n \to \infty, \text{ uniformly in any compact subinterval of } (p_c, 1]. \tag{12.40}$$

{e.NoLargeFini

Indeed, pointwise converging continuous functions on a compact interval converge uniformly, and both assumptions are satisfied here: convergence holds for any given $p$, because $\{n < |\mathscr{C}_o|\} \nearrow \{|\mathscr{C}_\infty| = \infty\}$; on the other hand, $\mathbf{P}_p[\, n < |\mathscr{C}_o| < \infty \,]$ is continuous in $(p_c, 1]$ for any given $n$, because we know from Corollary 12.14 that $\theta(p)$ is continuous in this interval, and $\mathbf{P}_p[\, n < |\mathscr{C}_o| \,]$ is clearly continuous, being a polynomial in $p$.

Now we introduce a new variable beside $p$, analogous to an external magnetic field in the Ising model. Color each vertex of $G$ green with probability $\gamma \in [0, 1)$, let $\mathcal{G}$ be the set of green vertices, and let

$$\theta(p, \gamma) := \begin{cases} \mathbf{P}_{p,\gamma}\big[\, \mathscr{C}_o \cap \mathcal{G} \neq \emptyset \,\big] & \text{if } \gamma > 0\,, \\ \theta(p) & \text{if } \gamma = 0\,. \end{cases}$$

It is easy to see that $\lim_{\gamma \searrow 0} \theta(p, \gamma) = \theta(p)$, hence this is a reasonable definition, and one should think of green vertices as being connected directly to infinity.

To be absolutely precise in what follows, we should work with finite approximations $\theta_N(p, \gamma)$ that look only at $\mathscr{C}_o \cap B_N(o)$. These are polynomials hence smooth, and thus can be manipulated without worries; on the other hand, they are so close to $\theta(p, \gamma)$ that the differentiability of that follows in both variables, with the derivatives of $\theta_N$ converging to those of $\theta$. These technicalities are detailed in [Gri99, Appendix I], and we will ignore them here.

Now, for $\theta = \theta(p, \gamma)$, we have the differential inequality

$$\theta \leq \mathbf{P}_{p,\gamma}\big[\, |\mathscr{C}_o \cap \mathcal{G}| = 1 \,\big] + \theta^2 + \theta p \frac{\partial \theta}{\partial p}. \tag{12.41}$$

{e.greendiffin

The first term is self-explanatory. The second term is an upper bound by the BK-inequality on the probability that there exist two open paths from $o$ to $\mathcal{G}$ that are disjoint except for $o$; in other words, that the event $\{\mathscr{C}_o \cap \mathcal{G} \neq \emptyset\}$ holds but there is no pivotal edge for it. Finally, the third term is an upper bound on the probability there is a pivotal edge. Here is a vague intuitive explanation for this bound. By Russo's formula (12.17), the expected number of pivotals is $\frac{\partial \theta}{\partial p}$. This is of course an upper bound on the probability that there exists any, but we want to improve on this by putting in the factor $\theta$ (the factor of $p$ will be irrelevant to us). If there exist pivotal edges, then there is a linear ordering of them, $\{(x_i, y_i)\}_{i=1}^t$, such that each $x_j$ lies on all the open paths from $y_j$ to $o$, and each $(x_i, y_i)$, for $i < j$, lies on all the open paths from $x_j$ to $o$. Thus $(x_t, y_t)$ may be called the "last pivotal edge", and $y_t$ has two disjoint open paths to $\mathcal{G}$. But now it is quite natural that we need to collect at least about $1/\theta$ pivotals before one of them succeeds in sending a second open path to $\mathcal{G}$ — boosting the expected number of pivotals, as desired. An actual proof can be found in [AiB87] or [Gri99, Section 5.3], using a decomposition along the possibilities for the last pivotal edge $(x_t, y_t)$.

A much simpler inequality is that, for $p < 1 - \delta/2$,

$$\theta \le C(\delta) \frac{\partial \theta}{\partial p} \,. \tag{12.42}$$ {e.thetaderiv}

As before, $\partial\theta/\partial p$ is the expected number of pivotal edges. But if we have a configuration where $\mathscr{C}_o$ contains a green site (which happens with probability $\theta$), then we can make a local modification around $o$ — closing all edges but one — and get a configuration with at least one pivotal edge.

Now take any $\epsilon \in (0, \delta/2)$. If $\gamma$ is sufficiently close to 0, we easily get from (12.40) that the first term on the right side of (12.41) is smaller than $\theta/2$ for all $p \in [p_c + \epsilon/2, p_c + \epsilon]$. Rearranging, dividing by $\theta$, then using (12.42), we get

$$1/2 \le (C(\delta) + p) \frac{\partial \theta}{\partial p} \,.$$

This implies that $\theta(p_c + \epsilon, \gamma) \ge \theta(p_c, \gamma) + c\epsilon$, with $c$ independent of $\gamma$. Taking $\gamma$ to zero proves the theorem. $\square$

Mean-field criticality has been proved without lace expansion in some non-amenable cases other than regular trees: for "highly non-amenable graphs" [PaSN00, Scho01], for groups with infinitely many ends and for planar transitive non-amenable graphs with one end [Scho02], and for a direct product of two trees [Koz11]. For tree-like graphs even $p_c$ can be computed in many cases, using multi-type branching processes [Špa09]. However, for these non-amenable cases it is not clear what the scaling limit should be: it still probably should be Integrated Super-Brownian Excursion, but on what object? Some sort of scaling limit of Cayley graphs, a bit similarly to how $\mathbb{R}^d$ arises from $\mathbb{Z}^d$, is the construction of **asymptotic cones**; in particular, Mark Sapir conjectures [Sap07] that if two groups have isometric asymptotic cones, then they have the same critical exponents. An issue regarding the use of asymptotic cones in probability theory was pointed out by Itai Benjamini [Ben08]: the asymptotic cone of $\mathbb{Z}^d$ is $\mathbb{R}^d$ with the $\ell^1$-metric, while the scaling limits of interesting stochastic processes are usually rotationally invariant, hence the $\ell^2$-distance should be more relevant. A possible solution Benjamini suggested was to consider somehow random geodesics in the definition of the asymptotic cone; for instance, a uniform random $\ell^1$-geodesic in $\mathbb{Z}^2$ between $(0,0)$ and $(n,n)$ is very likely to be $O(\sqrt{n})$-close to the $\ell^2$ geodesic, the straight diagonal line. However, it is far from clear that this idea would work in other groups.

## 12.5 Geometry and random walks on percolation clusters {ss.percgeom}

When we take an infinite cluster $\mathscr{C}_\infty$ at $p > p_c(G)$ on a transitive graph, are the large-scale geometric and random walks properties of $G$ inherited? Of course, no percolation cluster (at $p < 1$) can be non-amenable, or could satisfy any $IP_d$, since arbitrarily bad pieces occur because of the randomness. There are relaxed versions of isoperimetric inequalities that are conjectured to survive (known in some cases) but I don't want to discuss them right now. Instead, let me just give some of my favourite conjectures, and see [Pet08] for more details. As we will see in a minute, these questions and ideas are related also to such fundamental questions as Conjecture 12.15 on $p_c < p_u$ above, and Conjecture 14.18 on the locality of $p_c$ below.

**Conjecture 12.30** ([BenLS99]). **(i)** *If $G$ is a transient transitive graph, then all infinite percolation clusters are also transient.* **(ii)** *Similarly, positive speed and zero speed also survive.* {c.BLS}

Part (i) is known for graphs with only exponentially many minimal cutsets (such as Cayley graphs of finitely presented groups) for $p$ close enough to 1 [Pet08]. Part (ii) and hence also (i) are known for non-amenable Cayley graphs for all $p > p_c$ [BenLS99]. Part (ii) is known also for $p$ close enough to 1 on the lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$, $d \geq 3$, and for all $p > p_c$ for $d \leq 2$ [ChPP04].

▷ **Exercise 12.66.** * *Without consulting [LyPer16], show that any supercritical GW tree (you may assume $\mathbf{E}[\,\xi^2\,] < \infty$ for the offspring distribution, if needed), conditioned on survival, is a.s. transient.*

**Conjecture 12.31** (folklore). *The giant cluster at $p = (1+\epsilon)/n$ on the hypercube $\{0,1\}^n$ has poly$(n)$ mixing time, probably $C_\epsilon\, n^2$.*

▷ **Exercise 12.67.** *By exhibiting long "hanging paths" in the giant cluster in $\{0,1\}^n$, show that the mixing time is at least $c_\epsilon n^2$.*

By the result of [AngB07] mentioned in the last paragraph of Section 12.3, the conjecture is proved for $p > n^{-1/2+\epsilon}$, for any $\epsilon > 0$.

A simple discovery I made (in 2003, literally on the margin of a paper by my advisor) was that good isoperimetry inside $\mathscr{C}_\infty$ would follow from certain large deviation results saying that it is unlikely for the cluster $\mathscr{C}_o$ of the origin to be large but finite. My favourite formulation is the following, called **exponential cluster repulsion**.

{c.repulsion}

**Conjecture 12.32** ([Pet08]). *If $G$ is a transitive (unimodular?) graph, then for all $p > p_c(G)$,*

$$\mathbf{P}_p\big[|\mathscr{C}_o| < \infty \text{ but } \exists \text{ an infinite } \mathscr{C}_\infty \text{ with } e(\mathscr{C}_o, \mathscr{C}_\infty) > n\big] < C_p \exp(-c_p n),$$

*where $e(\mathscr{C}_1, \mathscr{C}_2)$ is the number of edges with one endpoint in $\mathscr{C}_1$ another in $\mathscr{C}_2$.*

In [Pet08], I proved this result for $\mathbb{Z}^d$, all $p > p_c(\mathbb{Z}^d)$, and for any infinite graph with only exponentially many minimal cutsets for $p$ close enough to 1. The result implies that a weaker (the so-called anchored) version of any isoperimetric inequality satisfied by the original graph is still satisfied by any infinite cluster. These anchored isoperimetric inequalities are enough to prove, e.g., transience (using Thomassen's criterion (5.2)), therefore, Conjecture 12.32 would imply part (i) of Conjecture 12.30. However, they are only conjecturally strong enough for return probabilities. But, on $\mathbb{Z}^d$, not only the survival of the anchored $d$-dimensional isoperimetric inequality follows, but also the survival of almost the entire isoperimetric profile (as defined in Section 8.2), hence, using the evolving sets theorem, the return probabilities inside $\mathscr{C}_\infty$ are $p_n(x,x) \leq C_d n^{-d/2}$. This was known before [Pet08], but this is the simplest proof by far, or at least the most conceptual.

A nice result related to Conjecture 12.32 (but very far from the exponential decay) is a theorem by Timár: in a non-amenable unimodular transitive graph, any two infinite clusters touch each other only finitely many times, a.s. [Tim06a]

The reason that Conjecture 12.32 is known on $\mathbb{Z}^d$ for $p$ arbitrarily close to $p_c$ is due to a fundamental method of statistical physics (in particular, percolation theory), presently available only on $\mathbb{Z}^d$, called **renormalization**. See [Pet08] for a short description and [Gri99] for a thorough one. This technique uses that $\mathbb{Z}^d$ has a tiling with large boxes such that the tiling graph again looks like $\mathbb{Z}^d$ itself. One algebraic reason for this tiling is the subgroup sequence $(2^k\mathbb{Z}^d)_{k=0}^\infty$, given by the expanding endomorphism $\overline{x} \mapsto 2\overline{x}$, already mentioned in Section 4.4. It is unclear on what groups such tilings are possible:

{q.tiling}

**Question 12.33** ([NekP09]). *A* **scale-invariant tiling** *of a transitive graph $G$ is a decomposition of its vertex set into finite sets $\{T_i : i \in I\}$ such that (1) the subgraphs induced by these tiles $T_i$ are connected and all isomorphic to each other; (2) the following tiling graph $\widehat{G}$ is isomorphic to $G$: the vertex set is $I$, and $(i,j)$ is an edge of $\widehat{G}$ iff there is an edge of $G$ connecting $T_i$ with $T_j$; (3) for each $n \geq 1$, there is such a tiling graph $\widehat{G}^{n+1}$ on $\widehat{G}^n$ in such a way that the resulting nested sequence of tiles $T^n(x) \in \widehat{G}^n$ containing any fixed vertex $x$ of $G$ exhausts $G$.*

*Furthermore, $G$ has a* **strongly scale-invariant tiling** *if each $T^n$ is isomorphic to $T^{n+1}$.*

*If $G$ has a scale-invariant tiling, is it necessarily of polynomial growth?*

From the expanding endomorphism of $\mathbb{Z}^d$ it is trivial to construct a strongly scale-invariant tiling, but this uses the commutativity of the group very much. A harder result proved in [NekP09] is that the Heisenberg group also has Cayley graphs with strongly scale-ivariant tilings.

Although this geometric property of the existence of scale-invariant tilings looks like a main motivation and ingredient for percolation renormalization, it is actually not that important for the presently existing forms of the method. See [NekP09] for more on this. On the other hand, there is a key probabilistic ingredient missing, whose proof appears to be a huge challenge:

{q.unique}

**Question 12.34** ([NekP09]). *Let $G$ be an amenable transitive graph, and let $\mathscr{C}_\infty$ be its unique infinite percolation cluster at some $p > p_c(G)$, with density $\theta(p)$. For a finite vertex set $W \subset G$, let $c_i(W)$ denote the number of vertices in the $i^{th}$ largest connected component of $W$. Does there exist a connected Følner sequence $F_n \nearrow G$ such that for almost all percolation configurations,*

$$\lim_{n\to\infty} \frac{c_2(F_n \cap \mathscr{C}_\infty)}{c_1(F_n \cap \mathscr{C}_\infty)} = 0\,,$$

*moreover,*

$$\lim_{n\to\infty} \frac{c_1(F_n \cap \mathscr{C}_\infty)}{|F_n|} = \lim_{n\to\infty} \frac{|F_n \cap \mathscr{C}_\infty|}{|F_n|} = \theta(p)\,?$$

Why would this be true? The main idea is that the intersection of the unique percolation cluster with a large Følner set should not fall apart into small pieces, because if two points are connected to each other in the percolation configuration, then the shortest connection should not be very long. This is also the underlying idea why random walk on the infinite cluster should behave similarly to the original graph. Furthermore, it is closely related to Conjecture 12.25 on the uniqueness of the giant cluster in finite transitive graphs, and to Oded Schramm's conjecture on the locality of $p_c$, at least in the amenable setting; see Conjecture 14.18 below.

An affirmative answer would follow from an affirmative answer to the following question, which can be asked for any transitive graph, not only amenable ones.

Let $\omega$ denote the percolation configuration, and $\mathrm{dist}_\omega(x,y)$ the **chemical distance**, i.e., the distance measured inside the percolation clusters (infinite if $x$ and $y$ are not connected).

{q.chemical}

**Question 12.35.** *Is it true that for any transitive (unimodular?) graph, for any $p > p_u(G)$ there is a $K(p) < \infty$ such that for any $x, y \in V(G)$,*

$$K_{x,y}(p) := \mathbf{E}\Big[\frac{\mathrm{dist}_\omega(x,y)}{\mathrm{dist}(x,y)} \;\Big|\; \mathrm{dist}_\omega(x,y) < \infty\Big] < K(p) < \infty, \tag{12.43}$$ {e.chemical}

*moreover, there is a $\kappa < 1$ with $K(p) < O(1)\,(p-p_u)^{-\kappa+o(1)}$ as $p \searrow p_u$? The finiteness $K(p) < \infty$ might hold for almost all values $p > p_c$, and the exponent $\kappa < 1$ might hold with $p_u$ replaced by $p_c$; however, there could exist special values (such as $p = p_u$ on certain graphs) with $K(p) = \infty$.*

The existence of $K(p) < \infty$ is known on $\mathbb{Z}^d$ [AntP96], but the $p \searrow p_c$ behaviour is not known.

I do not presently see a conceptual reason for $\kappa < 1$ to hold — this condition is there only because we will need it below. On the other hand, I do not know any example where some $\kappa > 0$ is actually needed. A non-trivial example where one might hope for explicit calculations is site percolation on the hexagonal lattice [Wer07]. If $x$ and $y$ are neighbours in $G$, then the event that they are connected in $\omega$ and the shortest path between them leaves the $r$-ball but does not leave the $2r$-ball around them is comparable to the largest radius of an alternating 4-arm event around them being between $r$ and $2r$. (This rough equivalence uses the so-called separation of arms phenomenon.) On the other hand, the outer boundary of a closed cluster of radius roughly $r$ is an open path having length $r^{4/3+o(1)}$ with large probability. (This follows from the dimension being $4/3$ and from [GarPS13a].) These, together with the 4-arm exponent $5/4$ and the obvious observation that $\mathbf{P}[\,\mathrm{dist}_\omega(x,y) < \infty\,] > c > 0$, imply that

$$n^{-5/4+o(1)} = c_1\,\alpha_4(n) \le \mathbf{P}\big[\,\mathrm{dist}_\omega(x,y) > n \,\big|\, \mathrm{dist}_\omega(x,y) < \infty\,\big] \le c_2\,\alpha_4(n^{3/4}) = n^{-15/16+o(1)},$$

which does not decide the finiteness of the mean at $p_c$, but, using the theory of near-critical percolation and the Antal-Pisztora chemical distance results (which are easy in 2 dimensions), it implies that $\kappa \le 1/9$ in (12.43). This value comes from the following facts: on the level of exponents, ignoring smaller order effects, (1) at $p = p_c + \epsilon$, percolation looks critical up to scales $\epsilon^{-4/3}$ and supercritical at larger scales; (2) the probability that the radius of the 4-arm event is about $\epsilon^{-4/3}$ is $\epsilon^{4/3\cdot 5/4}$; and (3) at this scale, the length of the path is at most $\epsilon^{-4/3\cdot 4/3}$; so, altogether, we get an expectation at most $\epsilon^{5/3-16/9} = \epsilon^{-1/9}$.

Further examples to try are the groups where critical percolation has the mean-field behaviour: here, whenever the mean-field behaviour is actually proved (such as $\mathbb{Z}^d$, $d \ge 19$, or planar non-amenable unimodular transitive graphs), the expected chemical distance between neighbours is known to be finite [KozN09]. However, this is not known to imply that $\lim_{p\searrow p_c} K(p) < \infty$, because continuity here is far from clear.

As Gady Kozma pointed out to me, a natural candidate to kill Question 12.35 for all $p > p_c$ could be percolation near $p_u$ on graphs where there is already a unique infinite cluster at $p_u$. The canonical example of such a graph is a non-amenable planar graph $G$ with one end — here $p_u(G)$ is equal to $1 - p_c(G^*)$, where $G^*$ is the planar dual to $G$. Here, at $p_u(G)$, many things can be understood using the mean-field theory at $p_c(G^*)$, and, our back-of-the-envelope arguments, based on [KozN11], suggest that $\mathbf{P}_{p_u}\big[\,\infty > \mathrm{dist}_\omega(x,y) > r\,\big] \asymp 1/r$ for neighbours $x, y$. Hence $K(p_u) = \infty$, though barely. So, probably any $\kappa > 0$ would do for $p \searrow p_u$; but, again, proving this seems hard.

▷ **Exercise 12.68.** *** *Is it true that the chemical distance in percolation at $p_u$ between neighbours in a planar unimodular transitive non-amenable graph with one end satisfies $K_{x,y}(p_u) = \infty$?*

An affirmative answer to Question 12.35 would also prove that $p_c < p_u$ on non-amenable transitive graphs, via the following unpublished gem of Oded Schramm (which can now be watched in [Per09] or read in [Koz11]):

{t.odedpcpu}

**Theorem 12.36** (O. Schramm)**.** *Let $G$ be a transitive unimodular non-amenable graph, with spectral radius $\rho < 1$. Consider percolation at $p_c$, where we know from Theorem 12.11 that there are only finite clusters. Take a SRW $(X_n)_{n=0}^\infty$ on $G$, independent from the percolation. Then, in the joint probability space of the* *percolation and the SRW,*

$$\mathbf{P}\big[\, X_0 \text{ and } X_n \text{ are in the same percolation cluster at } p_c \,\big] \le 2\rho^n \,.$$

*Proof.* Consider the following branching random walk on $G$. Start $m+1$ particles from $o$, doing independent simple random walks for $n$ steps. Then each particle of this first generation branches into $m$ particles, and all the $(m+1)m$ particles (the second generation) continue with independent SRWs for $n$ more steps, then each particle branches into $m$ particles again, and so on. The expected number of particles at $o$ at the end of the $t^{\text{th}}$ generation, i.e., after $tn$ SRW steps, is $p_{tn}(o,o)\,(m+1)m^{t-1} \approx \rho^{tn} m^t$, where "$\approx$" means equal exponential growth rates. If $m < \rho^{-n}$, then this expectation decays exponentially fast in $t$. Since $p_{2n}(o,x) \le p_{2n}(o,o)$ for any $x \in V(G)$ by Lemma 9.1, we get that for any fixed radius $r > 0$, the expected number of total visits to $B_r(o)$ by all the particles is finite, i.e., the branching random walk is transient.

Now, using this branching walk and $\mathsf{Ber}(p_c)$ percolation on $G$, we define a bond percolation process $\xi$ on the rooted $(m+1)$-regular tree $\mathbb{T}_{m+1}$ that indexes the branching random walk: if $v$ is a child of $u$ in $\mathbb{T}_{m+1}$, then the edge $(u,v)$ will be open in $\xi$ if the branching random walk particles corresponding to $u$ and $v$ are at two vertices of $G$ that are connected in the $\mathsf{Ber}(p_c)$ percolation. So, the probability for any given edge of $\mathbb{T}_{m+1}$ to be open in $\xi$ is $\mathbf{P}[X_0 \overset{p_c}{\longleftrightarrow} X_n]$, but this is, of course, not a Bernoulli percolation. At first sight, it does not even look invariant, since the tree is rooted and the flow of time for the random walk has a direction. However, using that $G$ is unimodular and SRW is reversible, it can be shown that $\xi$ is in fact $\mathrm{Aut}(\mathbb{T}_{m+1})$-invariant:

▷ **Exercise 12.69.** *Show that the above percolation process $\xi$ on $\mathbb{T}_{m+1}$ is automorphism-invariant.*

We have now all ingredients ready for the proof. We know from Theorem 12.11 that we have only finite clusters in $\mathsf{Ber}(p_c)$. Together with the transience of the branching random walk for $m < \rho^{-n}$, we get that $\xi$ has only finite clusters. So, by Exercise 12.31, the average degree in $\xi$ is at most 2. That is, $(m+1)\,\mathbf{P}[X_0 \overset{p_c}{\longleftrightarrow} X_n] \le 2$. So, taking $m := \lceil \rho^{-n} \rceil - 1$, we get that $\mathbf{P}[X_0 \overset{p_c}{\longleftrightarrow} X_n] \le 2/\lceil \rho^{-n} \rceil \le 2\rho^n$, as desired. □

The way this theorem is related to $p_c < p_u$ is the following. (1) The theorem implies that we have a definite exponential decay of connectivity at $p_c$ in certain directions; (2) it seems reasonable that we still should have a decay at $p_c + \epsilon$ for some small enough $\epsilon > 0$ (we will explain this in a second); (3) but then $p_c + \epsilon < p_u$, since from the FKG-inequality it is clear that at $p > p_u$ any two points are connected with a uniformly positive probability: at least with $\theta(p)^2$.

The chemical distance result would be used in making the second step rigorous. Assuming $p_c = p_u$ and taking $p = p_u + \epsilon$ with small $\epsilon > 0$, we would have $X_0$ and $X_n$ in the same cluster with probability $\theta(p)^2 > 0$. On the other hand, their distance in $G$ is at most $n$, hence, by (12.43) and Markov's inequality, $\mathbf{P}_p\big[\mathrm{dist}_\omega(X_0,X_n) > 2K(p)n \;\big|\; \mathrm{dist}_\omega(X_0,X_n) < \infty\big] < 1/2$. So, when lowering the level of percolation from $p$ to $p_u$, the probability that we have a connection between $X_0$ and $X_n$ at $p_u = p_c$ is at least $\theta(p)^2/2 \cdot (1-\epsilon)^{2K(p)n} \asymp \exp(-\epsilon^{1-\kappa+o(1)} n)$, which contradicts Theorem 12.36 if $\kappa < 1$ and $\epsilon$ is chosen sufficiently small.

Finally, as pointed out in [Tim06a], good bounds on the chemical distance (see Question 12.35) can imply good cluster repulsion (see Conjecture 12.32) not only through the renormalization method (see Question 12.34), but also directly. The idea is that for neighbours $x, y \in V(G)$, if $\mathscr{C}_x \neq \mathscr{C}_y$, then a large touching set $e(\mathscr{C}_x, \mathscr{C}_y) = \{e_i : i \in I\}$ would mean that after inserting the edge $\{x,y\}$ into the bond percolation configuration, the endpoints of the edges $e_i$ would be connected but with large chemical distances:

▷ **Exercise 12.70** (Ádám Timár). *Let $G$ be a unimodular transitive graph. Assume that, for some* $\mathsf{Ber}(p)$ *bond percolation, there are infinitely many infinite clusters and for any neighbours $x, y \in V(G)$ we have* $\mathbf{E}_p\big[\operatorname{dist}_\omega(x,y) \bigm| \mathscr{C}_x = \mathscr{C}_y\big] < \infty$. *Show that* $\mathbf{E}_p\big[e(\mathscr{C}_x, \mathscr{C}_y) \bigm| \mathscr{C}_x \neq \mathscr{C}_y\big] < \infty$. *(Hint: use the idea described above, implemented using a suitable Mass Transport.)*

# 13 Further spatial models

{s.spatial}

## 13.1 Ising, Potts, and the FK random cluster models

{ss.Ising}

The Ising and the more general Potts models have a huge literature, partly because they are more important for physics than Bernoulli percolation is. The FK random cluster models, in some sense, form a joint generalization of Potts, percolation, and the Uniform Spanning Trees and Forests, and hence are truly fundamental. Since percolation and the USF already exhibit many of their main features, we will discuss them rather briefly. See [GeHM01] for a great introduction and survey of these models (and further ones, e.g., the hard core model), [Lyo00] specifically for their phase transitions on non-amenable graphs, and [BerKMP05, Sly10] for relationships between phase transitions in spatial, dynamical, and computational complexity behaviour. A future version of our notes will hopefully expand on these phase transitions a bit more.

The **Ising model** is the most natural site percolation model with correlations between the sites. The **Potts**($q$) model is the obvious generalization with $q$ states for each vertex instead of the Ising case $q = 2$. For the definition of these models, we first need to focus on finite graphs $G(V, E)$, with a possibly empty subset $\partial V \subset V$ of so-called boundary vertices. For spin configurations $\sigma : V \longrightarrow \{0, 1, \ldots, q-1\}$, consider the **Hamiltonian** (or **energy function**)

$$H(\sigma) := \sum_{(x,y)\in E(G)} \mathbf{1}_{\{\sigma(x)\neq\sigma(y)\}}\,, \tag{13.1}$$ {e.Hamq}

then fix $\beta \geq 0$ and define the following probability measure (called **Gibbs measure**) on configurations that agree with some given boundary configuration $\eta$ on $\partial V$:

$$\mathbf{P}^\eta_\beta[\sigma] := \frac{\exp(-2\beta H(\sigma))}{Z^\eta_\beta}\,, \quad \text{where} \quad Z^\eta_\beta := \sum_{\sigma:\sigma|_{\partial V}=\eta} \exp(-2\beta H(\sigma))\,. \tag{13.2}$$ {e.Gibbsq}

(The reason for the factor 2 in the exponent will become clear in the next paragraph.) This $Z_\beta$ is called the **partition function**. The more disagreements between spins there are, the larger the Hamiltonian and the smaller the probability of the configuration is. The interpretation of $\beta$ is the inverse temperature $1/T$: at large $\beta$, disagreements are punished more, the system prefers order, while at small $\beta$ the system does not care that much: thermal noise takes over. In particular, $\beta = 0$ gives the uniform measure on all configurations. See Figure 13.1 for an illustration.

A natural extension is to add an external field with which spins like to agree. From now on, we will focus on the Ising case, $q = 2$, and therefore switch to the usual Ising setup $\sigma : V(G) \longrightarrow \{-1, +1\}$, to be interpreted as $+$ and $-$ magnetic spins, instead of $\sigma(x) \in \{0, 1, \ldots, q-1\}$. So, we will consider the Hamiltonian

$$H(\sigma, h) := -h \sum_{x\in V(G)} \sigma(x) - \sum_{(x,y)\in E(G)} \sigma(x)\sigma(y)\,, \tag{13.3}$$ {e.Hamh}

where $h > 0$ means spins like to be 1, while $h < 0$ means they like to be $-1$. Note that $H(\sigma, 0) = 2H(\sigma) - |E(G)|$ with the definition of (13.1), hence, if we now define the corresponding measure and

{f.Ising3}

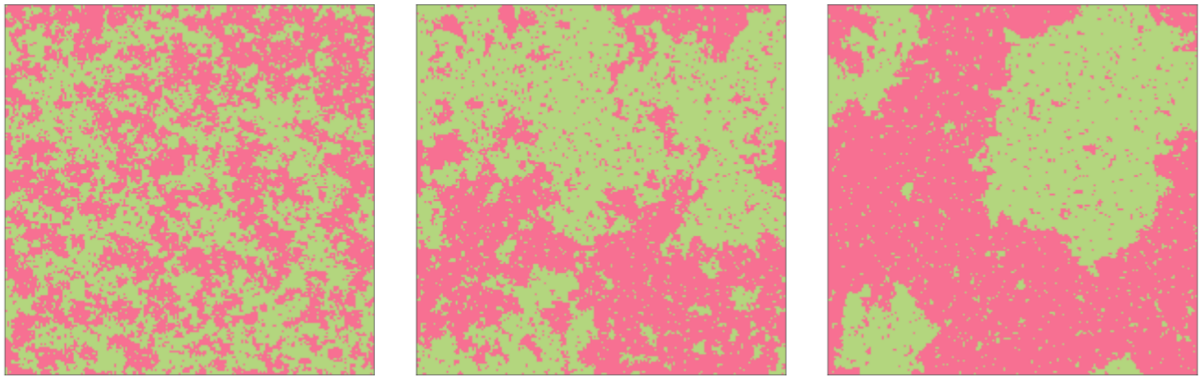


Figure 13.1: The Ising model on a $200 \times 200$ torus at three different inverse temperatures: at $\beta = 0.4$, at $\beta_c = \frac{1}{2}\ln(1+\sqrt{2}) \approx 0.44$, and at $\beta = 0.48$. (Thanks to Gergely Lukáts for the pictures.)

partition function without the factor 2, as

$$\mathbf{P}^{\eta}_{\beta,h}[\sigma] := \frac{\exp(-\beta H(\sigma,h))}{Z^{\eta}_{\beta,h}} \quad \text{and} \quad Z^{\eta}_{\beta,h} := \sum_{\sigma:\sigma|_{\partial V}=\eta} \exp(-\beta H(\sigma,h))\,, \tag{13.4}$$ {e.Gibbsh}

then $\exp(-\beta H(\sigma,0)) = K_\beta \exp(-2\beta H(\sigma))$ with a constant $K_\beta$, and hence $\mathbf{P}^{\eta}_{\beta,0}[\sigma] = \mathbf{P}^{\eta}_{\beta}[\sigma]$.

A further generalization is where edges $(x,y)$ and vertices $x$ have their own "coupling strengths" $J_{x,y}$ and $J_x$, instead of constant 1 and $h$, respectively. If $J_{x,y} > 0$ for all $(x,y) \in E(G)$, i.e., neighbouring spins like to agree, the model is called **ferromagnetic**, while if $J_{x,y} < 0$ for all $(x,y)$, then it is called **antiferromagnetic**. But we will stick to (13.3).

If this is the first time the Reader sees a model like this, they might think that the way of defining the measure $\mathbf{P}^{\eta}_{\beta,h}$ from the Hamiltonian was a bit arbitrary, and $Z_{\beta,h}$ is just a boring normalization factor. However, these are not at all true. First of all, this way the measure $\mathbf{P}^{\eta}_{\beta,h}$ is defined using a product over edges of $G$, hence it clearly satisfies the **spatial Markov property**, or in other words, it is a **Markov random field**: if $U \subset V(G)$ and we set $\partial V = V \setminus U$ and $\partial U = \partial^{\mathrm{out}}_V U \subset \partial V$, then for any boundary condition $\eta$ on $\partial V$, the measure $\mathbf{P}^{\eta}_{\beta,h}$ is already determined by $\eta|_{\partial U}$. In words, if we cut the graph into two by a subset of vertices, then one side depends on the other only through the boundary vertices, just like in a Markov chain, where the future depends on the past only through the present. (The Reader should check this property right now, if it is not clear at first sight!) In fact, the Hammersley-Clifford theorem says that for any graph $G(V,E)$, the Markov random fields satisfying the finite energy condition (as defined in the second proof of Lemma 12.2) are exactly the **Gibbs random fields**: measures that are given by an exponential of a Hamiltonian that is a sum over cliques (complete subgraphs) of $G$. See [Gri10, Section 7.2]. The role of the finite energy condition will be clear from the following example. A colouring of the vertices of a graph with $q$ colours is called **proper** if no neighbours share their colours. The uniform distribution on proper $q$-colourings of a graph is clearly a Markov field, but it does not have the finite energy property. And it is not a Gibbs measure, only in a certain $\beta \to \infty$ limit: it can be considered as the zero temperature antiferromagnetic Potts($q$) model.

Another good reason for defining the measure from the Hamiltonian via $\exp(-\beta H)$ is that, for any given energy level $E \in \mathbb{R}$, among all probability measures $\mu$ on $\{\pm 1\}^{V(G)}$ that satisfy the boundary condition $\eta$ and have $\mathbf{E}_\mu[\,H(\sigma)\,] = E$, our Gibbs measures $\mathbf{P}^{\eta}_{\beta,h}$ maximize the entropy $\mathsf{Ent}(\mu) := -\sum_\omega \mu(\omega)\log\mu(\omega)$, for some $\beta \geq 0$. This is probably due to Boltzmann, proved using Lagrange multipliers. See, e.g., [CovT06, Section 12.1]. So, if we accept the Second Law of Thermodynamics, then we are "forced" to consider these Gibbs measures. Of course, this entropy maximization is closely related to the spatial Markov property: both conditions describe the property of having as much independence as possible.

Similarly to generating functions in combinatorics, the partition function contains a lot of information about the model. The first signs of this are the following:

{ex.partition}

▷ **Exercise 13.1.**

**(a)** *Show that the* **expected total energy** *is*

$$\mathbf{E}_{\beta,h}[\,H\,] = -\frac{\partial}{\partial \beta} \ln Z_{\beta,h}\,, \text{ with variance } \mathrm{Var}_{\beta,h}[H] = -\frac{\partial}{\partial \beta}\mathbf{E}_{\beta,h}[\,H\,]\,.$$

**(b)** *The* **average free energy** *or* **pressure** *is defined by* $f(\beta,h) := (\beta|V|)^{-1} \ln Z_{\beta,h}$. *Show that for the* **total average magnetization** $M(\sigma) := |V|^{-1}\sum_{x\in V}\sigma(x)$, *we have*

$$m(\beta,h) := \mathbf{E}_{\beta,h}[\,M\,] = \frac{\partial}{\partial h} f(\beta,h)\,.$$

**(c)** *The* **susceptibility** *of the total magnetization to a change in the external magnetic field is*

$$\chi(\beta,h) := \frac{1}{\beta}\frac{\partial}{\partial h} m(\beta,h) = \frac{1}{\beta}\frac{\partial^2}{\partial h^2} f(\beta,h)\,.$$

*Relate this quantity to* $\mathrm{Var}_{\beta,h}[M]$. *Deduce that* $f(\beta,h)$ *is convex in* $h$.

So far, we have been talking about the Ising (and Potts) model on finite graphs, only. As opposed to Bernoulli percolation, it is not obvious what the measure on an infinite graph should be. We certainly want any infinite volume measure to satisfy the spatial Markov property, and, for any finite $U \subset V(G)$ and boundary condition $\eta$ on $\partial_V^{\mathrm{out}} U$, the distribution should follow (13.4). Note that this requirement already suggests how to construct such measures: if $\mathbf{P}^\infty_{\beta,h}$ is such a Markov field on the infinite graph $G(V,E)$, and $V_n \nearrow V$ is an exhaustion by finite connected subsets, with induced subgraphs $G_n(V_n,E_n)$, and $\eta_n$ are random boundary conditions on $\partial_V^{\mathrm{out}} V_n$ sampled according to $\mathbf{P}^\infty_{\beta,h}$, then the laws $\mathbf{E}\big[\,\mathbf{P}^{\eta_n}_{\beta,h}\,\big]$ (with the expectation taken over $\eta_n$) will converge weakly to $\mathbf{P}^\infty_{\beta,h}$. Vice versa, any weak limit point of a sequence of measures $\mathbf{P}^{\eta_n}_{\beta,h}$ will be a suitable Markov field. That limit points exist is clear from the Banach-Alaoglu theorem, since $\{-1,1\}^V$ or $\{0,1,\dots,q-1\}^V$ with the product topology is compact. Therefore, the question is: **what is the set of (the convex combinations of) the limit points given by finite exhaustions?** Is there only one limit point, or several? In particular, do different boundary conditions have an effect even in the limit? Intuitively, a larger $\beta$ increases correlations and hence helps the effect travel farther, while $\beta = 0$ is just the product measure, hence there are no correlations and there is a single limit measure. Is there a **phase transition** in $\beta$, i.e., a non-trivial critical $\beta_c \in (0,\infty)$, below which there is only one limit measure, while there are multiple limit measures above? This certainly seems easier to achieve when $h = 0$, since setting $h > 0$ affects every single spin directly, which probably overrides the effect of even a completely negative boundary condition, at least for amenable graphs. This also suggests another way of producing different infinite volume Gibbs measures for $h = 0$: the limits $\lim_{h\to 0+}\lim_{n\to\infty}\mathbf{P}^{\eta_n}_{\beta,h}$ and $\lim_{h\to 0-}\lim_{n\to\infty}\mathbf{P}^{\eta_n}_{\beta,h}$, assuming that they exist, may differ.

To summarize this discussion, here are three possible ways to formalize the existence of a phase transition, both on finite and infinite graphs. It is the existence of some $\beta_c < \infty$ such that for all $\beta > \beta_c$, some of the following properties holds (which is clearly not the case for $\beta = 0$):

(MGS) **Multiple Gibbs states:** On an infinite graph $G$, there exist at least two different Markov random fields $\mathbf{P}^G_{\beta,h}$ whose marginals on finite subsets satisfy the prescribed conditional distributions.

(SSB) **Spontaneous symmetry-breaking:** Let $G_n$ be a sequence of finite graphs, with no boundary condition, and consider the total average magnetization $M_n(\sigma) := |V(G_n)|^{-1} \sum_{x\in V(G_n)} \sigma(x)$. For any $\beta \geq 0$, by the global $\pm$ spin flip symmetry, $\mathbf{E}_{\beta,h=0}[\, M_n \,] = 0$. If

$$\lim_{h\to 0+} \lim_{n\to\infty} \mathbf{E}_{\beta,h}[\, M_n \,] > 0 > \lim_{h\to 0-} \lim_{n\to\infty} \mathbf{E}_{\beta,h}[\, M_n \,],$$

then we say that the $\pm$ symmetry is broken in the limit. (Of course, for an arbitrary graph sequence $G_n$, these limits might not exist. One can take liminfs and limsups in that case.)

(LRO) **Long range order:** Again, let $G_n$ be a sequence of finite graphs, and let $M_n(\sigma)$ be the total average magnetization, with $\mathbf{E}_{\beta,h=0}[\, M_n \,] = 0$. Is there a sort of ergodicity, a Weak Law of Large Numbers, where distant parts of the configuration $\sigma$ are independent enough so that the spatial average converges to the theoretical mean, i.e., $M_n \to 0$ in probability? Or, instead, is there a long range order, where $\sigma$ globally favours either $+$ or $-$, so that $M_n$ is typically of constant order? A clean formulation, almost equivalent to the above alternative: we have (LRO) if $\mathrm{Var}_{\beta,h=0}[M_n]$ does not converge to 0. The virtue of this formulation is that the variance can be calculated once we know the **two-point correlations**:

$$\mathrm{Var}_{\beta,h=0}[M_n(\sigma)] = \frac{1}{|V(G_n)|^2} \sum_{x,y\in V(G_n)} \mathbf{E}_{\beta,h=0}[\, \sigma(x)\, \sigma(y)\,].$$

Therefore, (LRO) means that correlations, at least on average, do not decay with the distance.

In any setting where both (MGS) and (SSB) apply, with the finite graphs $G_n$ forming an exhaustion of $G$, it is obvious that the second implies the first. It is also not surprising that, in such a setting, (LRO) implies (MGS), since the former is about detecting a difference in one specific observable of the measures: total magnetization. It is a bit harder to prove [Leb77] that (LRO) implies (SSB) in great generality. And what is much harder, e.g., proved on the Ising model on $\mathbb{Z}^3$ at $\beta_c$ only recently [AiDCS15], is that having no (LRO), which had been known earlier for all $\beta \leq \beta_c$, implies no (SSB).

Just like in Bernoulli percolation, the simplest examples of finite graphs to analyze are the complete graphs. The **Curie-Weiss model** is the Ising model on the complete graph $K_n$, with edge weights $1/n$, so that the Hamiltonian is

$$H_{n,h}(\sigma) := -h \sum_{i=1}^{n} \sigma_i - \frac{1}{2n} \sum_{i,j=1}^{n} \sigma_i \sigma_j \,.$$

The $1/2$ factor is to make up for having each pair $\{i,j\}$ with $i \neq j$ twice in the sum. The appearance of the terms $i = j$ causes just a shift of $H$ by a constant, which is not visible in $\mathbf{P}_{\beta,h}$. And the weight $1/n$ is to make the effect of the external field to be comparable to the effect of the neighbours.

In terms of the average magnetization $M(\sigma) = \sum_i \sigma_i / n$, note that we can write

$$H_{n,h}(\sigma) = -\big(hM(\sigma) + M(\sigma)^2/2\big)n\,,$$

and the number of $\sigma$'s with $M(\sigma) = x \in \{-1, \frac{-n+2}{n}, \dots, \frac{n-2}{n}, 1\}$ is $\binom{n}{n(1+x)/2}$. Thus,

$$Z_{n,\beta,h} = \sum_x c_{n,\beta,h}(x)\,, \quad \text{where} \quad c_{n,\beta,h}(x) := \binom{n}{n(1+x)/2} \exp\Big(\beta n\big(hx + x^2/2\big)\Big)\,.$$

We can now perform explicit Large Deviations type calculations to show (SSB) and (LRO) phase transition at $\beta_c = 1$, done in the following exercise.

▷ **Exercise 13.2.**

**(a)** *Show that* $f(\beta, h) := \lim_{n\to\infty} f_n(\beta, h) = \lim_{n\to\infty} \frac{\max_x \ln c_{n,\beta,h}(x)}{\beta n}$.

**(b)** *Similarly to Exercise 1.9, show that* $\ln c_{n,\beta,h}(x) = n\big(\beta h x - \Phi_\beta(x)\big) + o(n)$, *where*

$$\Phi_\beta(x) = \frac{1-x}{2}\ln\frac{1-x}{2} + \frac{1+x}{2}\ln\frac{1+x}{2} - \frac{\beta x^2}{2} \qquad \text{for } x \in [-1,1]\,.$$

**(c)** *Sketch the curves* $\Phi_\beta(x)$ *and* $\Phi'_\beta(x)$ *on* $x \in [-1,1]$, *for some parameters* $\beta < 1$, $\beta = 1$, *and* $\beta > 1$.

**(d)** *By choosing the appropriate root* $x = x_0(\beta, h)$ *of* $\Phi'_\beta(x) = \beta h$, *find* $\lim_{n\to\infty} \arg\max_x \ln c_{n,\beta,h}(x)$. *Note that parts (a) and (b) give*

$$\frac{\partial}{\partial h} f(\beta,h) = \frac{\partial}{\partial h}\left( h x_0(\beta,h) - \frac{\Phi_\beta(x_0(\beta,h))}{\beta}\right) = x_0(\beta,h)\,.$$

**(e)** *By part (b) of Exercise 13.1,* $m_n(\beta,h) = \frac{\partial}{\partial h} f_n(\beta,h)$. *Assuming that* $m(\beta,h) := \lim_{n\to\infty} m_n(\beta,h) = \frac{\partial}{\partial h} f(\beta,h)$ *holds for* $h \neq 0$ *(which is indeed the case), deduce from the above that*

$$\lim_{h\to 0+} m(\beta,h) > 0 \qquad \text{and} \qquad \lim_{h\to 0-} m(\beta,h) < 0 \qquad \text{for } \beta > 1\,,$$

*while these limits equal* 0 *for* $\beta \le 1$. *Hence* $m(\beta,h)$ *is discontinuous at* $h = 0$ *iff* $\beta > 1$.

**(f)** *Show that*

$$\frac{1}{\beta}\frac{\partial^2}{\partial h^2} f(\beta,h) = \frac{1}{\beta}\frac{\partial}{\partial h} x_0(\beta,h) = \frac{1 - x_0(\beta,h)^2}{1 - \beta(1 - x_0(\beta,h)^2)}\,.$$

*For* $\beta = 1$, *deduce that* $\frac{\partial}{\partial h} x_0(\beta,h) = \infty$. *That is,* $m(1,h)$ *is continuous but not analytic at* $h = 0$. *Assuming that the limiting susceptibility* $\chi(\beta,h) := \lim_{n\to\infty} \chi_n(\beta,h)$ *equals* $\frac{1}{\beta}\frac{\partial^2}{\partial h^2} f(\beta,h)$, *we get that the limiting susceptibility is* $\chi(1,0) = \infty$. *What does that mean for the variance of the average magnetization?*

**(g)*** *Show that* $\frac{\partial}{\partial h} x_0(\beta, 0+) < \infty$ *for* $\beta > 1$, *so that the limiting susceptibility is finite.*

Back to exhaustions of bounded degree infinite graph, we can now ask: in the case $h = 0$, do we expect a phase transition, in the simplest form of (MGS), for all larger-than-one-dimensional graphs, as in percolation? And actually, is this question related in any way to the connectivity phase transition in percolation?

Before starting to answer these questions, let us prove the **FKG inequality** for the Ising model, as promised in Section 12.2. This beautiful dynamical proof, applicable to a wide range of models, is due to Holley [Hol74].

{t.FKGIsing}

**Theorem 13.1.** *Take the Ising model on any finite graph* $G(V,E)$, *with any boundary condition* $\eta$ *on* $\partial V \subset V$, *and consider the natural partial order on the configuration space* $\{-1,+1\}^V$. *Then, any two increasing events* $A$ *and* $B$ *are positively correlated:* $\mathbf{P}^\eta_{\beta,h}[A \mid B] \ge \mathbf{P}^\eta_{\beta,h}[A]$.

*Proof.* Recall that a probability measure $\mu$ on a poset $(\mathcal{P}, \ge)$ is said to **stochastically dominate** another probability measure $\nu$ if $\mu(A) \ge \nu(A)$ for any increasing measurable set $A \subseteq \mathcal{P}$. Strassen's theorem says that this domination $\mu \ge \nu$ is equivalent to the existence of a **monotone coupling** $\phi$ of the measures $\mu$ and $\nu$ on $\mathcal{P}\times\mathcal{P}$; i.e., a coupling such that $x \ge y$ for $\phi$-almost every pair $(x,y) \in \mathcal{P}\times\mathcal{P}$. The simplest possible example, which we will actually need in a minute, is the following: for $\mathcal{P} = \{-1,+1\}$, we have $\mu \geq \nu$ iff $\mu(+1) \geq \nu(+1)$, and in this case,

$$\phi(+1,+1) := \nu(+1), \qquad \phi(+1,-1) := \mu(+1) - \nu(+1), \qquad \phi(-1,-1) := \mu(-1) \tag{13.5}$$ {e.pm1example}

is the unique monotone coupling $\phi$. (In general, Strassen's coupling need not be unique. This is one reason for the issues around Exercise 11.5.)

Now, the FKG inequality says that $\mathbf{P}[\,\cdot \mid B\,]$ stochastically dominates $\mathbf{P}[\,\cdot\,]$. We will prove this using the **heat-bath dynamics** or **Gibbs sampler**, which is the most natural Markov chain with the Ising model at a given temperature as stationary measure: each vertex has an independent exponential clock of rate 1, and when a clock rings, the spin at that vertex is updated according to the Ising measure conditioned on the current spins of the neighbours. One can easily check that for $\beta < \infty$ the chain is reversible. This is an example of **Glauber dynamics**, which is the class of Markov chains that use independent local updates and keep Ising stationary; another example is the Metropolis algorithm, which we do not define here.

We will run two Markov chains, $\{X_i^+\}_{i\geq 0}$ and $\{X_i^-\}_{i\geq 0}$, started from the all-plus and all-minus configurations on $V \setminus \partial V$, respectively, and fixed to equal $\eta$ on $\partial V$ forever. $\{X_i^-\}_{i\geq 0}$ is just standard heat-bath dynamics, with stationary measure $\mathbf{P}^\eta_{\beta,h}$, while $\{X_i^+\}_{i\geq 0}$ is a modified heat-bath dynamics, where steps in which a $+1$ would change into a $-1$ such that the resulting configuration would cease to satisfy $B$ are simply suppressed (or in other words, replaced by a $+1$ update). A simple general claim about reversible Markov chains (immediate from the electric network representation) is that the stationary measure of the latter chain is simply the stationary measure of the original chain conditioned on the event that we are keeping: $\mathbf{P}^\eta_{\beta,h}[\,\cdot \mid B\,]$. Now, we are not running these two Markov chains independently, but coupled in the following way: the clocks on the vertices ring at the same time in the two chains, and we couple the updates such that $X_i^+ \geq X_i^-$ is maintained for all $i \geq 0$. Why is this possible? If $X_i^+ \geq X_i^-$ and the clock of a vertex $v$ rings, then $v$ has at least as many $+1$ neighbours in $X_i^+$ as in $X_i^-$, hence the probability of the outcome $+1$ in a standard heat-bath dynamics update would be at least as big for $X_i^+$ as for $X_i^-$, and it is even more so if we take into account that some of the $-1$ moves in $\{X_i^+\}_{i\geq 0}$ are suppressed. Hence, by example (13.5), we can still have $X_{i+1}^+ \geq X_{i+1}^-$ after the update.

The Markov chains $\{X_i^+\}_{i\geq 0}$ and $\{X_i^-\}_{i\geq 0}$ are clearly ergodic, converging to their stationary distributions. Since $X_i^+ \geq X_i^-$ holds in the coupling for $i$, this stochastic domination also holds for the stationary measures, and we are done. $\square$

An immediate corollary is that if $\eta \geq \eta'$ on $\partial V$, then $\mathbf{P}^\eta_{\beta,h} \geq \mathbf{P}^{\eta'}_{\beta,h}$. In particular, if $U \subset U' \subset V(G)$ and $\eta$ and $\eta'$ are the all-plus configurations on $\partial^{\mathrm{out}} U$ and $\partial^{\mathrm{out}} U'$, respectively, then $\eta \geq \mathbf{P}^{\eta'}_{\beta,h}$ on $\partial^{\mathrm{out}} U$, and hence $\mathbf{P}^\eta_{\beta,h} \geq \mathbf{P}^{\eta'}_{\beta,h}$ on $U$. Consider now an exhaustion $V_n \nearrow V(G)$ with $\eta_n^+ \equiv +1_{\partial^{\mathrm{out}} V_n}$, giving rise to the monotone decreasing (w.r.t. stochastic domination) sequence of measures $\mathbf{P}^{\eta_n^+}_{\beta,h}$. Any weak limit point of this sequence dominates all other possible limits, given by any sequence $V_n' \nearrow V(G)$ and any $\eta_n'$ on $\partial^{\mathrm{out}} V_n'$, since for any $V_n$ there exists an $m_0(n)$ such that $V_n \subseteq V_m'$ for all $m \geq m_0(n)$. Therefore (see Exercise 13.3), this limit point for $(V_n, \eta_n^+)$ must be unique, and cannot depend even on the exhaustion. It will be denoted by $\mathbf{P}^+_{\beta,h}$. Similarly, there is a unique minimal measure, denoted by $\mathbf{P}^-_{\beta,h}$. It is also natural to consider limit points of the measures $\mathbf{P}^{V_n}_{\beta,h}$, which are the Ising measures on the subgraphs spanned by $V_n$, without any boundary condition. It is not clear at this point that there is a unique limit measure, but we will see in Exercise 13.7 that this is in fact the case; this limit measure will be denoted by $\mathbf{P}^{\mathsf{free}}_{\beta,h}$.

{ex.stochdom}

▷ **Exercise 13.3.**

**(a)** *Show that if $\mu$ and $\nu$ are probability measures on a finite poset $(\mathcal{P}, \geq)$, and both $\mu \geq \nu$ and $\mu \leq \nu$, then $\mu = \nu$. Conclude that if two infinite volume limits of Ising measures on an infinite graph dominate each other, then they are equal.*

**(b)** *Let $\mathcal{P} = \{-1,+1\}^V$ with coordinatewise ordering. Show that if $\mu \leq \nu$ on $\mathcal{P}$, and $\mu|_x = \nu|_x$ for all $x \in V$ (i.e., all the marginals coincide), then $\mu = \nu$. (Hint: use Strassen's coupling.)*

**(c)** *On any transitive infinite graph, the limit measures $\mathbf{P}^+_{\beta,h}$ and $\mathbf{P}^-_{\beta,h}$ are translation invariant. They are equal iff $\mathbf{E}^+_{\beta,h}\sigma(x) = \mathbf{E}^-_{\beta,h}\sigma(x)$ for one or any $x \in V(G)$.*

▷ **Exercise 13.4.** *On any transitive infinite graph, the limit measures $\mathbf{P}^+_{\beta,h}$ and $\mathbf{P}^-_{\beta,h}$ are ergodic.*

In summary, at given $\beta$ and $h$, all infinite volume measures are sandwiched between $\mathbf{P}^-_{\beta,h}$ and $\mathbf{P}^+_{\beta,h}$. So, to answer the question of the uniqueness of infinite volume measures, we "just" need to decide if $\mathbf{P}^-_{\beta,h} \neq \mathbf{P}^+_{\beta,h}$. Ernst Ising proved in 1924 in his PhD thesis that the Ising model on $\mathbb{Z}$ has no phase transition: there is a unique infinite volume limit for any given $h \in \mathbb{R}$ and $\beta \in \mathbb{R}_{\geq 0}$. Based on this, he guessed that there is no phase transition in any dimension. However, he turned out to be wrong: using a variant of the contour method that we saw in the elementary percolation result $1/3 \leq p_c(\mathbb{Z}^2, \mathrm{bond}) \leq 2/3$, Rudolph Peierls showed in 1933 that for $h = 0$ there is a phase transition in $\beta$ on $\mathbb{Z}^d$, $d \geq 2$. More precisely, he proved the existence of some values $0 < \beta_-(d) < \beta_+(d) < \infty$ such that if $\eta_n$ is the all-plus spin configuration on $\partial^{\mathrm{out}}_{\mathbb{Z}^d}[-n,n]^d$, then

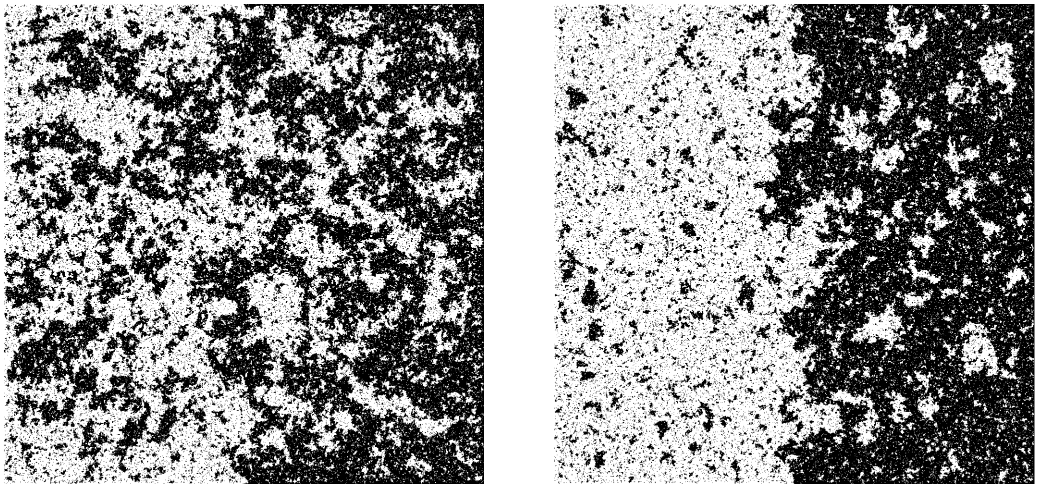
$$\begin{aligned} \liminf_{n\to\infty} \mathbf{E}^{\eta_n}_{\beta,0}\big[\sigma(0)\big] &> 0 \qquad \text{for } \beta > \beta_+\,, \\ \lim_{n\to\infty} \mathbf{E}^{\eta_n}_{\beta,0}\big[\sigma(0)\big] &= 0 \qquad \text{for } \beta < \beta_-\,. \end{aligned} \tag{13.6}$$ {e.IsingPeierl

In particular, for $\beta > \beta_+(d)$, there are at least two ergodic translation-invariant infinite volume measures on $\mathbb{Z}^d$, while for $\beta < \beta_-(d)$ there is uniqueness (see Exercise 13.3 (c)). We will prove (13.6) directly from $1/3 \leq p_c(\mathbb{Z}^2, \mathrm{bond}) \leq 2/3$ once we have defined the FK random cluster measures. A similar result holds for the Potts($q$) models, as well.

In 1944, in one the most fundamental works of statistical mechanics, Lars Onsager showed (employing partly non-rigorous math, with the gaps filled in during the next few decades) that $\beta_c(\mathbb{Z}^2) = \frac{1}{2}\ln(1+\sqrt{2}) \approx 0.440687$ for $h = 0$: for $\beta \leq \beta_c$, there is a unique infinite volume measure, while there is non-uniqueness for $\beta > \beta_c$. He also computed critical exponents like $\mathbf{E}^{\eta_n}_{\beta_c}[\sigma(0)] = n^{-1/8+o(1)}$.

Onsager proved his results by looking at the partition function: the critical points need to occur at the singularities of the limiting average free energy $f_\infty(\beta,h) = -\lim_{|V_n|\to\infty}(\beta|V_n|)^{-1}\log Z_{\beta,h}$, and the critical behaviour must be encoded in the analytic properties of the singularities. Why is this so? From Exercise 13.1, we can see that a big change of the free energy corresponds to big changes of quantities like the total energy and the average magnetization. These quantities are "global", involving the entire finite domain $V_n$, not just a fixed window inside the domain, far from the boundary, hence it is not immediately clear that "local" quantities that behave well under taking weak limits, like $\mathbf{E}[\,\sigma(0)\,]$ or even the average magnetization in the infinite limit measure, will also have interesting changes in their behaviour. Nevertheless, using that $\mathbb{Z}^d$ is amenable, the contribution of the boundary to these "global" quantities turns out to be negligible, and it can be proved that the singularities of $f_\infty(\beta,h)$ describe the critical points.

That there is no phase transition on $\mathbb{Z}^d$ for $h \neq 0$ was first proved in 1972 by Lebowitz and Martin-Löf and by Ruelle, using the **Lee-Yang circle theorem**: if $A = (a_{i,j})_{i,j=1}^n$ is a real symmetric matrix, then all the roots of the polynomial $P(z) := \sum_{S\subseteq[n]} z^{|S|}\prod_{i\in S, j\notin S} a_{i,j}$ lie on the unit circle. Therefore, the Ising partition function $Z_{\beta,h}$ on any finite graph, as a polynomial in $h$, can have roots only at purely



$\beta = 0.440687$ $\beta = 0.45$

Figure 13.2: The Ising model with Dobrushin boundary conditions (black on the right side of the box and white on the left), at the critical and a slightly supercritical inverse temperature. (Vincent Beffara's pictures.

imaginary values of $h$. This can be used to prove that, for any $h \neq 0$, any limit $f_\infty(\beta, h)$ is differentiable in $\beta$. This works for any infinite graph $G(V, E)$. A different approach (by Preston in 1974) is to use the so-called **GHS concavity inequalities** [GHS70] to prove the same differentiability. However, the connection between differentiability and uniqueness works only on amenable transitive graphs: Jonasson and Steif proved that a transitive graph is nonamenable iff there is an $h \neq 0$ such that $\mathbf{P}^+_{\beta,h} \neq \mathbf{P}^-_{\beta,h}$ for some $\beta < \infty$. See [JoS99] and the references there for pointers to the above discussion.

We have hinted a couple of times at the Ising correlation and uniqueness of measure questions being analogous to the phase transition in the existence of infinite clusters in Bernoulli percolation. Indeed, correlations between Ising spins can be interpreted as connectivity in a different model, the $\mathbf{FK}(p, q)$ **random cluster model** with $q = 2$. Here is the definition of the model, due to Fortuin and Kasteleyn [ForK72]; see [Gri06] for a thorough treatment of the model, including the history. On a finite graph $G(V, E)$, with a so-called boundary $\partial V \subset V$ together with a partition $\pi$ of it into disjoint subsets, for any $\omega \subset E$, let

$$\mathbf{P}^\pi_{\mathsf{FK}(p,q)}[\omega] := \frac{p^{|\omega|}\,(1-p)^{|E\setminus\omega|}\,q^{k_\pi(\omega)}}{Z^\pi_{\mathsf{FK}(p,q)}} \qquad \text{with} \quad Z^\pi_{\mathsf{FK}(p,q)} := \sum_{\omega\subseteq E} p^{|\omega|}\,(1-p)^{|E\setminus\omega|}\,q^{k_\pi(\omega)}\,, \tag{13.7}$$ {e.FK}

where $k_\pi(\omega)$ is the number of clusters of $\omega/\pi$, i.e., the clusters given by $\omega$ in the graph where the vertices in each part of $\pi$ are collapsed into a single vertex.

The $q = 1$ case is clearly just Bernoulli($p$) bond percolation. Furthermore, if we let $q \to 0$, then we punish a larger number of components more and more, so we get a single spanning cluster in the limit, and then if we let $p \to 0$, then we will have as few edges as possible, but otherwise all configurations will have the same probability; that is, we recover the Uniform Spanning Tree, see Section 11.2. If $G(V, E)$ is infinite, and $V_n \nearrow V$ is an exhaustion by finite subsets, then taking $\partial V_n := \emptyset$ gives $\mathsf{FUSF}$ in the limit, while taking $\partial V_n := \partial^{\mathrm{in}}_V V_n$ with $\pi_n := \{\partial V_n\}$ gives $\mathsf{WUSF}$.

For $q \in \{2, 3, \dots\}$, we can retrieve the Potts($q$) model via the **Edwards-Sokal coupling**, which was introduced somewhat implicitly in [SwW87] and explicitly in [EdS88]. Given $G(V, E)$ and a boundary partition $\pi$, color each cluster of $\omega/\pi$ in the $\mathsf{FK}(p, q)$ model independently with one of $q$ colors, then forget $\omega$, just look at the $q$-coloring of the vertices. We will prove in a second that we indeed get the Potts($q$) model, with

$$\beta = \beta(p) = -\frac{1}{2}\ln(1-p)\,, \tag{13.8}$$ {e.betap}

except that the boundary condition on $\partial V$, instead of a function $\eta : \partial V \longrightarrow \{0, 1, \dots, q-1\}$, will be only a partition telling which spins in $\partial V$ have to agree with each other. Note that, in the resulting Potts($q$) model, the correlation between the spins $\sigma(x)$ and $\sigma(y)$ is exactly the probability $\mathbf{P}^{\pi}_{\mathsf{FK}(p,q)}[x \longleftrightarrow y]$, since if they are connected, they get the same color in the coupling, and if they are not, then their colors will be independent. In particular, the Ising expectation $\mathbf{E}^{+}_{\beta(p),2}[\sigma(0)]$ is the probability of the connection $\{0 \longleftrightarrow \partial[-n, n]^d\}$ in the wired FK measure $\mathsf{FK}(p, 2)$ on $[-n, n]^d$. (In the Edwards-Sokal coupling, to get the Ising $+$ measure, we need to condition the measure to have the $+1$ spin on $\partial[-n, n]^d$, but, by symmetry, that does not change the probability of the connection.) The interpretation of correlations as connections also shows that it is more than natural that a larger $p$ gives a higher $\beta$ in formula (13.8).

Let us denote the Edwards-Sokal coupling of the $\mathsf{FK}(p, q)$ configuration $\omega$ and the $q$-coloring $\sigma$ of its clusters by

$$\mathbf{P}^{\pi}_{\mathrm{ES}(p,q)}[\omega, \sigma] = \frac{p^{|\omega|}(1-p)^{|E\setminus\omega|}\prod_{(x,y)\in\omega\cup\pi}\mathbf{1}_{\{\sigma(x)=\sigma(y)\}}}{Z^{\pi}_{\mathsf{FK}(p,q)}}\,, \tag{13.9}$$ {e.ES}

where the formula is clear from the facts that for any given $\omega$, the number of compatible $q$-colorings is $q^{k_\pi(\omega)}$, and that we need to arrive at (13.7) after summing over these colorings $\sigma$.

So, we need to prove that the marginal on $\sigma$ is the Potts($q$) model with $\beta = \beta(p)$ and boundary partition $\pi$. Fix a $q$-coloring $\sigma$; we may assume that it is compatible with $\pi$ (i.e., $\sigma(x) = \sigma(y)$ for all $(x, y) \in \pi$), otherwise $\sum_\omega \mathbf{P}^{\pi}_{\mathrm{ES}(p,q)}[\omega, \sigma] = 0$. If $(x, y) \in E$ and $\sigma(x) = \sigma(y)$, then the configurations $\omega$ with $\mathbf{P}^{\pi}_{\mathrm{ES}(p,q)}[\omega, \sigma] \neq 0$ come in pairs: $(x, y)$ can be kept in $\omega$ or deleted, leaving other edges intact. On the other hand, if $\sigma(x) \neq \sigma(y)$, then $(x, y) \notin \omega$ whenever $\mathbf{P}^{\pi}_{\mathrm{ES}(p,q)}[\omega, \sigma] \neq 0$, i.e., such an edge $(x, y)$ always contributes a factor $1 - p$ to the probability of the configuration. Summarizing,

$$\begin{aligned}
\sum_\omega \mathbf{P}^{\pi}_{\mathrm{ES}(p,q)}[\omega, \sigma] &= \frac{1}{Z^{\pi}_{\mathsf{FK}(p,q)}} \prod_{\substack{(x,y)\in E\\ \sigma(x)=\sigma(y)}} (p + 1 - p) \prod_{\substack{(x,y)\in E\setminus\pi\\ \sigma(x)\neq\sigma(y)}} (1-p) \prod_{(x,y)\in\pi} \mathbf{1}_{\{\sigma(x)=\sigma(y)\}} \\
&= \frac{1}{Z^{\pi}_{\mathsf{FK}(p,q)}} \exp\left(-2\beta \sum_{(x,y)\in E\setminus\pi} \mathbf{1}_{\{\sigma(x)\neq\sigma(y)\}}\right) \prod_{(x,y)\in\pi} \mathbf{1}_{\{\sigma(x)=\sigma(y)\}}\,,
\end{aligned}$$

using the choice $(1 - p) = \exp(-2\beta)$, which is (13.8). This measure on $\sigma$ is exactly the Potts($q$)-measure (13.2), just with boundary partition $\pi$, and with normalization $Z_{\mathsf{FK}(p,q)}$ instead of $Z_{\beta,q}$, but since we are talking about probability measures, these normalizations actually have to agree:

$$Z^{\pi}_{\mathsf{FK}(p,q)} = Z^{\pi}_{\beta(p),q}\,. \tag{13.10}$$ {e.PottsFKpart

This concludes the verification of the Edwards-Sokal coupling.

The $\mathsf{FK}(p, q)$ model satisfies the **FKG-inequality** for $q \geq 1$. First of all, why is $q \geq 1$ important? Well,

$$\mathbf{P}_{\mathsf{FK}(p,q)}\Big[(x,y)\in\omega \,\Big|\, \omega|_{E\setminus\{e\}}\Big] = \begin{cases} p & \text{if } \{x \overset{\omega}{\longleftrightarrow} y\} \text{ in } E\setminus\{e\} \\ \frac{p}{p+(1-p)q} & \text{otherwise}\,; \end{cases} \tag{13.11}$$ {e.FKFKG}

i.e., an existing connection increases the probability of an edge iff $q > 1$. Having noted this, the proof of the FKG inequality becomes very similar to the Ising case, Theorem 13.1: consider the FK heat-bath dynamics with independent exponential clocks on the edges, and updates following (13.11). For $q \geq 1$, this dynamics $\{\omega_i\}_{i\geq 0}$ is attractive in the sense that if $\omega_i \geq \omega'_i$ in the natural partial order on $\{0,1\}^E$, then $\mathbf{P}\big[\, e \in \omega_{i+1} \bigm| \omega_i \,\big] \geq \mathbf{P}\big[\, e \in \omega'_{i+1} \bigm| \omega'_i \,\big]$, and hence we can maintain the monotone coupling of the proof of Theorem 13.1 for all $i \geq 0$, proving the FKG inequality.

For $q < 1$, there should be negative correlations, but this is proved only for the $\mathsf{UST}$, which is a determinantal process. This is an open problem that have been bugging quite a few people for quite a long time; see [BorBL09] for recent results on negative correlations.

{ex.mon2}

▷ **Exercise 13.5.** *For the* $\mathsf{FK}(p,q)$ *model on any finite graph* $G(V,E)$ *with boundary* $\partial V \subset V$*, for* $q \geq 1$*, show the following two types of stochastic domination:*

**(a)** *If* $\pi \leq \pi'$ *on* $\partial V$*, then* $\mathbf{P}^{\pi}_{\mathsf{FK}(p,q)} \leq \mathbf{P}^{\pi'}_{\mathsf{FK}(p,q)}$ *on* $V$*.*

**(b)** *Given any* $\pi$ *on* $\partial V$*, if* $p \leq p'$*, then* $\mathbf{P}^{\pi}_{\mathsf{FK}(p,q)} \leq \mathbf{P}^{\pi}_{\mathsf{FK}(p',q)}$ *on* $V$*.*

**(c)** *Conclude for the + limit Ising measure on any infinite graph* $G(V,E)$ *that if* $\mathbf{E}^{+}_{\beta,h}[\sigma(x)] > 0$ *for some* $x \in V$*, then the same holds for any* $\beta' > \beta$*. Consequently, the uniqueness of the Ising limit measures is monotone in* $\beta$*.*

{ex.mon3}

▷ **Exercise 13.6.** *Show a third type of stochastic domination for the* $\mathsf{FK}(p,q)$ *model on a finite graph: if* $p \in (0,1)$ *and* $1 \leq q \leq q'$*, then* $\mathbf{P}^{\pi}_{\mathsf{FK}(p,q')} \leq \mathbf{P}^{\pi}_{\mathsf{FK}(p,q)}$*.*

Similarly to the Ising model, the FKG inequality implies that the fully wired boundary condition (where $\pi$ has just one part, $\partial V$) dominates all other boundary conditions, and the free boundary (where $\pi$ consists of singletons) is dominated by all other conditions. Therefore, in any infinite graph, the limits of free and wired FK measures along any finite exhaustion exist and are unique, denoted by $\mathsf{FFK}(p,q)$ and $\mathsf{WFK}(p,q)$. On a transitive graph, they are translation invariant and ergodic.

However, regarding limit measures, there is an important difference compared to the Ising model. In formulating the spatial Markov property over finite domains $U \subset V(G)$ for an infinite volume measure, the boundary conditioning is on all the connections in $V \setminus U$ (just like in (13.11)), which is not as local as it was for the Potts($q$) model. Therefore, it is not clear that any infinite volume limit measure actually satisfies the spatial Markov property, neither that any Markov measure is a convex combination of limit measures. Although these statements are expected to hold, they have not been proved; see [Gri06, Chapter 4] for more information. Nevertheless, it is at least known that all Markov measures and all limit measures are sandwiched between $\mathsf{FFK}(p,q)$ and $\mathsf{WFK}(p,q)$ in terms of stochastic domination.

{ex.inftyFKIsi

▷ **Exercise 13.7** (Ising and $\mathsf{FK}$ limit measures)**.**

**(a)** *Show that the + Ising measure* $\mathbf{P}^{+}_{\beta,0}$ *on any infinite graph is given by taking the* $\mathsf{WFK}(p,2)$ *measure, with* $p = p(\beta)$ *according to (13.8), then coloring all the vertices in all the infinite* $\mathsf{WFK}$*-clusters +, while coloring the finite clusters randomly with a fair coin. The − measure* $\mathbf{P}^{-}_{\beta,0}$ *is given analogously, by coloring the infinite clusters −, while the free measure* $\mathbf{P}^{\mathsf{free}}_{\beta,0}$ *is given by coloring all the clusters of* $\mathsf{FFK}(p,2)$ *measure randomly.*

**(b)** *How can we get the +, − and free Ising measures from* $\mathsf{FK}$ *measures in the case* $h \neq 0$*?*

Since we are interested in the connectivity properties of the FK model, it is natural to define the **critical point** $p_c(q)$ in any infinite volume limit measure as the infimum of $p$ values with an infinite cluster. Fortunately, on any amenable transitive graph, there is actually only one $p_c(q)$, independently of which limit measure is taken, because of the following argument, see [Gri06]. Using convexity, any of the limiting average free energy functions has only a countable number of singularities in $p$, which implies that, for any $q$, there is only a countable number of $p$ values where $\mathsf{FFK}(p,q) \neq \mathsf{WFK}(p,q)$. It is clear that $p_c^{\mathsf{F}}(q) \geq p_c^{\mathsf{W}}(q)$, but if this was a strict inequality, then the free and wired measures would differ for the entire interval $p \in (p_c^{\mathsf{W}}(q), p_c^{\mathsf{F}}(q))$, contradicting countability.

We can easily show that $0 < p_c(q) < 1$ on $\mathbb{Z}^d$, for any $d \geq 2$ and $q \geq 1$. The key is to notice that (13.11) implies that $\mathbf{P}^{\pi}_{\mathsf{FK}(p,q)}$ stochastically dominates Bernoulli($\tilde{p}$) for $\tilde{p} = \frac{p}{p+(1-p)q}$, and is stochastically dominated by Bernoulli($p$) bond percolation, for any $\pi$, and then the claim follows from $0 < p_c(\mathbb{Z}^d) < 1$ in Bernoulli percolation. The proof of (13.6), namely the existence of a phase transition for the Ising model on $\mathbb{Z}^d$, $d \geq 2$, is also clear now: as we noticed above, the Ising expectation $\mathbf{E}^{+}_{\beta(p),2}[\sigma(0)]$ is exactly the probability of the connection $\{0 \longleftrightarrow \partial[-n,n]^d\}$ in the wired FK measure $\mathsf{FK}(p,2)$ on $[-n,n]^d$. But this probability is bounded from above and below by the probabilities of the same event in Bernoulli($p$) and Bernoulli($\tilde{p}$) percolation, respectively, and we are done.

The random cluster model also provides us with an explanation where Onsager's value $\beta_c(\mathbb{Z}^2) = \frac{1}{2}\ln(1+\sqrt{2}) \approx 0.440687$ comes from. Recall that the percolation critical values $p_c(\mathbb{Z}^2, \text{bond}) = p_c(\mathsf{TG}, \text{site}) = 1/2$ came from planar self-duality, plus two main probabilistic ingredients: RSW bounds proved using the FKG-inequality, and the Margulis-Russo formula; see Theorem 12.28. So, to start with: is there some planar self-duality in $\mathsf{FK}(p,q)$? Consider the planar dual to a configuration $\omega$ on a box with free boundary, say: we get a configuration $\omega^*$ on a box with wired boundary. See Figure 13.3. What is the law of $\omega^*$?

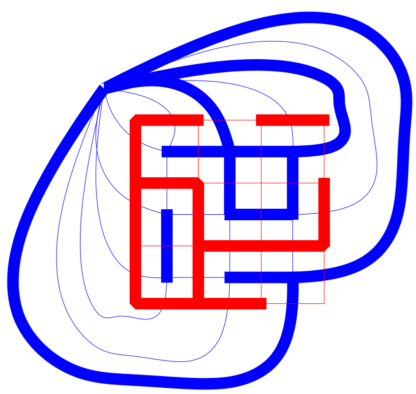


Figure 13.3: The planar dual of an FK configuration on $\mathbb{Z}^2$. {f.FWFK}

Clearly, $|\omega^*| = |E| - |\omega|$, and $k(\omega^*)$ equals the number of faces in $\omega$ (which is 2 in the figure). So, by Euler's formula, $|V| - |\omega| + k(\omega^*) = 1 + k(\omega)$. If we now let $y = p/(1-p)$, then

$$\mathbf{P}_{\mathsf{FK}(p,q)}[\omega] \;\propto\; y^{|\omega|}\, q^{k(\omega)} \;\propto\; y^{-|\omega^*|}\, q^{k(\omega^*)+|\omega^*|} = \left(\frac{q}{y}\right)^{|\omega^*|} q^{k(\omega^*)}\,.$$

Well, this is a random cluster model for $\omega^*$, with the same $q$ and $y^* = q/y$! Or, in terms of $p$ and $p^*$, we get

$$\frac{p\, p*}{(1-p)\,(1-p^*)} = q\,.$$

Therefore, $p = p_{\mathrm{sd}}(q) = \frac{\sqrt{q}}{1+\sqrt{q}}$ is the **self-dual point** on $\mathbb{Z}^2$: the planar dual of the free measure becomes the wired measure at the same value $p_{\mathrm{sd}}$. Just as in the case of percolation with $p = 1/2$, one naturally expects that this is also the critical point $p_c(q)$. This was proved for all $q \geq 1$ only rather recently, by

Vincent Beffara and Hugo Duminil-Copin [BefDC10]. Substituting $q = 2$ and $\beta(p) = -\frac{1}{2}\ln(1-p)$ gives Onsager's value. Note that $p_c(q) < p_c(q')$ for $q < q'$ is in line with Exercise 13.6.

The main obstacle Beffara and Duminil-Copin needed to overcome was that the previously known RSW proofs for percolation (e.g., the one we presented in Proposition 12.27) do not work in the presence of dependencies: an exploration path has two sides, one having positive, the other having negative influence on increasing events, so we cannot use FKG. Nevertheless, they found an only slightly more complicated argument, exploring crossings and gluing them, where the information revealed can be easily compared with symmetric domains with free-wired-free-wired boundary conditions, and hence one can use symmetry, just as in the percolation case. However, here comes another issue: the dual of the free measure is wired, so, what measure should we work in to have exact self-duality and the symmetries needed? The solution is to use periodic boundary conditions, i.e., to work on a large torus, and draw the rectangle that we want to cross inside there. Finally, the replacement for the sharp threshold results for product measures via the Margulis-Russo formula or the BKKKL theorem (see (12.17) and Theorem 12.22) is provided by [GraG06].

What about the **uniqueness of infinite volume measures** for $\mathsf{FK}(p,q)$? On $\mathbb{Z}^2$, uniqueness is known for $q = 2$ and all $p$. Non-uniqueness at the single value $p_c(q)$ is expected for $q > 4$, and proved for $q > 25.72$. See [DuCS11, Proposition 3.10] and [Gri06]. Note that assuming uniqueness of the measures at the self-dual point $p_c(q)$, Zhang's argument works again, yielding that there is no percolation at $p_c(q)$. Conversely, non-uniqueness of the measures implies that there is an infinite cluster in the wired measure (basically a supercritical system), and there is no infinite cluster in the free measure (basically a subcritical system). Indeed, if there are only finite clusters even in the wired measure, then any finite box $B$ is surrounded by a dual circuit $\gamma$; but the configuration in $B$ conditioned on any specific such $\gamma$ is dominated by the infinite volume free measure, implying that in fact we are in the free measure; see [DuC14, Corollary 4.40]. In summary, non-uniqueness of measures at $p_c(q)$ is the same as **discontinuity of the phase transition**.

How does the intuitive "inherited geometry" argument presented after Conjecture 12.10 on the continuous phase transition in the case of percolation break for these $\mathsf{FK}(p,q)$ models with a discontinuous phase transition? The answer is that, for $\omega_p \subset E(\mathbb{Z}^2)$ sampled from the $\mathsf{WFK}(p,q)$ measure, which has a unique infinite cluster at $p_c(q)$ but no infinite cluster at any $p < p_c(q)$, the difference between the edge marginals of $\omega_{p_c}$ and $\omega_p$ does not tend to zero as $p \nearrow p_c$ [Gri06, Proposition 8.59], hence one cannot obtain $\omega_p$ from $\omega_{p_c}$ by deleting a tiny fraction of the edges, and there is no reason why the infinite cluster of $\omega_{p_c}$ could not be ruined.

When looking at the $\mathsf{FK}(p,q)$ measure for the first time, it is easy to think that $p$ should roughly stand for edge density, hence this discontinuity in the edge marginals is somewhat hard to believe. However, it turns out that any **Markovian monotone coupling** $(\omega_p)_{p\in[0,1]}$ of the $\mathsf{FK}(p,q)$ measures for $q > 1$ must be quite different from the standard coupling for $q = 1$. Here is the only such coupling that I know of; it was implicit in [Hol74], made explicit in [Gri95], and proved to be Markovian in $p$ in [DuCGP13] (but maybe this was also known to many people before). Let $G(V,E)$ be any finite graph and $\Omega$ be the space $[0,1]^E$. The goal is to find a measure $\mu = \mu_G$ on $\Omega$ in a such a way that all the "projections" $\omega_p(Z)$ with $Z \sim \mu$, defined by

$$\omega_p(Z)(e) := \mathbf{1}_{\{Z(e)\leq p\}}\,, \qquad p \in [0,1],\ e \in E\,,$$

follow the random-cluster probability measure of parameters $(p,q)$ on $\{0,1\}^E$ with some given boundary conditions. It turns out that it is non-trivial to construct such a measure $\mu$ explicitly (although its existence follows from a generalized Strassen's theorem). Instead, [Gri95] obtains it as the invariant measure of a natural Markov process $Z_t$ on the space $\Omega := [0,1]^E$.

Let $Z_t$ be a Markov chain on $\Omega$ where labels on the edges are updated at rate one according to the conditional law defined below. For any $e = (x,y) \in E$, let $\mathcal{D}_e \subset \{0,1\}^E$ be the event that there is a path of open edges in $E \setminus \{e\}$ connecting $x$ and $y$. For any $e \in E$ and any $Z \in \Omega$, define

$$T_e(Z) := \inf\{p \in [0,1] \text{ s.t. } \omega_p(Z) \in \mathcal{D}_e\} .$$

Let $U_e = Z_t(e)$ be the new label at $e$ and time $t$ knowing the current configuration $Z_{t-}$ (before the update), given by the law

$$\mathbf{P}\big[\, U_e \le p \,\big] := \begin{cases} p & \text{if } p \ge T \\ \frac{p}{p+(1-p)q} & \text{if } p < T\,, \end{cases} \tag{13.12}$$ {e.Ue}

where $T = T_e(Z_{t-})$. Note that this is the only possible update rule if we want that the projections $\omega_p(Z_t)_{t\ge 0}$ follow the heat-bath dynamics for $\mathsf{FK}(p,q)$ given by (13.11) simultaneously for all $p \in [0,1]$. The condition $q \ge 1$ implies that this is a valid distribution function, hence we can simply define $U_e$ to be a sample from this distribution. Note that $U_e$ has an absolutely continuous part plus **an atom** (for $q > 1$) at $T$, namely $[T - \frac{T}{T+(1-T)q}]\delta_T$.

Now, one can get the edge marginal $\mathbf{P}[\, e \in \omega_p(Z)\,]$ from this coupling by first conditioning on all the labels $Z(f)$ for $f \neq e$, then taking the expectation of the conditional probability (13.12). The location $T$ of the atom turns out to be quite concentrated around $p_c(q)$, which makes the edge intensity change faster than linearly around $p_c$ for $q \ge 2$, and even produces a discontinuity at $p_c$ in the infinite volume limit for $q > 4$ (proved for $q > 25.72$), making the discontinuity of the percolation probability possible. An additional fascinating property of this monotone coupling was discovered in [DuCGP13]: due to the atoms, there are many edges $e$ sharing the same label $Z(e)$, and these edges arrive not in a roughly uniform Poissonian way (as the edges in the Gibbs sampler), but under some complicated **self-organization scheme**. This scheme is responsible for the fact that near-critical window in the Ising-FK model is not given by the expected number of pivotal edges at criticality, but is narrower, as already mentioned after (12.39). However, the details of this self-organization scheme are quite mysterious at present.

The case $0 \le q \le 1$ is even more mysterious, as we already remarked around Exercises 13.5 and 13.6 in relation with correlation inequalities.

**Question 13.2.** *Is there a monotone coupling $(\omega_p)_{p\in[0,1]}$ of the $\mathsf{FK}(p,q)$ measures for any $q \in [0,1)$?*

▷ **Exercise 13.8.** *Consider the graph $G$ with 6 vertices and 7 edges that looks like a figure 8 on a digital display. Consider the $\mathsf{UST}$ measure on the 15 spanning trees of $G$, and the uniform measure $\mathsf{UST}+1$ on the 7 connected subgraphs with 6 edges (one more than a spanning tree). Find an explicit monotone coupling between the two measures.*

Some further results and questions on the FK model will be mentioned in Section 14.2.

## 13.2 Extremal Gibbs measures and factor of IID processes

{ss.fiid}

On any infinite graph $G(V,E)$, consider the set of Markov random fields, with values in some $\Omega = S^V$, corresponding to a given set of compatible finite dimensional conditional distributions (sometimes called **local specifications**); for instance, think of the set of Gibbs measures corresponding to a given Hamiltonian (that is given by a sum over cliques in $G$, as in the Hammersley-Clifford theorem). This is clearly a closed convex set in the space of all measures on $\Omega$, hence, by the Krein-Milman theorem, it is given by the closure of the convex hull of its extremal points. These **extremal** Markov random fields are also called **pure states** in the statistical physics literature. Clearly, "understanding a model" should typically include a description of its extremal measures and the main properties of those. The following result says that extremal measures can be recognized intrinsically:

{pr.extrtail}

**Proposition 13.3** (Extremality is tail-triviality [LaRu69])**.** *A Markov random field $\{X_v\}_{v\in V}$ on some countable set $V$ is extremal iff its* **tail-$\sigma$-field** *$\mathcal{T}$ is trivial.*

*Here, similarly to Kolmogorov's 0-1 Theorem 9.20, the definition is $\mathcal{T} = \bigcap_{n\geq 1} \sigma\{X_v : v \notin V_n\}$, where $V_n$ is any exhaustion of $V$ by finite sets, and triviality means that any event in $\mathcal{T}$ has probability 0 or 1.*

*Proof.* The key step in the proof is the following lemma:

{l.extrtail}

**Lemma 13.4.** *Let $\mathbf{P}$ be a Markov random field on $\Omega = S^V$, and assume that $\mu$ is another measure on $\Omega$ that is absolutely continuous w.r.t. $\mathbf{P}$. Then $\mu$ is a Markov random field (with the same local specifications) iff the Radon-Nikodym derivative $\phi(\omega) = \frac{d\mu}{d\mathbf{P}}(\omega)$ has a $\mathcal{T}$-measurable version (i.e., a tail-measurable function $\tilde{\phi}(\omega)$ that $\mathbf{P}$-almost surely coincides with $\phi$).*

From the lemma, the proof of the proposition is simple. On the one hand, if $(\Omega, \mathcal{T}, \mathbf{P})$ is non-trivial, then take any event $A \in \mathcal{T}$ with $\mathbf{P}[\,A\,] = p \in (0,1)$, and write $\mathbf{P}[\,\cdot\,] = p\,\mathbf{P}[\,\cdot \mid A\,] + (1-p)\,\mathbf{P}[\,\cdot \mid A^c\,]$. By the lemma, this is a convex combination of two Markov random fields, hence $\mathbf{P}$ is non-extremal.

The reverse direction is similar: if we can write $\mathbf{P} = p\mathbf{P}_0 + (1-p)\mathbf{P}_1$, a combination of Markov random fields, then $\mathbf{P}_i \ll \mathbf{P}$, hence the lemma says that both $\phi_i = d\mathbf{P}_i/d\mathbf{P}$ are $\mathcal{T}$-measurable. If $\mathcal{T}$ is trivial, then both $\phi_i$ are constant $\mathbf{P}$-a.s., but both are density functions, hence both are constant 1, hence both $\mathbf{P}_i = \mathbf{P}$, and the convex combination was in fact trivial.

To prove the lemma, assume that $\mu$ is Markov, with the same local specifications as $\mathbf{P}$. Take an exhaustion $V_n \nearrow V$ by finite sets. If $f : \Omega \longrightarrow \mathbb{R}$ is a function that depends on finitely many coordinates only, then for large enough $n$, these coordinates are all contained in $V_n$. Let $\mu_n$ be the marginal of $\mu$ on $V_n^c$. It is clear from $\mu \ll \mathbf{P}$ that $\mu_n \ll \mathbf{P}$; denote the Radon-Nikodym derivative by $\phi_n = d\mu_n/d\mathbf{P}$. Now,

$$\begin{aligned}
\mathbf{E}_{\mathbf{P}}[\,\phi(\omega) f(\omega)\,] = \mathbf{E}_{\mu}[\,f(\omega)\,] &= \mathbf{E}_{\mu}\big[\,\mathbf{E}_{\mu}[\,f(\omega) \mid \omega_{V_n^c}\,]\,\big] \\
&= \mathbf{E}_{\mu_n}\big[\,\mathbf{E}_{\mathbf{P}}[\,f(\omega) \mid \omega_{V_n^c}\,]\,\big] \\
&= \mathbf{E}_{\mathbf{P}}\big[\,\phi_n(\omega_{V_n^c})\,\mathbf{E}_{\mathbf{P}}[\,f(\omega) \mid \omega_{V_n^c}\,]\,\big] = \mathbf{E}_{\mathbf{P}}[\,\phi_n(\omega) f(\omega)\,]\,,
\end{aligned}$$

where we used the $\mathbf{P}$-Markov-property of $\mu$ and the definition of $\mu_n$ to get the second line, the $\omega_{V_n}^c$-measurability of $\phi_n$ in the third line, and the definition $\phi_n(\omega) := \phi_n(\omega_{V_n^c})$ at the end. This identity clearly implies that $\phi_n(\omega) = \phi(\omega)$ for $\mathbf{P}$-almost all $\omega$. By taking $\tilde{\phi}(\omega) := \liminf_{n\to\infty} \phi_n(\omega)$, we get a $\mathcal{T}$-measurable version of $\phi$, as desired.

The backward direction of the lemma is again similar to the forward direction: if there is a $\mathcal{T}$-measurable RN-derivative $\phi$, then, for any measurable $f : \Omega \longrightarrow \mathbb{R}$ and finite $U \subset V$,

$$\begin{aligned}
\mathbf{E}_{\mu}[\,f(\omega)\,] = \mathbf{E}_{\mathbf{P}}[\,\phi(\omega) f(\omega)\,] &= \mathbf{E}_{\mathbf{P}}\big[\,\mathbf{E}_{\mathbf{P}}[\,\phi(\omega) f(\omega) \mid \omega_{U^c}\,]\,\big] \\
&= \mathbf{E}_{\mathbf{P}}\big[\,\phi(\omega)\,\mathbf{E}_{\mathbf{P}}[\,f(\omega) \mid \omega_{U^c}\,]\,\big] = \mathbf{E}_{\mu}\big[\,\mathbf{E}_{\mathbf{P}}[\,f(\omega) \mid \omega_{U^c}\,]\,\big]\,,
\end{aligned}$$

which means that $\mu$ is indeed Markov, with the same local specifications as $\mathbf{P}$. $\square$

We should emphasize that being non-extremal does not automatically mean non-canonical. For instance, as we will see later, the free Ising measure on the regular tree $\mathbb{T}_d$ is non-extremal for low temperatures, but is still a natural and interesting ergodic $\mathrm{Aut}(\mathbb{T}_d)$-invariant measure; the only strange thing about it is that, being non-extremal but ergodic, it must be a convex combination of non-translation-invariant Gibbs measures. (As opposed to $\mathbb{Z}^d$, where it turns out to be simply the average of the + and − measures.) Another instructive example is the following. It is easy to see that there is a unique $\mathrm{Aut}(\mathbb{T}_3)$-invariant measure on perfect matchings on $\mathbb{T}_3$. However, a perfect matching configuration outside a finite subset determines the configuration inside, which easily implies (proved carefully in [Lyo14]) that the measure cannot be tail trivial; rather, it has full tail. Hence, by Proposition 13.3, it cannot be extremal. Indeed, the deterministic behavior given the configuration in the complement of any finite subset implies that the atomic measure on any single configuration is a Markov random field, hence any non-constant measure on perfect matchings of $\mathbb{T}_3$ is in fact non-extremal.

... ... ...

## 13.3 Bootstrap percolation and zero temperature Glauber dynamics

{ss.bootstrap}

Bootstrap percolation on an arbitrary graph has a $\mathsf{Ber}(p)$ initial configuration, and a deterministic spreading rule with a fixed parameter $k$: if a vacant site has at least $k$ occupied neighbors at a certain time step, then it becomes occupied in the next step. **Complete occupation** is the event that every vertex becomes occupied during the process. The main problem is to determine the critical probability $p(G, k)$ for complete occupation: the infimum of the initial probabilities $p$ that make $\mathbf{P}_p[\text{complete occupation}] > 0$.

▷ **Exercise 13.9** ([vEnt87]). *Show that $p(\mathbb{Z}^2, 2) = 0$ and $p(\mathbb{Z}^2, 3) = 1$. (Hint for $k = 2$: show that a single large enough completely occupied box (a "seed") has a positive chance to occupy everything.)*

▷ **Exercise 13.10** ([Scho92]). * *Show that $p(\mathbb{Z}^d, k) = 0$ for $k \le d$ and $= 1$ for $k \ge d + 1$. (Hint: use the $d = 2$ idea, the threshold result Exercise 12.46, and induction on the dimension.)*

▷ **Exercise 13.11** ([BalPP06]). * {ex.BPTd}

**(a)** *Show that the 3-regular tree has $p(\mathbb{T}_3, 2) = 1/2$. More generally, show that for $2 \le k \le d$, $p(\mathbb{T}_{d+1}, k)$ is the supremum of all $p$ for which the equation*

$$\mathbf{P}\Big[\,\mathsf{Binom}(d, (1-x)(1-p)) \le d - k\,\Big] = x$$

*has a real root $x \in (0, 1)$.*

**(b)** *Deduce from part (a) that for any constant $\gamma \in [0, 1]$ and a sequence of integers $k_d$ with $\lim_{d\to\infty} k_d/d = \gamma$,*

$$\lim_{d\to\infty} p(\mathbb{T}_d, k_d) = \gamma.$$

The previous exercises show a clear difference between the behaviour of the critical probability on $\mathbb{Z}^d$ and $\mathbb{T}_d$: on $\mathbb{Z}^d$, once $k$ is small enough so that there are no local obstacles that obviously make $p(\mathbb{Z}^d, k) = 1$, we already have $p(\mathbb{Z}^d, k) = 0$. So, one can ask the usual question:

**Question 13.5.** *Is a group amenable if and only if for any finite generating set, the resulting $r$-regular Cayley graph has $p(G_r, k) \in \{0, 1\}$ for any $k$-neighbor rule?*

The answer is known to be affirmative for symmetric generating sets of $\mathbb{Z}^2$ on one hand [GraG96], and for any finitely generated non-amenable group that contains a free subgroup on two elements on the other [BalPP06].

▷ **Exercise 13.12.** *** *Find the truth for at least one more group (that is not a finite extension of $\mathbb{Z}^d$, of course).*

See [Scho92, BalPP06, BalP07, Holr07, BalBM10, BalBDCM10] and the references there for more on this model, both on infinite and finite graphs.

Bootstrap percolation results have often been applied to the study of the **zero temperature Glauber dynamics** of the Ising model. This Glauber dynamics was already defined in Section 13.1, but the zero temperature case can actually be described without mentioning the Ising model at all. Given a locally finite infinite graph $G(V,E)$ with an initial spin configuration $\omega_0 \in \{+,-\}^V$, the dynamics is that each site has an independent Poissonian clock, and if the clock of some site $x \in V$ rings at some time $t > 0$, then $\omega_t(x)$ becomes the majority of the spins of the neighbours of $x$, or, if there is an equal number of neighbours in each state, then the new state of $x$ is chosen uniformly at random. Now let $p_{\mathrm{fix}}(G)$ be the infimum of $p$ values for which this dynamics started from a $\mathsf{Ber}(p)$ initial configuration of "+" spins fixates at "+" (i.e., the spin of each site $x$ will be "+" from a finite time $T(x)$ onwards) almost surely.

From the symmetry of the two competing colours, it is clear that $p_{\mathrm{fix}}(G) \geq 1/2$. For what graphs is it equal to $1/2$? It is non-trivial to prove that $p_{\mathrm{fix}}(\mathbb{Z}) = 1$, see [Arr83].

**Question 13.6.** *Is it true that $p_{\mathrm{fix}}(\mathbb{Z}^d) = 1/2$ for all $d \geq 2$? And $p_{\mathrm{fix}}(\mathbb{T}_d) = 1/2$ for $d \geq 4$?*

Here it is what is known. Using very refined knowledge of bootstrap percolation, [Mor10] proved that $\lim_{d\to\infty} p_{\mathrm{fix}}(\mathbb{Z}^d) = 1/2$. On the other hand, $p_{\mathrm{fix}}(\mathbb{T}_3) > 1/2$ [How00]; the reason is that a density $1/2-\epsilon$ for the "$-$" phase is just barely subcritical for producing a bi-infinite path, while the "+" phase is just barely supercritical, so in a short time in the dynamics, bi-infinite "$-$" paths will form (somewhere in the huge non-amenable tree), making "+" fixation impossible.

▷ **Exercise 13.13.** * *Show that $\lim_{d\to\infty} p_{\mathrm{fix}}(\mathbb{T}_d) = 1/2$. (This follows from [CapM06], but here is a hint for a simpler proof, coming from Rob Morris. Exercise 13.15 (b) says that for $p > 1/2+\epsilon$, if $d$ is large enough, then the probability of everything becoming "$-$" in $\lceil d/2 \rceil$-neighbour "$-$"-bootstrap is zero. Prove that, moreover, the probability that an initially "+" site ever becomes "$-$" is tending to 0, as $d \to \infty$. This implies that the probability that a given vertex fixates at "+" is tending to 1. But then, the majority of the neighbours of any given vertex fixate at "+", fixing the state of that vertex, as well.)*

It is also interesting what happens at the critical density. By definition, non-fixation can happen in two ways: either some sites fixate at "+" while some other sites fixate at "$-$", or every site changes its state infinitely often. For $p = 1/2$ on $\mathbb{Z}^2$, it is known that every site changes its state infinitely often [NaNS00], while, for $p = 1/2$ on the hexagonal lattice, some sites fixate at "+" while all other sites fixate at "$-$" [HowN03]. The reason for the difference is the odd degree on the hexagonal lattice. Of course, these results do not imply that $p_{\mathrm{fix}} = 1/2$ on these graphs. What happens on $\mathbb{T}_d$ for $d \geq 4$ at $p = 1/2$ is not known, either.

## 13.4 Minimal Spanning Forests

{ss.MSF}

Our main references here will be [LyPS06] and [LyPer16].

While the Uniform Spanning Forests are related to random walks and harmonic functions, the Minimal Spanning Forests are related to percolation. The Minimal Spanning Tree ($\mathsf{MST}$) on a finite graph is constructed by taking i.i.d. $\mathsf{Unif}[0,1]$ labels $U(e)$ on the edges, then taking the spanning tree with the minimal sum of labels. Note that this is naturally coupled to $\mathsf{Ber}(p)$ bond percolation for all $p \in [0,1]$ at once.

▷ **Exercise 13.14.** *Give a finite graph on which $\mathsf{MST} \neq \mathsf{UST}$ with positive probability.*

For an infinite graph $G$, we again have two options: we can try to take the weak limit of the $\mathsf{MST}$ along any finite exhaustion, with free or wired boundary conditions. The limiting measures can be directly constructed. For any $e \in E(G)$, define

$$Z_{\mathsf{F}}(e) := \inf_{\gamma} \max\{U(f) : f \in \gamma\},$$

where the infimum is taken over paths $\gamma$ in $G \setminus \{e\}$ that connect the endpoints of $e$, and define

$$Z_{\mathsf{W}}(e) := \inf_{\gamma} \sup\{U(f) : f \in \gamma\},$$

where the infimum is taken over "generalized paths" $\gamma$ in $G \setminus \{e\}$ that connect the endpoints of $e$, i.e., $\gamma$ can also be a disjoint union of two half-infinite paths, one emanating from each endpoint of $e$. Then, the **Free** and **Wired Minimal Spanning Forests** are

$$\mathsf{FMSF} := \{e : U(e) \le Z_{\mathsf{F}}(e)\} \quad \text{and} \quad \mathsf{WMSF} := \{e : U(e) \le Z_{\mathsf{W}}(e)\}.$$

The connection between $\mathsf{WMSF}$ and critical percolation becomes clear through **invasion percolation**. For a vertex $v$ and the labels $\{U(e)\}$, let $T_0 = \{v\}$, then, inductively, given $T_n$, let $T_{n+1} = T_n \cup \{e_{n+1}\}$, where $e_{n+1}$ is the edge in $\partial_E T_n$ with the smallest label $U$. The **Invasion Tree** of $v$ is then $\mathsf{IT}(v) := \bigcup_{n \ge 0} T_n$. We now have the following deterministic result:

{ex.IT}

▷ **Exercise 13.15.** *Prove that if $U : E(G) \longrightarrow \mathbb{R}$ is an injective labelling of a locally finite graph, then* $\mathsf{WMSF} = \bigcup_{v \in V(G)} \mathsf{IT}(v)$.

Once the invasion tree enters an infinite $p$-percolation cluster $\mathscr{C} \subseteq \omega_p := \{e : U(e) \le p\}$, it will not use edges outside it. Furthermore, it is not surprising (though non-trivial to prove, see [HäPS99]) that for any transitive graph $G$ and any $p > p_c(G)$, the invasion tree eventually enters an infinite $p$-cluster. Therefore, $\limsup\{U(e) : e \in \mathsf{IT}(v)\} = p_c$ for any $v \in V(G)$. This already suggests that invasion percolation is a "self-organized criticality" version of critical percolation.

{ex.ITsparse}

▷ **Exercise 13.16.**** *Show that for $G$ transitive amenable, $\theta(p_c) = 0$ is equivalent to $\mathsf{IT}(v)$ having density zero, measured along any Følner exhaustion of $G$.*

Non-percolation at $p_c$ also has an interpretation as the smallness of the $\mathsf{WMSF}$ trees, which we state without a proof:

{t.WMSFoneend}

**Theorem 13.7** ([LyPS06])**.** *On a transitive unimodular graph, $\theta(p_c) = 0$ implies that a.s. each tree of* $\mathsf{WMSF}$ *has one end.*

By Benjamini-Lyons-Peres-Schramm's Theorem 12.11, in the non-amenable case this gives that each tree of $\mathsf{WMSF}$ has one end. For the amenable case, the following two exercises, combined with the fact (obvious from Exercise 13.19) that all the trees of $\mathsf{WMSF}$ are infinite, almost give this "one end" result:

{ex.deg2}

▷ **Exercise 13.17.** *Show that for any invariant spanning forest $\mathfrak{F}$ on a transitive amenable $G$, the expected degree is at most 2. Moreover, if all trees of $\mathfrak{F}$ are infinite a.s., then the expected degree is exactly 2.*

{ex.end2}

▷ **Exercise 13.18.** *Show that for any invariant spanning forest $\mathfrak{F}$ on a transitive amenable $G$, it is not possible that a.s. all trees of $\mathfrak{F}$ have at least 3 ends. Moreover, if all the trees are infinite, then each has 1 or 2 ends. (Hint: use the previous exercise.)*

{ex.end1}

▷ **Exercise 13.19.** *** *Show that all the trees in the* WMSF *on any transitive graph have one end almost surely.*

On the other hand, FMSF is more related to percolation at $p_u$. On a transitive unimodular graph, each tree in it can intersect at most one infinite cluster of $p_u$-percolation in the standard coupling, moreover, if $p_u > p_c$, then each tree intersects exactly one infinite $p_u$-cluster. Furthermore, adding an independent $\mathsf{Ber}(\epsilon)$ bond percolation to FMSF makes it connected. Finally, we have the following:

{t.MSFpcpu}

**Theorem 13.8** ([LyPS06])**.** *On any connected graph $G$, we have a.s. at most one infinite cluster for almost all $p \in [0,1]$ if and only if* FMSF = WMSF *a.s. In particular, for transitive unimodular graphs, $p_c = p_u$ is equivalent to* FMSF = WMSF *a.s.*

*Proof.* Since $G$ is countable, WMSF $\subsetneq$ FMSF is equivalent to having some $e \in E(G)$ with the property that $\mathbf{P}[\, Z_{\mathsf{W}}(e) < U(e) \le Z_{\mathsf{F}}(e)\,] > 0$. This probability equals to the probability of the event $A(e)$ that the two endpoints of $e$ are in different infinite clusters of $U(e)$-percolation on $G \setminus \{e\}$. If this is positive, then, by the independence of $U(e)$ from other edges, there is a positive Lebesgue measure set of possibilities for this $U(e)$, and for these percolation parameters we clearly have at least two infinite clusters with positive probability. Conversely, if there is a positive measure set of $p$ values for which there are more than one infinite $p$-clusters with positive probability, then, by insertion tolerance, there is an edge $e$ whose endpoints are in different infinite $p$-clusters with positive probability, and hence, by the independence of $U(e)$ from other edges, there is a positive probability for $A(e)$. □

There are many open questions here; see [LyPS06] or [LyPer16]. For instance, must the number of trees in the FMSF and the WMSF in a transitive graph be either 1 or $\infty$ a.s.? In which $\mathbb{Z}^d$ is the MSF a single tree? The answer is yes for $d = 2$, using planarity. On the other hand, for $d \ge 19$, where the lace expansion shows $\theta(p_c) = 0$ and many other results, one definitely expects infinitely many trees. Regarding the critical dimension, where the change from one to infinity happens, some contradictory conjectures have been made: $d = 6$ [JaR09] and $d = 8$ [NewS96]. It might also be that both answers are right in some sense: for $\mathbb{Z}^d$, the critical dimension might be 8, but for the scaling limit of the MSF, which is a spanning tree of $\mathbb{R}^d$ in a well-defined sense (very roughly, for any finite collection of points there is a tree connecting them, with some natural compatibility relations between the different trees), the critical dimension might be 6 [AiBNW99]: in $d = 7$, say, there could be quite long connections between some points of $\mathbb{Z}^d$, escaping to infinity in the scaling limit.

Finally, on the triangular grid, the scaling limit of a version of the MST (adapted to site percolation) is known to exist, is rotationally and scale invariant, but is conjectured not to be conformally invariant [GarPS10b] — a behaviour that goes against the physicists' rule of thumb about conformal invariance.

## 13.5 Measurable group theory and orbit equivalence

{ss.MeGrTh}

Consider a (right) action $x \mapsto g(x) = x^g$ of a discrete group $\Gamma$ on some probability space $(X, \mathcal{B}, \mu)$ by measure-preserving transformations. We will usually assume that the action is ergodic (i.e., if $U \subseteq X$ satisfies $g(U) = U$ for all $g \in \Gamma$, then $\mu(U) \in \{0,1\}$) and essentially free (i.e., $\mu\{x \in X : g(x) = x\} = 0$ for any $g \ne 1 \in \Gamma$). These conditions are satisfied for the natural translation action on bond or site percolation configurations under an ergodic probability measure, say $\mathsf{Ber}(p)$ percolation: $\omega^g(h) := \omega(gh)$, as we did in the second proof of Corollary 12.20. This action of a group $\Gamma$ on $\{0,1\}^\Gamma$, or $S^\Gamma$ with a countable $S$, or $[0,1]^\Gamma$, with the product of $\mathsf{Ber}(p)$ or $\{p_s\}_{s \in S}$ or $\mathsf{Leb}[0,1]$ measures, is usually called a **Bernoulli shift**.

The obvious notion for probability measure preserving (abbreviated p.m.p.) actions of some groups $\Gamma_i$ $(i = 1, 2)$ on some $(X_i, \mathcal{B}_i, \mu_i)$ being the same is that there is a group isomorphism $\iota : \Gamma_1 \longrightarrow \Gamma_2$ and a measure-preserving map $\varphi : X_1 \longrightarrow X_2$ such that $\varphi(x^g) = \varphi(x)^{\iota(g)}$ for almost all $x \in X_1$. There is a famous theorem of Ornstein and Weiss [OW87] that two Bernoulli shifts of a given f.g. amenable group (in most cases) are equivalent in this sense iff the entropies (generalizing $-\sum_{s\in S} p_s \log(p_s)$ from the case of $\mathbb{Z}$-actions suitably) are equal. We will consider here a cruder equivalence relation, which is the natural measure-theoretical analogue of the virtual isomorphism of groups, exactly as quasi-isometry was the geometric analogue — see Exercises 3.7 and 3.8.

{d.measequiv}

**Definition 13.9.** *Two f.g. groups, $\Gamma_1$ and $\Gamma_2$ are* **measure equivalent** *if they admit commuting (not necessarily probability) measure preserving essentially free actions on some measure space $(X, \mathcal{B}, \mu)$, each with a positive finite measure fundamental domain.*

Another natural notion is the following:

{d.orbiteq}

**Definition 13.10.** *Two p.m.p. actions, $\Gamma_i$ acting on $(X_i, \mathcal{B}_i, \mu_i)$ for $i = 1, 2$, are called* **orbit equivalent** *if there is a measure-preserving map $\varphi : X_1 \longrightarrow X_2$ such that $\varphi(x^{\Gamma_1}) = \varphi(x)^{\Gamma_2}$ for almost all $x \in X_1$. (Note that this is indeed an equivalence relation.)*

*A small relaxation is that the actions are* **stably orbit equivalent***: for each $i = 1, 2$ there exists a Borel subset $Y_i \subseteq X_i$ that meets each orbit of $\Gamma_i$ and there is a measure-scaling isomorphism $\varphi : Y_1 \longrightarrow Y_2$ such that $\varphi(x^{\Gamma_1} \cap Y_1) = \varphi(x)^{\Gamma_2} \cap Y_2$ for a.e. $x \in X_1$.*

A good reason for looking at this notion is the theorem of Feldman and Moore [FeM75] that every countable Borel equivalence relation on a standard Borel space can be realized as the orbit equivalence relation of a Borel action of some countable group. Another good reason is that two f.g. groups are measure equivalent in the sense of Definition 13.14 iff they admit stably orbit equivalent actions. Or, almost the same claim: they are measure equivalent in such a stronger way that the fundamental domains are actually the same iff they are orbit equivalent. The proofs, similar to Exercise 3.7, can be found in [Gab02], which is a great introduction to orbit equivalence.

Of course, if the two actions are equivalent in the usual sense, then they are also orbit equivalent. But this notion is much more flexible, as shown, for instance, by the following actions of $\mathbb{Z}$ and $\mathbb{Z}^2$.

Consider the set $X = \{0, 1\}^{\mathbb{N}}$ of infinite binary sequences, with the $\mathsf{Ber}(1/2)$ product measure. The **adding machine** action of $\mathbb{Z}$ on $X$ is defined by the following recursive rule: for the generator $a$ of $\mathbb{Z}$, and any $w \in X$,

$$\begin{aligned} (0w)^a &= 1w \\ (1w)^a &= 0w^a . \end{aligned} \tag{13.13}$$ {e.adding}

Note that this definition can also be made for finite sequences $w$, and the actions on the starting finite segments of an infinite word are compatible with each other, hence the definition indeed makes sense for infinite words. For a finite word $w = w_0 w_1 \dots w_k$, if we write $\beta(w) := \sum_{i=0}^{k} w_i 2^i$, then this action has the interpretation of adding 1 in binary expansion, $\beta(w^a) = \beta(w) + 1$, hence the name. (One thing to be careful about is that for a finite sequence $w$ of all 1's this $\beta$-interpretation breaks down: one needs to add at least one zero at the end of $w$ to get it right.) The action of $\mathbb{Z}$ on $X$ is clearly measure-preserving.

We can similarly define an action of $\mathbb{Z}^2$ on $X \times X = \{0, 1\}^{\mathbb{N}\times\mathbb{N}}$, simply doing the $\mathbb{Z}$-action coordinate-wise. Now, these two actions are orbit-equivalent. Why? The orbit of a word $w \in X$ by the $\mathbb{Z}$ action is the set of all words with the same tail as $w$. (Note that we can apply $a$ or $a^{-1}$ only finitely many times.) Similarly, the orbit of some $(w, w') \in X \times X$ is the set of all pairs with each coordinate having the correct tail. Now, the interlacing map $\varphi : X \times X \longrightarrow X$ defined by $\varphi(w_0 w_1 \dots, w'_0 w'_1 \dots) = w_0 w'_0 w_1 w'_1 \dots$ is clearly measure-preserving and establishes an orbit-equivalence.

An extreme generalization of the previous example is another famous result of Ornstein and Weiss from 1980, using some of the machinery of the entropy/isomorphism theorem, see [KecM04]: any two ergodic free p.m.p. actions of amenable groups are orbit equivalent to each other. A closely related probabilistic statement is the following:

{ex.amenZ}

▷ **Exercise 13.20.** *Show that a Cayley graph $G(\Gamma, S)$ is amenable iff it has a $\Gamma$-invariant random spanning $\mathbb{Z}$ subgraph. (Hint: for one direction, produce the invariant $\mathbb{Z}$ using coarser and coarser "quasi-tilings" that come from Exercise 12.34; for the other direction, produce an invariant mean from the invariant $\mathbb{Z}$.)*

On the other hand, as a combination of the work of several people, see [Gab10, Section 10], a recent result is that any non-amenable group has continuum many orbit-inequivalent actions. A key ingredient is the following theorem, which has a percolation-theoretical proof, to be discussed in a future version of these notes:

{t.GaboLyo}

**Theorem 13.11** ([GabL09]). *For any non-amenable countable group $\Gamma$, the orbit equivalence relation of the Bernoulli shift action on $([0,1], \mathsf{Leb})^\Gamma$ contains a subrelation generated by a free ergodic p.m.p. action of the free group $F_2$. In other words, the orbits of the Bernoulli shift can be decomposed into the orbits of an ergodic $F_2$ action. Or more probabilistically, there is an invariant spanning forest of 4-regular trees on $\Gamma$ that is a factor of i.i.d. $\mathsf{Unif}[0,1]$ labels on the vertices $\Gamma$, and the trees are indistinguishable.*

A nice interpretation of this result is that any non-amenable group has an $F_2$ **randosubgroup**, in the following sense. Consider the set $S_{G,H} := \{f : G \longrightarrow H, f(1) = 1\}$. The group $\Gamma$ acts on this set by

$$f^g(t) := f(g)^{-1} f(gt)\,; \tag{13.14}$$ {e.SGH^G}

it is easy to check that this is indeed an action from the right. A group-homomorphism is just a fixed point of this $\Gamma$-action. A **randomorphism** is a $\Gamma$-invariant probability distribution on $S_{G,H}$. So, $\Gamma$ is a randosubgroup of $H$ if there is a $\Gamma$-invariant distribution on injective maps in $S_{G,H}$. (It cannot be called a "random subgroup", since it is not a distribution on actual subgroups.)

▷ **Exercise 13.21.** *A randosubgroup of an amenable group is also amenable.*

Now, what is the connection to orbit equivalence? If the orbit equivalence relation on $X$ generated by $\Gamma$ is a subrelation of the one on $Y$ generated by $H$, i.e., there is a p.m.p. map $\varphi : X \longrightarrow Y$ such that $\varphi(x^\Gamma) \subseteq \varphi(x)^H$ for almost all $x \in X$, then for a.e. $x \in X$ and every $g \in \Gamma$ there is an $\alpha(x,g) \in H$ such that $\varphi(x^g) = \varphi(x)^{\alpha(x,g)}$. Moreover, if the $H$-action is free, then this $\alpha(x,g)$ is uniquely determined, and it satisfies the so-called **cocycle equation**

$$\alpha(x, gh) = \alpha(x,g)\,\alpha(x^g, h)\,. \tag{13.15}$$ {e.cocycle}

By writing $\alpha_x(g) = \alpha(x,g)$, for a.e. $x \in X$ we get a map $\alpha_x \in S_{G,H}$. Now, what is the action of $\Gamma$ on such elements of $S_{G,H}$? By (13.17), $\alpha_x^g(t) = \alpha_x(g)^{-1}\alpha_x(gt)$, which, by (13.18), is $\alpha_x(g)^{-1}\alpha_x(g)\,\alpha_{x^g}(t) = \alpha_{x^g}(t)$. So, we have a $\Gamma$-equivariant action on the set $\{\alpha_x : x \in X\}$, and if we take a random point $x \in X$ w.r.t. the $\Gamma$-invariant probability measure $\mu$, then we get a $\Gamma$-invariant measure on $\{\alpha_x : x \in X\}$, i.e., a randoembedding of $\Gamma$ into $H$.

Given that all amenable groups are measure equivalent, in order to distinguish non-amenable groups from each other, one seem to need rather non-trivial invariants. The $\ell^2$-Betti numbers and the cost of groups mentioned in Chapter 11 are such examples. A future version of these notes will hopefully discuss them in a bit more detail, but see [Gab02, Gab10] for now. Well, let's try.

Given a probability space $(X, \mathcal{B}, \mu)$, a **graphing** on $X$ is simply a measurable oriented graph: a countable set of "edges" $\Phi = \{\varphi_i : A_i \longrightarrow B_i\}_{i \in I}$, which are measure-preserving isomorphisms between measurable subsets $A_i$ and $B_i$. The **cost of a graphing**, which we can also think of as the average out-degree of a random vertex in $X$, is

$$\mathrm{cost}(\Phi) := \sum_{i \in I} \mu(A_i) = \int_X \sum_{i \in I} \mathbf{1}_{A_i}(x)\, d\mu(x)\,.$$

Any graphing $\Phi$ generates a measurable equivalence relation $\mathcal{R}_\Phi$ on $X$: the equivalence classes are the connected components of $\overline{\Phi}$, which is the graph obtained from $\Phi$ by forgetting the orientations. The **cost of an equivalence relation** $\mathcal{R} \subseteq X \times X$ is

$$\mathrm{cost}(\mathcal{R}) := \inf\{\mathrm{cost}(\Phi) : \Phi \text{ generates } \mathcal{R}\}\,.$$

A usual way of obtaining a graphing is to consider the "Schreier graph" of a p.m.p. action of a group $\Gamma$ on $X$, with $\Phi = \{\varphi_i : X \longrightarrow X\}_{i \in I}$ given by a generating set $\{\gamma_i\}_{i \in I}$ of $\Gamma$. Then the equivalence relation generated by $\Phi$ is the orbit equivalence relation of the action. Now, the **cost of a group** $\Gamma$ is defined as

$$\mathrm{cost}(\Gamma) := \inf\big\{\mathrm{cost}(\mathcal{R}) : \mathcal{R} \text{ is the orbit equiv. rel. of some free p.m.p. action } \Gamma \curvearrowright X\big\}\,. \tag{13.16}$$ {e.cost1}

A group is said to have a **fixed price** if the orbit equivalence relations of all free p.m.p. actions have the same cost, and it is not known whether all groups have a fixed price.

Here is a more tangible definition of the cost of a group $\Gamma$ for probabilists:

$$\mathrm{cost}(\Gamma) = \frac{1}{2} \inf\big\{\mathbf{E}_\mu[\deg(o)] : \mu \text{ is the law of a } \Gamma\text{-invariant random spanning graph on } \Gamma\big\}\,. \tag{13.17}$$ {e.cost2}

How is this the same cost as before? A $\Gamma$-invariant random graph is a probability measure $\mu$ on $\Omega = \{0,1\}^{\Gamma \times \Gamma}$ that is concentrated on symmetric functions on $\Gamma \times \Gamma$ and is invariant under the diagonal action of $\Gamma$. A corresponding graphing (which may be called the **cluster graphing**) is the following. Fix an element $o \in \Gamma$, then let $\omega, \eta \in \Omega$ be connected by $\gamma \in \Gamma$ if $\omega^\gamma = \eta$ and the edge from $o$ to $\gamma o$ is open in the graph $\omega$. The domain $A_\gamma$ of this measurable edge consists of those $\omega \in \Omega$ in which the edge $(o, \gamma o)$ is open, hence the cost of this graphing (measured in $\mu$) is the $\mu$-expected out-degree of $o$, or one half of the expected total degree. (Note here that $\gamma^{-1}$ is an edge from $\eta$ to $\omega$, since $\omega = \eta^{\gamma^{-1}}$ and $\omega(o, \gamma o) = \eta(\gamma^{-1} o, o) = \eta(o, \gamma^{-1} o)$.) This is a sub-graphing of the **full graphing**, in which $\omega, \eta$ are connected by $\gamma$ iff $\omega^\gamma = \eta$, and which is just the Schreier graphing of the action of $\Gamma$ on $\Omega$. Clearly, the cluster graphing $\mu$-almost surely generates the orbit equivalence relation of the action iff $\mu$-almost surely the graph spans the entire $\Gamma$. Thus, (13.20) is indeed the infimum of some costs; however, we considered here only some special p.m.p. actions $\Gamma \curvearrowright (\Omega, \mu)$, and only some special graphings generating the corresponding orbit equivalence relations. And we do not even get that (13.20) is at least as big as (13.19), because some of these actions on $(\Omega, \mu)$ might not be essentially free. But all these issues can be easily solved:

If $\Gamma \curvearrowright (X, \mu)$ is a free p.m.p. action, and $\Phi$ is a graphing generating its orbit equivalence relation, then for each $x \in X$ we can consider $\omega_x \in \Omega = \{0,1\}^{\Gamma \times \Gamma}$ given by

$$(g, h) \text{ is an open edge in } \omega_x \Leftrightarrow (x^g, x^h) \text{ is an edge in } \Phi\,.$$

Then the pushforward $\mu_\Phi$ of $\mu$ under this $x \mapsto \omega_x$ gives an invariant spanning graph on $\Gamma$, and $\mathbf{E}_{\mu_\Phi}[\deg o]/2 = \mathrm{cost}(\Phi)$. This shows that (13.20) is not larger than (13.19). For the other direction, given an invariant random spanning graph $\mu$ on $\Omega$, we can produce a free p.m.p. action by taking a direct product of $(\Omega, \mu)$ with a fixed free action (say, a Bernoulli shift) $(Y, \nu)$, and considering the natural extension of the cluster graphing that we had before, to $\Omega \times Y$. For the details, see [KecM04, Proposition 29.5].

The notion of cost resembles the **rank** (i.e., the minimal number of generators) $d(\Gamma)$ of a group. Here is an explicit formulation of this idea, which has several nice applications, for hyperbolic groups etc; see [AbN07].

Let $\Gamma = \Gamma_0 \geq \Gamma_1 \geq \dots$ be a chain of finite index subgroups. Corresponding to such a subgroup sequence, the **right coset tree** $\mathcal{T}$ has the root $\Gamma = \Gamma_0$, and a coset $\Gamma_{n+1}y$ is a child of $\Gamma_n x$ if $\Gamma_{n+1}y \subset \Gamma_n x$. The number of children of $\Gamma_n x$ is $[\Gamma_n : \Gamma_{n+1}]$. The set of rays $\Gamma = \Gamma_0 x_0 \supset \Gamma_1 x_1 \supset \Gamma_2 x_2 \supset \dots$ in $\mathcal{T}$ is the boundary $\partial\mathcal{T}$ of the tree, equipped with the usual metrizable topology. If we have normal subgroups, $\Gamma_n \triangleleft \Gamma$ $\forall n$, then $\partial\mathcal{T}$ can be equipped with a group structure, and is called the profinite completion of $\Gamma$ with respect to the series $\{\Gamma_n\}_{n\geq 0}$; see e.g., [Wil98].

One often assumes the so-called **Farber condition** on the sequence $\{\Gamma_n\}_{n\geq 0}$: the natural action of $\Gamma$ on the boundary $\partial T(\Gamma, \{\Gamma_n\})$ of the coset tree, with the natural Borel probability measure, is essentially free.

{ex.Farber}

▷ **Exercise 13.22.** *The Farber condition holds if each $\Gamma_n$ is normal in $\Gamma$, and $\bigcap_{n\geq 1} \Gamma_n = \{1\}$.*

The **rank gradient** of a subgroup chain is defined by

$$\mathrm{RG}(\Gamma, \{\Gamma_n\}) := \lim_{n\to\infty} \frac{d(\Gamma_n) - 1}{|\Gamma : \Gamma_n|} \,. \tag{13.18}$$ {e.rankgrad}

The following exercise shows that this is a good definition. To start with, recall the Schreier index formula, Theorem 2.18: for any $H \leq F_k$ of finite index, one has $H \simeq F_\ell$, hence $d(F_k) = k$ and $d(H) = \ell$, and $d(H) - 1 = \big(d(F_k) - 1\big)\,[F_k : H]$.

▷ **Exercise 13.23.**

**(a)** *Show that for any finitely generated groups $H \leq G$ with $[G : H] < \infty$, one has*

$$d(H) - 1 \leq \big(d(G) - 1\big)\,[G : H] \,.$$

**(b)** *Conclude that the limit in the definition of* RG *always exists.*

**(c)** *Conclude that, for the free group on $k$ generators,* $\mathrm{RG}(F_k, \{\Gamma_n\}) = k - 1$*, regardless of $\Gamma_n$.*

**Theorem 13.12** ([AbN07])**.** *Let $\mathcal{R}$ denote the orbit equivalence relation of $\Gamma \curvearrowright \partial T(\Gamma, \{\Gamma_n\})$. Then $\mathrm{RG}(\Gamma, \{\Gamma_n\}) = \mathrm{cost}(\mathcal{R}) - 1$.*

We can also obtain a **random rooted graph** from a graphing: pick a random root $x \in X$ according to $\mu$, and take its connected component $\overline{\Phi}(x)$ in $\overline{\Phi}$. This random rooted graph will be **unimodular** (with a definition more general than what we gave before, which applies to non-regular graphs): by the $\varphi_i$'s in $\Phi$ being measure-preserving, the equivalence relation $\mathcal{R}_\Phi$ is also **measure-preserving**, i.e., for any measurable $F : X \times X \longrightarrow \mathbb{R}$, we have

$$\int_X \sum_{y \in \mathcal{R}_\Phi[x]} F(x, y)\, d\mu(x) = \int_X \sum_{y \in \mathcal{R}_\Phi[x]} F(y, x)\, d\mu(x) \,,$$

where $\mathcal{R}_\Phi[x] = \overline{\Phi}(x)$ is the equivalence class or connected component of $x$. Now, this can be taken as the definition of the Mass Transport Principle hence unimodularity for the random rooted graph $\overline{\Phi}(x)$. See also Definition 14.1 and the MTP (14.1) in Section 14.1.

# 14 Local approximations to Cayley graphs

{s.local}

For many probabilistic models, it is easy to think that understanding the model in a large box of $\mathbb{Z}^d$ is basically the same as understanding it on the infinite lattice. Well, sometimes the finite problem is actually harder (for instance, compare Conjecture 12.25 with Lemma 12.5), but they are certainly closely related. Why exactly is this so? In what sense do the boxes $[n]^d$ converge to $\mathbb{Z}^d$?

## 14.1 Unimodularity and soficity

{ss.sofic}

A sequence of finite graphs $G_n$ is said to converge to a transitive graph $G$ in the **Benjamini-Schramm sense** [BenS01] (also called **local weak convergence** [AldS04]) if for any $\epsilon > 0$ and $r \in \mathbb{N}_+$ there is an $n_0(\epsilon, r)$ such that for all $n > n_0$, at least a $(1-\epsilon)$-proportion of the vertices of $G_n$ have an $r$-neighbourhood isomorphic to the $r$-ball of $G$. We will sometimes abbreviate this as **BSch-convergence**.

For instance, the cubes $\{1, \dots, n\}^d$ converge to $\mathbb{Z}^d$. On the other hand, if we take the balls $B_n(o)$ in the $d$-regular tree $\mathbb{T}_d$, then the proportion of leaves in $B_n(o)$ converges to $(d-2)/(d-1)$ as $n \to \infty$, and more generally, the proportion of vertices at distance $k \in \mathbb{N}$ from the set of leaves (i.e., on the $(n-k)$th level $L_{n-k}$) converges to $p_{-k} := (d-2)/(d-1)^{k+1}$. And, for a vertex in $L_{n-k}$, the sequence of its $r$-neighbourhoods in $B_n(o)$ (for $r = 1, 2, \dots$) is not at all the same as in an infinite regular tree, and depends on the value of $k$. Therefore, the limit of the balls $B_n(o)$ is certainly not $\mathbb{T}_d$, or any other transitive graph. More generally, we have the following exercise:

{ex.amensofic}

▷ **Exercise 14.1.** *Show that a transitive graph $G$ has a sequence $G_n$ of* subgraphs *converging to it in the local weak sense iff it is amenable.*

The sequence of balls in a regular tree does not converge to any transitive graph, but there is still a meaningful limit structure, a **random rooted graph**. Namely, generalizing our previous definition, we say that a sequence of finite graphs $G_n$ converges in the local weak sense to a probability distribution on rooted bounded degree graphs $(G, \rho)$, where $\rho \in V(G)$ is the root, if for any $r \in \mathbb{N}$, taking a uniform random root $\rho_n \in V(G_n)$, the distribution we get on the $r$-neighbourhoods around $\rho_n$ in $G_n$ converges to the distribution of $r$-neighbourhoods around $\rho$ in $G$. There is a further obvious generalization, for graphs whose edges and/or vertices are labelled by elements of some finite set, or more generally, of some compact metric space: two such labelled rooted graphs are close if the graphs agree in a large neighbourhood of the root and all the corresponding labels are close to each other. These structures are usually called **random rooted networks**. So, for instance, we can talk about the local weak convergence of edge-labeled digraphs (i.e., directed graphs) to Cayley diagrams of infinite groups (see Definition 2.6).

Continuing the previous example, the balls in $\mathbb{T}_d$ converge to the random rooted graph depicted on Figure 14.1, which is a fixed infinite tree $T_d^*$ (often called the canopy tree) with infinitely many leaves (denoted by level $L_0$) and one end, together with a random root that is on level $L_{-k}$ with probability $p_{-k} = (d-2)/(d-1)^{k+1}$ — the weights that we computed above. It does not matter how this probability is distributed among the vertices of the level; say, all of it could be given to a single vertex. The reason for this is that the vertices on a given level lie in a single orbit of $\mathrm{Aut}(T_d^*)$, and our notion of convergence looks only at isomorphism classes of $r$-neighbourhoods. So, in fact, the right abstract definition for our limiting "random rooted graph" is a Borel probability measure on rooted isomorphism classes of locally finite rooted graphs, denoted by $\mathcal{G}_*$, equipped with the obvious "local" topology.

It is clear how the probabilities $p_{-k}$ for the root $\rho$ in $T_d^*$ arise from the sequence of finite balls in $\mathbb{T}_d$. However, there is also a tempting interpretation in terms of $T_d^*$ itself: if $\rho$ was chosen "uniformly at random

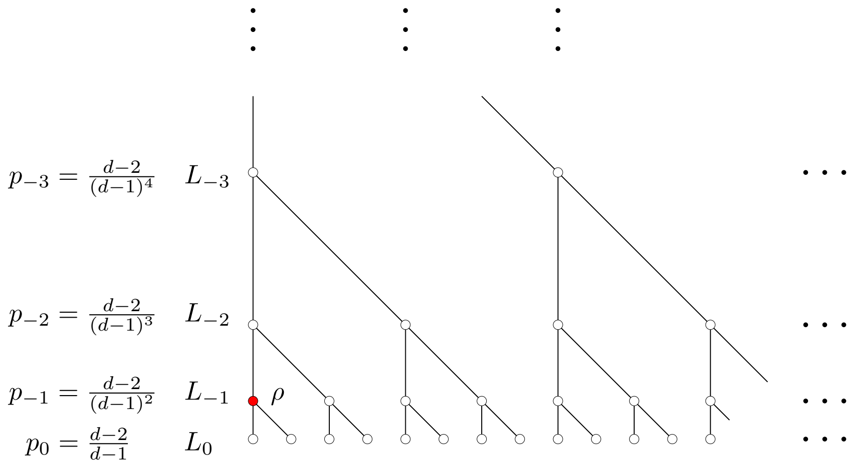



Figure 14.1: The "canopy tree" $T_d^*$ with a random root $\rho$ (now on level $L_{-1}$), which is the local weak limit of the balls in the $d$-regular tree $\mathbb{T}_d$, for $d = 3$. {f.2ary1end}

among all vertices of $T_d^*$", then it should have probability $p_{-k}$ to be on level $L_{-k}$, since "evidently" there are $p_{-k}/p_{-(k+1)} = d - 1$ times "more" vertices on level $L_{-k}$ than on $L_{-(k+1)}$. Now, even though with the counting measure on the vertices this argument does not make sense, there are ways to make it work: one should be reminded of the definition of unimodularity in Section 12.2, just after Theorem 12.11. In fact, one can make the following more general definition: {d.urn}

**Definition 14.1** (Unimodular random networks)**.**

**(a)** *A measure $\mu$ on the space $\mathcal{G}_*$ of rooted networks is called* **unimodular** *if the* **Mass Transport Principle** *holds in the following form. Let $\mathcal{G}_{**}$ be the set of all double-rooted isomorphism-classes of triples $(G, x, y)$, where $G$ is a locally finite graph and $x, y \in V(G)$, equipped with the natural local topology. For any Borel-measurable function $F$ on $\mathcal{G}_{**}$, we require that*

$$\int \sum_{x \in V(G)} F(G, \rho, x)\, d\mu(G, \rho) = \int \sum_{y \in V(G)} F(G, y, \rho)\, d\mu(G, \rho)\,. \tag{14.1}$$ {e.MTPgen}

**(b)** *Given a $d$-regular random rooted graph (or network) $(G, \rho)$ sampled from a measure $\mu$, choose a neighbour of $\rho$ uniformly at random, call it $\rho'$, and consider the joint distribution of $(G, \rho, \rho')$ on $\mathcal{G}_{**}$. Now "take the step to $\rho'$ and look back", i.e., take $(G, \rho', \rho)$, which is again a random rooted graph, with root $\rho'$, plus a neighbour $\rho$. If the two laws are the same, then the random rooted graph $(G, \rho)$, or rather the measure $\mu$, is called* **unimodular**.

**(c)** *For non-regular random graphs, if the degrees are $\mu$-a.s. bounded by some $d$, one can add $d - \deg(x)$ "half-loops" to each vertex $x$ to make the graph $d$-regular. In other words, in the above definition of the step from $\rho$ to $\rho'$, consider the* **delayed random walk** *that stays put with probability $(d - \deg(x))/d$. (This is a natural definition for random subgraphs of a given transitive graph.)*

**(d)** *If the degrees are not bounded, one still can take the following limiting procedure: take a large $d$, add the half-loops to each vertex with degree less than $d$ to get $G_d$, and then require that the total variation distance between $(G_d, \rho, \rho')$ and $(G_d, \rho', \rho)$, divided by $\mathbf{P}[\,\rho' \neq \rho\,]$ (with the randomness given by both* *sampling $(G, \rho)$ and then making the SRW step; we do this to make up for the laziness we introduced) tends to 0 as $d \to \infty$.*

*Another possibility in the unbounded case is to truncate $(G, \rho)$ in any reasonable way to have maximal degree $d$, and require that the resulting random graph be unimodular, for any $d$. One reasonable truncation is to remove uniformly at random the excess number of incident edges for each vertex with a degree larger than $d$. (Some edges might get deleted twice, which is fine, and we might get several components, which is also fine.)*

**(e)** *For the case $\mathbf{E}_\mu \deg(\rho) < \infty$, there is an alternative to using the delayed SRW. Let $\hat{\mu}$ be the probability measure $\mu$ on $(G, \rho)$* **size-biased** *by $\deg(\rho)$, i.e., with Radon-Nikodym derivative $\frac{d\hat{\mu}}{d\mu}(G, \rho) = \deg(\rho)/\mathbf{E}_\mu \deg(\rho)$. Now, if we sample $(G, \rho)$ from $\hat{\mu}$, and then take a non-delayed SRW step to a neighbor $\rho'$, then $(G, \rho', \rho)$ is required to have the same distribution as $(G, \rho, \rho')$.*

It is easy to see that if we take a transitive graph $G$, then it will be unimodular in our old sense ($|\Gamma_x y| = |\Gamma_y x|$ for all $x \sim y$, where $\Gamma = \mathrm{Aut}(G)$) if and only if the Dirac measure on $(G, o)$, with an arbitrary $o \in V(G)$, is unimodular in the sense of Definition 14.1. Indeed, for a double-rooted equivalence class $(G, x, y)$, with $(x, y) \in E(G)$, we will have $\mathbf{P}\big[\, (G, \rho, \rho') \simeq (G, x, y) \,\big] = |\Gamma_x y|/d$, where $d$ is the degree of the graph, while $\mathbf{P}\big[\, (G, \rho', \rho) \simeq (G, x, y) \,\big] = |\Gamma_y x|/d$, and these are equal for all pairs $(x, y) \in E(G)$ iff $G$ is unimodular.

We also get back our old Mass Transport Principle (12.9) for transitive graphs. In some sense, the transitivity of $G$ and the diagonal invariance of $f$ is now replaced by considering rooted networks and functions on double-rooted isomorphism-classes. Namely, if $f$ is a diagonally invariant random function on a transitive graph $G$, then, in (14.1), we can take $\mu$ to be a Dirac mass on $(G, o)$, with an arbitrary $o \in V(G)$, and $F(G, x, y) := \mathbf{E} f(x, y)$.

The simplest possible example of a unimodular random rooted graph is any fixed finite graph $G$ with a uniformly chosen root $\rho$. Here, the Mass Transport Principle (14.1) holds obviously, and part (e) of Definition 14.1 is also very transparent: if we size-bias the root $\rho$ by its degree, then the edge $(\rho, \rho')$ given by non-delayed SRW is simply a uniform random element of $E(G)$, obviously invariant under $\rho \leftrightarrow \rho'$.

The example of finite graphs is in fact a crucial one. It is easy to see that the Benjamini-Schramm limit of finite graphs is still a unimodular random rooted graph. For instance, the rooted tree $(T_d^*, \rho)$ of Figure 14.1 was a BSch-limit of finite graphs.

{ex.limisunimo

▷ **Exercise 14.2.** *If $G$ is a transitive graph with a sequence of finite $G_n$ converging to it in the Benjamini-Schram sense, then it must be unimodular. Same for random rooted networks that are BSch-limits.*

{ex.qtransTuni

▷ **Exercise 14.3.**

**(a)** *Show that the universal covering tree of any finite graph, with a natural root measure, is unimodular. On the other hand, give an example of a quasi-transitive tree that is not unimodular. (More precisely, there exists no unimodular root measure on it). Compare with Exercise 2.3.*

**(b)** ** *Is it true that every unimodular quasi-transitive tree is the universal covering tree of a finite graph?*

One is often interested in the local behaviour of large finite *random* graphs; for instance, for the giant cluster phase transition of the Erdős-Rényi random graph in Theorem 12.23, a key idea was that $G(n, \lambda/n)$ is locally similar to a Galton-Watson tree with Poisson($\lambda$) offspring distribution. Thus, one should clearly expect that the local weak limit is this GW-tree, denoted by $\mathsf{PGW}(\lambda)$ from now on. However, since the sequence of finite graphs $G_n$ is itself random, there are two possibilities what local convergence may actually mean. In the weaker sense (sometimes called "annealed"), we want that in the joint distribution of the randomness in $G_n$ and of picking a uniform random vertex of $G_n$, the $r$-ball around this vertex converges in law to the $r$-ball of $G$. In other words, the distribution of the $r$-ball around $\rho_n$ converges in the average over $G_n$. In the stronger sense (sometimes called "quenched"), we want that the distribution of the $r$-ball around $\rho_n$ converges in probability, again over $G_n$. A pair of simple-minded examples to compare these two notions is the following: $G_n$ is the square box $[0,n]^2$ or the path $[0,n]$, with probability 1/2 each; $H_n$ is a disjoint union of $[0,n]^2$ and $[0,n^2]$. In both cases, the limit is the 1/2-1/2 random mixture of $\mathbb{Z}^2$ and $\mathbb{Z}$, but $G_n$ converges only on average, while $H_n$ also converges in probability (obviously, since this sequence is not even random). Of course, in the most natural examples, both convergences hold:

{ex.ERPGW}

▷ **Exercise 14.4.** *Let $G_n$ be the Erdős-Rényi random graph $G(n,\lambda/n)$, with any $\lambda \in \mathbb{R}_+$ fixed.*

**(a)** *Show that the local weak limit of $G_n$ on average is the $\mathsf{PGW}(\lambda)$ tree, rooted as normally.*

**(b)** *Let $X_k(n)$ be the number of degree $k$ vertices in $G_n$. Show that $X_k(n)/n$ converges in probability, as $n \to \infty$, to $\mathbf{P}[\,\mathsf{Poisson}(\lambda) = k\,]$. (Hint: calculate the 2nd moment, or use the small variance Azuma-Hoeffding, Proposition 1.9, for the edge exposure martingale from Section 6.3.)*

**(c)** *Using part (b), strengthen part (a) to convergence in probability.*

Rather famous examples of truly random unimodular random rooted graphs are the Uniform Infinite Planar Triangulation and Uniform Infinite Planar Quadrangulation, which are the local weak limits of uniform planar maps with $n$ triangle or quadrangle faces, respectively, see [Ang03] and [BenC10], and the references there.

Unimodularity is clearly a strengthening of the **stationarity of the delayed random walk**, i.e., that the Markov chain $(G,\rho) \mapsto (G,\rho')$ given by the delayed SRW is stationary on $(\mathcal{G}_*,\mu)$. This strengthening is very similar to reversibility, and one could try an alternative description of "looking back" from $\rho'$ and working with $\mathcal{G}_{**}$: namely, one could require that the Markov chain $(G,\rho) \mapsto (G,\rho')$ given by the delayed SRW be **reversible** on $(\mathcal{G}_*,\mu)$. This does not always work: e.g., if $(G,\rho)$ is a deterministic transitive graph, then reversibility holds even without unimodularity. Nevertheless, we have the following simple equivalences:

▷ **Exercise 14.5.*** *Consider $\mathsf{Ber}(p)$ percolation with $0 < p < 1$ on a transitive graph $G$. Show that the following are equivalent:* {ex.percunimod

**(a)** *$G$ is unimodular;*

**(b)** *the cluster of a fixed vertex $\rho$ is a unimodular random graph with $\rho$ as its root;*

**(c)** *the delayed SRW on the cluster generates a reversible chain on $\mathcal{G}_*$.*

*Formulate a version for general invariant percolations on $G$.*

▷ **Exercise 14.6.** *If the previous exercise was too hard for you, just show that delayed SRW on the $\mathsf{Ber}(p)$ percolation cluster of the origin of the grandparent graph is not stationary on $\mathcal{G}_*$ (regardless of the cluster being finite or infinite).*

Let us remark that a random rooted graph $(G,\rho)$ is called stationary in [BenC10] if (non-delayed) SRW generates a stationary chain on $\mathcal{G}_*$, and is called reversible if the chain generated on $\mathcal{G}_{**}$ by the "looking back" procedure is stationary. In other words, using the inverse of Definition 14.1 (e): if we bias the distribution by $1/\deg(\rho)$, then we get a unimodular random rooted graph. Using this reweighting, one can often easily translate results between reversible random graphs and unimodular random graphs with finite expected degree; for different problems, one or the other may be the more natural setting.

{ex.PGW}

▷ **Exercise 14.7.**

**(a)** *Show that the Poisson Galton-Watson tree* $\mathsf{PGW}(\lambda)$*, rooted as normally, is unimodular. (One possibility is to just quote Exercises 14.2 and 14.4, but you are encouraged to check the Mass Transport Principle or the "looking back in the delayed random walk" definition directly.)*

**(b)** *Show that if we size-bias* $\mathsf{PGW}(\lambda)$ *by* $\deg(\rho)$*, we get a rooted tree given by connecting the roots of two i.i.d. copies of* $\mathsf{PGW}(\lambda)$ *by an edge, then choosing the root uniformly from the two.*

We have seen so far two large classes of unimodular random rooted networks: Cayley graphs (and their percolation clusters) and local weak limits of finite graphs and networks. We have also seen that amenable Cayley graphs are actually local weak limits of finite graphs. Here comes an obvious fundamental question: can we get all Cayley graphs and unimodular random rooted networks as local weak limits of finite graphs and networks? To start with, can we get regular trees as limits?

{ex.bipgirth}

▷ **Exercise 14.8.** *Fix* $d \in \mathbb{Z}_+$*, take* $d$ *independent uniformly random permutations* $\pi_1, \dots, \pi_d$ *on* $[n]$*, and consider the bipartite graph* $V = [2n]$*,* $E = \{(v, n + \pi_i(v)) : 1 \le i \le d, 1 \le v \le n\}$*.*

**(a)** *Show that the number of multiple edges remains tight as* $n \to \infty$*.*

**(b)** *Show that the local weak limit of these random bipartite graphs is the* $d$*-regular tree* $\mathbb{T}_d$*.*

It is also true that the sequence of uniformly chosen random $d$-regular graphs converges to the $d$-regular tree; this follows from Corollary 2.19 of [Bol01]. This model is usually handled by proving things for the so-called configuration model (which is rather close to a union of $d$ independent perfect matchings), and then verifying that the two measures are not far from each other from the point of view of what we are proving.

Now that we know that amenable transitive graphs and regular trees are local weak approximable (Exercises 14.1 and 14.8), it is not completely ridiculous to ask the converse to Exercise 14.2:

{q.sofic}

**Question 14.2** ([Gro99, Wei00],[AldL07])**.** *Is every f.g. group* **sofic***, i.e., does every Cayley diagram have a sequence of labelled finite digraphs* $G_n$ *converging to it in the Benjamini-Schramm sense? In particular, is every Cayley graph a local weak limit of finite graphs? And more generally, is every unimodular random rooted graph or network a local weak limit?*

We need to clarify a few things about these three questions. First of all, a group being sofic has a definition that is independent of its Cayley diagram, i.e., of the generating set considered. This definition is that the group $\Gamma$ has a sequence of almost-faithful almost-actions on finite permutation groups: a sequence $\{\sigma_i\}_{i=1}^\infty$ of maps $\sigma_i : G \longrightarrow \mathrm{Sym}(n_i)$ with $n_i \to \infty$ such that

$$
\begin{aligned}
\forall f, g \in \Gamma: \quad & \lim_{i\to\infty} \frac{1}{n_i} \left| \left\{ 1 \le p \le n_i : p^{\sigma_i(f)\sigma_i(g)} = p^{\sigma_i(fg)} \right\} \right| = 1\,; \\
\forall f \ne g \in \Gamma: \quad & \lim_{i\to\infty} \frac{1}{n_i} \left| \left\{ 1 \le p \le n_i : p^{\sigma_i(f)} \ne p^{\sigma_i(g)} \right\} \right| = 1\,.
\end{aligned}
\tag{14.2}
$$

{e.soficperm}

{ex.sofic}

▷ **Exercise 14.9.** *Show that a f.g. group is sofic in the "almost-action by permutations" sense (14.2) if and only if one or any Cayley diagram of it is local weak approximable by finite labelled digraphs.*

If we have a convergent sequence of labelled digraphs, we can just forget the edge orientations and labels, and get a convergent sequence of graphs. The converse is false: a result of Ádám Timár [Tim12] says that there is a sequence of graphs $G_n$ converging to a Cayley graph $G$ of the group $(\mathbb{Z} * \mathbb{Z}_2) \times \mathbb{Z}_4$ for which the edges of $G_n$ cannot be oriented and labelled in such a way that the resulting networks converge to the Cayley diagram that gave $G$. The construction uses Theorem 14.15 below: although the vertex set of the 3-regular tree $\mathbb{T}_3$, the standard Cayley graph of $\mathbb{Z} * \mathbb{Z}_2$, can trivially be decomposed into two independent sets in an alternating manner, there exists a sequence $\{H_n\}$ of finite 3-regular graphs with girth going to infinity (hence locally converging to $\mathbb{T}_3$), for which the density of any independent set is less than $1/2 - \epsilon$ for some absolute constant $\epsilon > 0$. Timár constructed a Cayley diagram of $(\mathbb{Z} * \mathbb{Z}_2) \times \mathbb{Z}_4$ that "sees" the alternating decomposition of $\mathbb{T}_3$, hence can be locally approximated by the graphs $H_n \times C_4$ only if one forgets the labels. Nevertheless, the Cayley diagram is still sofic: for a local approximation, one needs to use finite 3-regular bipartite graphs with high girth, for which the alternating decomposition into two independent sets does exist.

This example of Timár shows that going from the local approximability of graphs to the soficity of groups might be a complicated issue. Indeed, answering both of the following questions with "yes" is expected to be hard:

{q.GraphsToGro

**Question 14.3.**

**(a)** *Does the local weak approximability of one Cayley graph of a group implies the local weak approximability of all its Cayley graphs? (Note that the answer to the same question for approximations of Cayley diagrams by labelled digraphs is an easy "yes", by Exercise 14.9 above.)*

**(b)** *Is it true that if all finitely generated Cayley graphs of a group are local weak approximable, then the group is sofic?*

A probably simpler version of this question is the following:

▷ **Exercise 14.10.***** *By encoding edge labels by finite graphs, is it possible to show that every Cayley graph of every group being local weak approximable would imply that every group is sofic?*

It is also important to be aware of the fact that local approximability by Cayley diagrams of finite groups is a strictly stronger notion than soficity. This is called the **LEF property**, locally embeddable into finite groups, because, similarly to (14.2), it can be formalized as follows: for any finite subset $F \subseteq \Gamma$, there is a finite symmetric group $\mathrm{Sym}(n)$ and an injective map $\sigma_n : F \longrightarrow \mathrm{Sym}(n)$ such that $\sigma_n(fg) = \sigma_n(f)\, \sigma_n(g)$ whenever $f, g, fg \in F$. Now, there are groups that are known to be sofic but not LEF: there exist solvable non-LEF groups [GoV97]. Furthermore, the Burger-Mozes group (a nice non-trivial lattice in $\mathrm{Aut}(\mathbb{T}_m) \times \mathrm{Aut}(\mathbb{T}_n)$ [BurgM00]) is a finitely presented infinite simple group, hence cannot be LEF by the following exercise. (In fact, it might even be a non-sofic group.)

{ex.LEF}

▷ **Exercise 14.11.**

**(a)** *Show that any residually finite group (defined in Exercise 2.8) has the LEF property.*

**(b)** *Show that a finitely presented LEF group is residually finite. Conclude that a finitely presented infinite simple group cannot have the LEF property.*

▷ **Exercise 14.12.**

**(a)** *Let $\Gamma = \Gamma_0 > \Gamma_1 > \dots$ be a sequence of subgroups, and $S$ a finite generating set of $\Gamma$. Consider the action of $\Gamma$ on the corresponding coset tree $\mathcal{T}$ as defined just before (13.21). Recall also the Farber condition, defined just before Exercise 13.26.*

*Note that the Schreier graphs of the action of $\Gamma$ on the $n^{th}$ level of $\mathcal{T}$ is just the Schreier graph $G_n := G(\Gamma, \Gamma_n, S)$, as defined just before Exercise 2.11. Show that $\{G_n\}_{n \geq 0}$ converges in the Benjamini-Schramm sense to the Cayley graph $G(\Gamma, S)$ iff the sequence $\{\Gamma_n\}_{n \geq 0}$ satisfies the Farber condition.*

**(b)** *Using the residual finiteness of $F_2$ (see Exercise 2.9), give a sequence of finite transitive graphs converging locally to the tree $\mathbb{T}_4$.*

In part (b) of the previous exercise, we obtained a sequence of finite groups that resemble the free group $F_2$ not just locally but in many other senses: indeed, we used there that $F_2$ is a finite index subgroup of the residually finite $\mathsf{SL}_2(\mathbb{Z})$, and the finite groups $\mathsf{SL}_2(\mathbb{Z}_p)$ resemble $\mathsf{SL}_2(\mathbb{Z})$ in almost any possible sense. However, the local structure does not reveal in general the global shape of the group:

{ex.NonExpLocT

▷ **Exercise 14.13** (High girth non-expanders)**.**

**(a)** *Let the* **free $d$-step solvable group on $r$ generators** *be $S_{r,d} := F_r/F_r^{(d)}$, where $F_r^{(d)}$ is the $d^{th}$ element in the derived series of the free group $F_r$. Consider its natural Cayley graph. What is its girth?*

**(b)** *Let $\Gamma$ be some finite group with large girth and two generators. Let $\varphi_1 : F_2 \longrightarrow \Gamma$ be a quotient map. Let $K_1 := \varphi^{-1}(e)$ be its kernel. Now let $\varphi_2 : F_2 \longrightarrow \mathbb{Z}$ be some quotient map, and let $K_2$ be its kernel. Consider the group $F_2/(K_1 \cap K_2)$. Show that it is an infinite group with linear volume growth and high girth. (I learnt this example from Gady Kozma.)*

**(c)** *From either of the two infinite examples above, produce a sequence of finite non-expander Cayley graphs converging locally to $F_2$.*

**(d)** *On the other hand, show that if an expander sequence $\{G_n\}$ converges in the Benjamini-Schramm sense to a transitive infinite graph $G$, then $G$ is non-amenable.*

And here is an example where the local structure has a huge global influence:

{ex.Zloc}

▷ **Exercise 14.14.** **

**(a)** *Show that if $G$ is the Cayley graph of a group that has $\mathbb{Z}$ as a finite index subgroup, then there is some $r \in \mathbb{N}$ such that $B_r^{G'}(o') \simeq B_r^G(o)$ for any infinite transitive graph $G'$ implies that $G' \simeq G$.*

**(b)** *Are there less trivial examples? Say, a finitely presented infinite simple group? (See also Exercise 14.11.)*

Based on the isoperimetric and spectral characterizations of expanders and non-amenability, one may think that these two notions are the exact analogues of each other. However, from Exercise 14.13 (c) and (d), and also from Exercise 7.20, it may also seem that being an expander sequence must be strictly stronger than non-amenablity; in fact, Section 7.4 suggests that Kazhdan's property (T) might be the right infinite group-theoretical analogue. The following conjecture states that this should be the case also from the point of view of Benjamini-Schramm convergence:

{c.BowenKazhda

**Conjecture 14.4** (Lewis Bowen)**.** *Any Cayley graph $G$ of a Kazhdan (T) group can be approximated in the Benjamini-Schramm sense only with essential expander graphs $G_n$: there is a $c > 0$ such that for any $\epsilon > 0$ if $n > n_0(\epsilon)$, then there is a way to erase at most $\epsilon|V(G_n)|$ vertices from $G_n$ such that all the remaining connected components are $c$-expanders.*

As we mentioned in Section 11.1, regular trees, the group $\mathsf{SL}_2(\mathbb{Z})$, and locally finite circle packings of the hyperbolic plane all have non-trivial harmonic Dirichlet functions. On the other hand, Kazhdan (T) groups, such as $\mathsf{SL}_d(\mathbb{Z})$ with $d \geq 3$, do not have such functions. What is the apparent importance of planarity here? One possible semi-answer is that planar graphs cannot be expanders:

{t.LipTar}

**Theorem 14.5** (Lipton-Tarjan planar separator theorem [LipT79])**.** *Any planar graph $G$ on $n$ vertices has a subset $S \subset V(G)$ of at most $2\sqrt{2}\sqrt{n}$ vertices such that every connected component of $G \setminus S$ has at most $2n/3$ nodes.*

Instead of proving this theorem, here is an exercise of a similar flavor, due to [Ths83]:

▷ **Exercise 14.15.**

**(a)** *Show that there exists an absolute constant $\gamma > 0$ such that, for any $k \in \mathbb{Z}_+$, if a graph $G$ has minimum degree at least 3 and girth at least $\gamma k$, then $G$ can be contracted into a multigraph $H$ that has minimum degree $k$ and no two vertices are joined by more than two edges.*

**(b)** *Using Euler's formula $1 - |V| + |E| - |F| + 1 = 0$ for finite planar graphs, give an upper bound on the average degree of finite planar graphs. From this and part (a), deduce that there is an absolute constant $g > 0$ such that a finite planar graph cannot have minimum degree at least 3 and girth at least $g$.*

**(c)** ** *What is the smallest possible $g$ in part (b)?*

As usual, Theorem 14.5 and the previous exercise have generalizations from planarity to other excluded minors: see [KrSu09].

Having discussed the right graph sequence analogue of non-amenability and Kazhdan's property (T), a more than natural question arises: what is the graph sequence analogue of amenability?

{d.hyperfin}

**Definition 14.6** ([KecM04, Ele07])**.** *A sequence of finite graphs $G_n$ is called* **hyperfinite** *if for every $\epsilon > 0$ there is a $K < \infty$ such that, for all large enough $n$, one can remove at most $\epsilon|V(G_n)|$ edges from $E(G_n)$ so that each connected component of the resulting graph has at most $K$ vertices.*

The boxes $[-n, n]^d$ in $\mathbb{Z}^d$ clearly form a hyperfinite sequence. Using the Lipton-Tarjan Separator Theorem 14.5, it was shown in [LipT80] that any bounded degree sequence of planar graphs is hyperfinite. On the other hand, an expander sequence is clearly not hyperfinite. Two further simple examples are given in the next exercise:

{ex.hypfinX}

▷ **Exercise 14.16.**

**(a)** *Show that any sequence of finite trees with a uniform bound on the degrees is hyperfinite.*

**(b)** *Give a sequence of finite graphs that has neither a hyperfinite nor an expander subsequence.*

We saw in Exercise 14.13 (d) that if an expander sequence BSch-converges to a transitive graph, then the limit must be non-amenable. Similarly, if a hyperfinite sequence converges to a transitive graph, then the limit must be amenable:

{ex.hypfintran

▷ **Exercise 14.17.** *Show that a transitive graph $G$ has a hyperfinite graph sequence $G_n$ converging to it in the local weak sense iff $G$ is amenable.*

We should of course go beyond the world of transitive graphs, and define amenability for unimodular random rooted networks, as well:

{d.hyperfinurn

**Definition 14.7** ([Schr08, AldL07])**.** *A unimodular random graph $(G, \rho)$ is called* **amenable** *or* **hyperfinite** *if for every $\epsilon > 0$ there is a $K$ and a subset $S \subset E(G)$ such that $(G, S, \rho)$ is unimodular (as a 0/1-labeling of $(G, \rho)$), every component of $G \setminus S$ has size at most $K$, and the expected number of edges of $S$ adjacent to $\rho$ in $G$ is at most $2\epsilon$.*

This notion is obviously the right generalization of hyperfiniteness of finite graphs; the reason for using $2\epsilon$ is that an $(\epsilon, K)$-hyperfinite finite graph $G$ with a uniform random root thus becomes an $(\epsilon, K)$-hyperfinite random rooted graph, since each edge in $S$ contributes to the degree of two vertices. On the other hand, this is also the right generalization of amenability by Theorem 12.12. Another "proof" is the following exercise:

▷ **Exercise 14.18.** *Assume that a sequence $\{G_n\}$ of finite graphs BSch-converges to a unimodular random graph $(G, \rho)$. Prove the following generalizations of Exercises 14.13 (d) and 14.17:*

**(a)** *If the sequence $\{G_n\}$ is hyperfinite, then $(G, \rho)$ is amenable.*

**(b)** *If $\{G_n\}$ is an expander sequence, then $(G, \rho)$ is non-amenable.*

Given that non-amenability has turned out to be in some sense weaker than being an expander (losing against Kazhdan's property (T)), one may ask if amenability is not weaker than hyperfiniteness. The answer is no, by the following analogue of Conjecture 14.4; the proof is explained well in [Lov12, Section 21.1]: {t.hyperfinurn

**Theorem 14.8** ([Schr08])**.** *If a sequence $\{G_n\}$ BSch-converges to an amenable unimodular random graph $(G, \rho)$, then $\{G_n\}$ is hyperfinite.*

Amenability and residual finiteness are two very different reasons for soficity. One can also combine them easily:

▷ **Exercise 14.19.** *Define a good notion of* **residual amenability**, *and show that residually amenable groups are sofic.*

There are examples of sofic groups that are not residually amenable [ElSz06]. Further sources of provably sofic groups can be found in [Ele15].

Most people think that the answer to Question 14.2 is "no" (maybe everyone except for Russ Lyons, but he also has some good reasons). For instance, Gromov says that if a property is true for all groups, then it must be trivial. (A counterexample to this meta-statement is the Erschler-Lee-Peres theorem about the $c\sqrt{n}$ escape rate, Exercise 10.4 and Theorem 10.7, but maybe quantitative results do not count. But the $\theta(p_c) = 0$ percolation conjecture is certainly a serious candidate. Maybe probability does not count? Or these questions are not really about groups, but, say, transitive graphs? Note that these two things are not the same — see Chapter 16.) One reason that this is an important question is that there are many results known for all sofic groups.

Yet another view on soficity is Gottschalk's surjunctivity question [Got73]. A group $\Gamma$ is called **surjunctive** if for any integer $k \geq 2$, whenever a continuous self-map $\varphi$ of the colouring space $\{1, 2, \ldots, k\}^\Gamma$ that commutes with the $\Gamma$-action is injective, it is automatically surjective, as well. To start digesting this notion, do the following exercise: {ex.surj}

▷ **Exercise 14.20.**

**(a)** *Consider $\Gamma = \mathbb{Z}$ and $k = 2$, and the map $\varphi(f)(i) := f(i) + f(i+1) \pmod 2$. Show that this is surjective, but not injective.*

**(b)** *Show that the set of all periodic $k$-colourings of $\mathbb{Z}$ are dense in $\{1, 2, \ldots, k\}^{\mathbb{Z}}$. Note that any $\varphi$ that commutes with the $\mathbb{Z}$-action preserves the property of having a certain period $p$. Since the set of colourings of period $p$ is a finite set, if $\varphi$ is injective, then it is surjective on periodic colourings. Conclude that an injective $\varphi$ is surjective on $\{1, 2, \ldots, k\}^{\mathbb{Z}}$, hence $\mathbb{Z}$ is surjunctive.*

**(c)** *Similarly to part (b), show that any residually finite group is surjunctive.*

It was proved in [Wei00] that all sofic groups are also surjunctive; the converse is not known. Constructing any non-surjunctive group would also give a non-sofic one, and it might be easier to think about why a specific injective $\varphi$ is not surjective than to think about why there cannot exist *any* local approximation to the group by finite graphs. Also, Benjy Weiss [Wei00] used surjuncitvity to support his guess that non-sofic groups must exist: he showed that although any mixing subshift of finite type for $\Gamma = \mathbb{Z}$ is surjunctive, this already fails for $\Gamma = \mathbb{Z}^2$. Here, **subshift of finite type** means a closed $\Gamma$-invariant subset

$X$ of $\{1, 2, \ldots, k\}^\Gamma$ that is given by a list of how any finite window of $\ell$ consecutive symbols may look like, while **mixing** means that for any open sets $U, V \subset X$, there is a finite set $F \subset \Gamma$ such that $U \cap g(V) \neq \emptyset$ for all $g \in \Gamma \setminus F$.

▷ **Exercise 14.21.** *Show that the full shift $\{1, 2, \ldots, k\}^\Gamma$ for any infinite group $\Gamma$ and any integer $k \geq 2$ is mixing.*

▷ **Exercise 14.22.** *Show that the subshift of finite type $X \subset \{0, 1, 2\}^{\mathbb{Z}}$ given by the list of $\ell = 2$-windows $\{00, 01, 11, 12, 22\}$ is not mixing and not surjunctive. (Hint: let $\varphi$ rewrite the single appearance of $12222\ldots$, if it exists, to $11222\ldots$.)*

In [Wei00], the mixing non-surjunctive example for $\mathbb{Z}^2$ is given as a subshift of $\{0, 1, 2, *\}^{\mathbb{Z}^2}$. The non-$*$ symbols form infinite paths of width one, going diagonally from $(-\infty, -\infty)$ to $(+\infty, +\infty)$, with no more than three consecutive steps to the right or up, with the sequence of symbols belonging to the $\mathbb{Z}$-subshift of the previous exercise. These paths are then separated by some arbitrary width $\geq 5$ corridors of $*$-symbols. It is easy to see that this is a subshift of finite type, while the variable width of the $*$-corridors ensures the mixing property. This example supports the guess that non-surjenctive groups may exist because a strange mixing finite subshift of a certain group might possibly be modelled by the full shift of another, strange, group.

For the groundbreaking survey on local weak convergence and unimodularity, see [AldL07]; for a survey on soficity, see [Pes08]. In the next section, we will discuss the relevance of local weak convergence to probability on groups.

## 14.2 Spectral measures and other probabilistic questions {ss.spectral}

Although the question of local weak convergence is orthogonal to being quasi-isometric (in the sense that the former cares about local structure only, while the latter cares about global structure), it still preserves a lot of information about probabilistic behaviour. There are also examples where the information preserved is not quite what one may first expect. In fact, the Benjamini-Schramm paper where local weak convergence was introduced is about such an example: {t.BSrec}

**Theorem 14.9** ([BenS01])**.** *Let $G_n$ be a local weak convergent sequence of finite planar graphs with uniformly bounded degrees. Then the limiting unimodular random rooted graph $(G, o)$ is almost surely recurrent.*

For instance, the balls in a hyperbolic planar tiling converge locally to a recurrent unimodular random graph that is somewhat similar to the one-ended tree of Figure 14.1. The proof of the theorem is based on two theorems on circle packings. The first is Koebe's theorem that any planar graph can be represented as a circle packing; moreover, for triangulated graphs the representation is unique up to Möbius transformations. This helps normalize the circle packing representations of $G_n$ such that we get a circle packing for the limit graph $G$. Then one needs to show that this limiting circle packing has at most one accumulation point of centres, which implies by a result of He and Schramm that the graph is recurrent; see Theorem 11.1.

▷ **Exercise 14.23.** *Prove that any local weak limit of bounded degree finite trees is almost surely recurrent. (Hint: using a version of the Burton-Keane argument of Theorem 12.6, show that the root in the limit is a trifurcation point with probability 0, hence the limit tree has 1 or 2 ends.)*

A generalization of Theorem 14.9 is that if that there exists a finite graph such that none of $G_n$ contains it as a minor (e.g., planar graphs are characterized by not having either $K_5$ or $K_{3,3}$), then the limit is almost surely recurrent [AngSz]. The converse is false:

▷ **Exercise 14.24.** *Construct a local weak convergent sequence of uniformly bounded degree finite graphs $G_n$ such that for any finite graph $F$, if $n > n_F$, then $G_n$ contains $F$ as a minor, but the limit is almost surely recurrent. (Hint 1: decorate large finite pieces of $\mathbb{Z}$ with copies of the possible $F$'s. Hint 2: You get a probabilistically trivial but graph theoretically harder example by showing that the $\mathbb{Z}^2$ lattice with the diagonals added as edges contains any finite graph as a minor.)*

An extension of Theorem 14.9 in a different direction is to relax the condition on uniformly bounded degrees to sequences where the degree of the uniform random root has at most an exponential tail [GuGN13]. The importance of this result is that it shows that the uniform infinite planar triangulation and quadrangulation (UIPT and UIPQ, mentioned shortly after Definition 14.1) are recurrent.

What other properties are preserved by local weak convergence?

{d.testable}

**Definition 14.10.** *Let $p(G)$ be a graph parameter: a number or some more complicated object assigned to isomorphism classes of finite (or sometimes also infinite) graphs. It is said to be* **locally approximable***, or simply* **local***, or* **testable***, if it is continuous w.r.t. local weak convergence: whenever $\{G_n\}$ is a convergent sequence of finite graphs, $\{p(G_n)\}$ also converges in distribution.*

A simple but important example is the set of simple random walk return probabilities $p_k(o,o) = \mathbf{P}[\, X_k = o \mid X_0 = o\,]$: if the $r$-neighbourhood of a vertex $o$ in a graph $G$ is isomorphic to the $r$-neighbourhood of $o'$ in $G'$, and $r$ is at least $k/2$, then obviously $p_k^G(o,o) = p_k^{G'}(o',o')$ in the two graphs. In particular, if a sequence of finite graphs $G_n$, with a uniform random root $\rho_n$, converges to a unimodular random rooted graph $(G,\rho)$, then, for any $k \in \mathbb{Z}_+$ fixed, $\{p_k^{G_n}(\rho_n,\rho_n)\}_{n=1}^\infty$ converges in distribution to $p_k^G(\rho,\rho)$. A fancy reformulation of this observation is that the **spectral measure** of the random walk is locally approximable. Here is what we mean by this; a standard reference is [MohW89].

For any reversible Markov chain, e.g., simple random walk on a graph $G(V,E)$, the Markov operator $P$ is self-adjoint w.r.t. $(\cdot,\cdot)_\pi$, and hence we can take its spectral decomposition $P = \int_{-1}^1 t\, dE(t)$, where $E$ is a projection-valued measure on $\ell^2(V,\pi)$, a resolution of the identity, $I = \int_{-1}^1 dE(t)$. (If $G$ is a finite graph on $n$ vertices, then $P$ has $n$ eigenvalues $-1 \le \lambda_i \le 1$, with orthogonal eigenvectors $f_i$ with $\ell^2(V,\pi)$-norm 1, for each of them we have the projection $E_i(f) := (f,f_i)_\pi f_i$ on the eigenline spanned by $f_i$, and $P = \sum_{i=1}^n \lambda_i E_i$.) Now consider the unit vectors $\varphi_x := \delta_x/\sqrt{\pi(x)}$ for each $x \in V$, and define, for any measurable $S \subset [-1,1]$,

$$\sigma_{x,y}(S) := (\varphi_x, E(S)\varphi_y)_\pi = \int_S d(\varphi_x, E(t)\varphi_y)_\pi\,, \tag{14.3}$$ {e.Kestenmeas}

sometimes called the **Kesten spectral measures** or **Plancherel measures**. For $x = y$, these are probability measures on $[-1,1]$, while, for $x \neq y$, are signed measures with zero total mass.

{ex.specmeas}

▷ **Exercise 14.25.**

**(a)** *If $G(V,E)$ is a finite graph on $n$ vertices, it is natural to take the average of the above spectral measures: $\sigma_G := \frac{1}{n}\sum_{x\in V}\sigma_{x,x}$. On the other hand, it is also natural to take the normalized counting measure on the eigenvalues of the Markov operator, $\frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$, and call that the spectral measure of $G$. Show that $\sigma_{x,x}(\{\lambda_i\}) = f_i(x)^2\pi(x)$ with the unit eigenvectors $f_i$, and deduce that the two definitions give the same spectral measure $\sigma_G$.*

**(b)** ** *One may also consider taking the weighted average $\hat\sigma_G := \sum_{x\in V}\sigma_{x,x}\,\pi(x)$, similarly to the size-biasing in Definition 14.1 (e). Is there a nice way to express $\hat\sigma_G$ using the eigenvalues $\{\lambda_i\}$?*

Now notice that

$$\frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}}\, p_n(x,y) = (\varphi_x, P^n\varphi_y)_\pi = \int_{-1}^1 t^n\, d\sigma_{x,y}(t)\,, \tag{14.4}$$ {e.Kestenmom}

hence the return probabilities are given by the moments of the Kesten spectral measures $\sigma_{x,x}$. Since these measures have compact support, they are determined by their moments, and the local approximability of the return probabilities (and of the ratio $\pi(x)/\pi(y) = \deg(x)/\deg(y)$) implies that the spectral measures are also locally approximable. More precisely:

{ex.localspect

▷ **Exercise 14.26.** *Let $G_n$ be finite graphs converging to $(G,\rho)$ in the local weak sense. Then the spectral measure of $G_n$, as defined in Exercise 14.25 (a), converges weakly (weak convergence of measures) to the Kesten spectral measure $\mathbf{E}\sigma_{\rho,\rho}$ of $(G,\rho)$, averaged w.r.t. the randomness in the limit $(G,\rho)$.*

However, note that this does not mean that the supports of these measures also converge: for instance, each finite $G_n$ has 1 in its spectrum, while, if $G$ is a non-amenable transitive graph, then its spectral measure is bounded away from 1. And the eigenvalue 1 is not just a singular exception: as we saw in Exercise 14.13, one can find finite Cayley graphs that are very far from being expanders (that is, they have eigenvalues close to 1), nevertheless converge locally to the free group.

Although the supports of the spectral measures do not necessarily converge, more is true than just the weak convergence of measures proved in Exercise 14.26:

{t.Luck}

**Theorem 14.11** (Lück approximation for combinatorialists [Lüc94])**.** *Let $\{G_n\}_{n=1}^{\infty}$ be a sequence of finite graphs with degrees at most $D$, with edges labelled by integers from $\{-D,\dots,D\}$. Assume that $\{G_n\}$ converges in the local weak sense (with the obvious generalization handling the labels.) Let $A_n = \mathrm{Adj}(G_n)$ be the adjacency matrices with the labels being the entries. Then $\dim\big(\ker_{\mathbb{Q}} A_n\big)/|V(G_n)|$ converges.*

*Proof.* First of all, note that $\dim\ker_{\mathbb{Q}} A_n/|V(G_n)| = \sigma_n(\{0\})$, where $\sigma_n$ is the spectral measure of $A_n$. Consider now all the nonzero eigenvalues $\lambda_i \in \mathbb{R}$ of $A_n$. Their product is a coefficient of the characteristic polynomial, hence is a nonzero integer. But all the absolute values $|\lambda_i|$ are at most $D^2$, hence not too many of them can be very small. More precisely, the spectral measure of $A_n$ satisfies $\sigma_n\big([-\epsilon,\epsilon]\setminus\{0\}\big) < D'/\log(1/\epsilon)$ for any $\epsilon > 0$, with some $D'$ depending on $D$. Together with the weak convergence of the spectral measures, this implies that $\sigma_n(\{0\})$ also converges, and we are done. □

This was generalized by [Thm08] to atoms not only at 0, but at any $x \in \mathbb{R}$. In particular, this holds for the atoms in the spectral measure of the Markov operator of regular graphs. Note that this implies that the atoms in the spectral measure of a sofic Cayley graph are always located at algebraic integers.

After these convergence results, let us see some examples for the spectral measures of infinite graphs.

▷ **Exercise 14.27.**

**(a)** *Show that the $d$-regular tree $\mathbb{T}_d$ with $d \geq 2$ (i.e., including $\mathbb{Z}$) has no eigenvectors $\lambda f = Pf$ with $f \in \ell^2(\mathbb{T}_d)$, for any $\lambda \in \mathbb{R}$. (Hint: assuming there is one, show that there would also be one whose values depend only on the distance from the root; then exclude this by direct computation.)*

**(b)** *Show that the quasi-transitive tree $T$ on the left hand side of Figure 12.5, which is unimodular by Exercise 14.3, has an $\ell^2(T)$-eigenvector, with eigenvalue 0.*

{ex.Zspec}

▷ **Exercise 14.28** (The spectral measure of $\mathbb{Z}$)**.**
*For SRW on $\mathbb{Z}$, the Kesten spectral measure is $d\sigma_{x,x}(t) = \frac{1}{\pi\sqrt{1-t^2}}\mathbf{1}_{[-1,1]}(t)\,dt$. (Hint: you could do this in at least two ways: either from the spectrum of the cycle $C_n$, i.e., a combination of Exercises 7.5 and 14.26, or from (14.4), computing return probabilities and moments explicitly, and arguing that the spectral measure is determined by its moments.)*

{ex.prodspec}

▷ **Exercise 14.29.**

**(a)** *Show that for the spectral measures $\sigma^{G_i}_{x,y}$ associated to the adjacency matrices (as opposed to the Markov operators) of two graphs $G_i$, $i = 1, 2$, if $u = (u_1, u_2)$ and $v = (v_1, v_2)$ are two vertices in the direct product $G_1 \times G_2$, then $\sigma^{G_1 \times G_2}_{u,v} = \sigma^{G_1}_{u_1,v_1} * \sigma^{G_2}_{u_2,v_2}$, a convolution of measures.*

**(b)** *Note that the previous statement can be easily translated to the spectral measures of the Markov operators only when both $G_i$ are regular graphs. Deduce, for instance, that the direct product of two amenable groups is amenable.*

▷ **Exercise 14.30** (The spectral measure of $\mathbb{T}_d$)**.** * {ex.Tspec}

**(a)** *Consider the Kesten spectral measure $\sigma = \sigma_{x,x}$ of a Cayley graph. Its Stieltjes transform is*

$$S(z) := \int \frac{1}{t - z}\, d\sigma(t)\,,$$

*a finite integral for any $\Im z > 0$. Show that*

$$S(z) = \sum_{k=0}^{\infty} \frac{-m_k}{z^{k+1}}\,,$$

*where $m_k := \int t^k d\sigma(t)$ is the $k^{th}$ moment of the spectral measure.*

**(b)** *Prove the following nice interpretation of the Stieltjes transform (which I learnt from Charles Bordenave): for any continuous real function $f$ with bounded support,*

$$\frac{1}{\pi} \int f(t) \cdot \Im\, S(t + i\epsilon)\, dt = \mathbf{E} f(X + \epsilon Y)\,,$$

*where $X$ is distributed according to $\sigma$, and $Y$ is an independent Cauchy variable. Deduce the inversion formula*

$$\frac{d\sigma(t)}{dt} = \lim_{\epsilon \to 0+} \frac{\Im\, S(t + i\epsilon)}{\pi}\,,$$

*whenever the right hand side exists and is continuous in $t$.*

**(c)** *Using the strategy of Section 1.1, find $S(z)$ for the tree $\mathbb{T}_d$. Using the previous part, find the density of the Kesten spectral measure of $\mathbb{T}_d$.*

**(d)** *Note that the spectral measure of $\mathbb{T}_d$ you have just computed is supported on $[-\frac{2\sqrt{d-1}}{d}, \frac{2\sqrt{d-1}}{d}]$, and, as $d \to \infty$, it converges to Wigner's semicircle law (which is the $n \to \infty$ limit of the normalized counting measure on the eigenvalues of the $n \times n$ classical $\beta = 1, 2, 4$ random matrix ensembles).*

▷ **Exercise 14.31** ([Kai80])**.** * *Show that for any symmetric finitely supported $\mu$ on an amenable group, the associated random walk has not only $\rho(P) = 1$, but 1 also lies in the support of the spectral measure $d\sigma_{x,x}(t)$. In fact, for any $h < 1$ and $\epsilon > 0$,* {ex.amenspecme

$$\sigma_{x,x}[1 - h, 1] \geq \frac{1 - 2\epsilon/h^2}{|A_\epsilon|}\,,$$

*where $A_\epsilon$ is any Følner set with $|A_\epsilon g \Delta A_\epsilon| < \epsilon |A_\epsilon|$ for all $g \in \operatorname{supp} \mu$.*

For a long while, it was not known if simple random walk on a group can have a spectral measure not absolutely continuous with respect to Lebesgue measure. Discrete spectral measures are usually associated with random walk like operators on random underlying structures, e.g., with the adjacency or the transition matrices of random trees [BhES12], or with **random Schrödinger operators** $\Delta + \mathrm{diag}(V_i)$ on $\mathbb{Z}^d$ or $\mathbb{T}_d$, with $V_i$ being i.i.d. random potentials on the vertices [Kir07]. Note that the latter is a special case of the former in some sense, since the $V_i$ can be considered as loops with random weights added to the graph. Here, **Anderson localization** is the phenomenon that for large enough randomness in the potentials, the usual picture of absolutely continuous limiting spectral measure, with repulsing (or even lattice-like) eigenvalues and spatially extended eigenfunctions in the finite approximations, disappears, and Poisson statistics for the eigenvalues appears, with localized eigenfunctions, giving $L^2$-eigenfunctions, hence atoms, in the limiting spectral measure. However, the exact relationship between the behaviour of the limiting spectral measure and the eigenvalue statistics of the finite approximations has only been partially established, and it is also unclear when exactly localization and delocalization happen. Assuming that the random potentials have a nice distribution, localization is known for arbitrary small (but fixed) variance on $\mathbb{Z}$, and very large variance on other graphs, while delocalization is conjectured for small variance on $\mathbb{Z}^d, d \geq 3$, proved on $\mathbb{T}_d$. The case of $\mathbb{Z}^2$ is unclear even conjecturally. In the context of our present notes, it looks like a strange handicap for the subject that the role of the Benjamini-Schramm convergence was discovered as a big surprise only in [AiW06], but still without realizing that this local convergence has a well-established theory, e.g., for random walks.

Nevertheless, deterministic groups can also exhibit discrete spectrum: Grigorchuk and Żuk showed that the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}$ has a self-similar action on the infinite binary tree (see Section 15.1, around (15.7)), and, with the corresponding generating set, the resulting Cayley graph $G$ has a pure discrete spectral measure. They used the finite Schreier graphs $G_n$ for the action on the $n$th level of the tree converging locally to $G$. See [GriŻ01]. An alternative proof was found in [DicS02], later extended by [LeNW07], which interpret the return probabilities of SRW on the lamplighter group $F \wr \mathbb{Z}^d$ as the averaged return probabilities of a SRW on a $p$-percolation configuration on $\mathbb{Z}^d$, with parameter $p = 1/|F|$, where the walk is killed when it wants to exit the cluster of the starting vertex. (So, in a certain sense, the picture of a random walk on a random underlying structure is somehow still there.)

On the other hand, the following is still open. For a rough intuitive "definition" of the von Neumann dimension $\dim_\Gamma$, see the paragraph before Theorem 11.3.

{c.Atiyah}

**Conjecture 14.12** (Atiyah 1976)**.** *Simple random walk on a torsion-free group $\Gamma$ cannot have atoms in its spectral measure $d\sigma_{x,x}$, and more generally, any operator on $\ell^2(\Gamma)$ given by multiplication by a non-zero element of the group algebra $\mathbb{C}\Gamma$ has a trivial kernel. Even more generally, the kernel of any matrix $A \in \mathcal{M}_{n\times k}(\mathbb{C}\Gamma)$, as a linear operator $\ell^2(\Gamma)^n \longrightarrow \ell^2(\Gamma)^k$, has integer $\Gamma$-dimension*

$$\dim_\Gamma(\ker A) := \sum_{i=1}^{n} (\pi_{\ker A} e_i, e_i)_{\ell^2(\Gamma)^n}\,,$$

*where $\pi_H$ is the orthogonal projection onto the subspace $H \subseteq \ell^2(\Gamma)^n$ and $e_i$ is the standard basis vector of $\ell^2(\Gamma)^n$, with a Kronecker $\delta_e(g)$ function (where $e \in \Gamma$ is the identity) in the $i^{th}$ coordinate.*

How is the claim about the trivial kernel a generalization of the random walk spectral measure being atomless? The Markov operator $P$, generated by a finitely supported symmetric measure $\mu$, can be represented as multiplication of elements $\varphi = \sum_{g\in\Gamma} \varphi(g)g \in \ell^2(\Gamma)$ by $\mu = \sum_{g\in\Gamma} \mu(g)g \in \mathbb{R}_{\geq 0}\Gamma$. Now, because of group-invariance, having an atom in $d\sigma_{x,x}(t)$ at $t_0 \in [-1,1]$ is equivalent to having an atom at $dE(t_0)$, and that means there is a non-trivial eigenspace in $\ell^2(\Gamma)$ for $P$ with eigenvalue $t_0$. Indeed, $\dim_\Gamma(\ker(\mu - t_0 e))$ is exactly the size $\sigma_{x,x}(\{t_0\})$ of the atom.

The conjecture used to have a part predicting the sizes of atoms for groups with torsion; a strong version was disproved by the lamplighter result [GriŻ01], while all possible weaker versions have recently been disproved by Austin and Grabowski [Aus13, Grab10]. The proof of Austin is motivated by the above percolation picture.

Here is a nice proof of the Atiyah conjecture for $\mathbb{Z}$ that I learnt from Andreas Thom (which might well be the standard proof). The Fourier transform

$$a = (a_n)_{n\in\mathbb{Z}} \;\mapsto\; \mathbf{a}(t) = \sum_{n\in\mathbb{Z}} a_n \exp(2\pi i t n),\; t \in S^1$$

is a Hilbert space isomorphism between $\ell^2(\mathbb{Z})$ and $L^2(S^1)$, and also identifies multiplication in $\mathbb{C}\mathbb{Z}$ (which is a convolution of the coefficients) with pointwise multiplication of functions on $S^1$. So, the kernel $H_a$ for multiplication by $a \in \mathbb{C}\mathbb{Z}$ in $\ell^2(\mathbb{Z})$ is identified with $H_{\mathbf{a}} = \{f(t) \in L^2(S^1) : \mathbf{a}(t)f(t) = 0\}$, and projection on $H_{\mathbf{a}}$ is just multiplication by $\mathbf{1}_{Z_{\mathbf{a}}}(t)$, the indicator function of the zero set of $\mathbf{a}(t)$. Then, by the Hilbert space isomorphism,

$$\dim_{\mathbb{Z}}(H_a) = (\pi_{H_a} e, e)_{\ell^2(\mathbb{Z})} = \int_{S^1} \mathbf{1}_{Z_{\mathbf{a}}}(t)\, dt = \mathrm{Leb}(Z_{\mathbf{a}}) = 0\,, \tag{14.5}$$ {e.LebZa}

since $\mathbf{a}(t)$ is a nonzero trigonometric polynomial, and we are done.

{ex.arcsine}

▷ **Exercise 14.32.** ** *A variable $X \in [0,1]$ follows the* **arcsine law** *if $\mathbf{P}[\,X < x\,] = \frac{2}{\pi}\arcsin(\sqrt{x})$, or, in other words, has density $(\pi\sqrt{x(1-x)})^{-1}$. This distribution comes up in several ways for Brownian motion $B_t$ on $\mathbb{R}$, the scaling limit of SRW on $\mathbb{Z}$: the location of the maximum of $\{B_t : t \in [0,1]\}$, the location of the last zero in $[0,1]$, and the Lebesgue measure of $\{t \in [0,1] : B_t > 0\}$ all have this distribution. Is there a direct relation to the spectral measure density in Exercise 14.28? A possibly related question: is there a quantitative version of (14.5) relating the return probabilities $\asymp n^{-1/2}$ on $\mathbb{Z}$ to the dimension 1/2 of the zeroes of Brownian motion? This relationship is classical, but can you formulate it using projections? See [MöPe10] for background on Brownian motion.*

To see a probabilistic interpretation of atoms in the spectral measure, note that for SRW on a group $\Gamma$, by (14.4), there is no atom at the spectral radius $\pm\rho(P)$ iff $p_n(x,x) = o(\rho^n)$. And this bound is actually true, even in a stronger form, proved using random walks and harmonic functions: Theorem 7.8 of [Woe00], a result originated in the work of Guivar'ch and worked out by Woess, says that whenever a group is transient (i.e., not quasi-isometric to $\mathbb{Z}$ or $\mathbb{Z}^2$), then it is also $\rho$-transient, meaning that $G(x, y \,|\, 1/\rho) < \infty$, for Green's function evaluated at the spectral radius $\rho = \rho(P)$. This clearly implies that there is no atom at $\pm\rho$.

Let us turn for a second to the question what types of spectral behaviour are robust under quasi-isometries, or just under a change of generators. The spectral radius $\rho$ being less than 1 is of course robust, since it is the same as non-amenability. On the other hand, the polynomial correction in $p_n(x,x) = o(\rho^n)$ can already be sensitive: for the standard generators in the free product $\mathbb{Z}^d * \mathbb{Z}^d$, for $d \geq 5$, we have $p_n(x,x) \asymp n^{-5/2}\rho^n$, while, if we take a very large weight for one of the generators in each factor $\mathbb{Z}^d$, then random walk in the free product will behave like random walk on a regular tree, giving $p_n(x,x) \asymp n^{-3/2}\rho^n$, see Exercise 1.7. This instability of the exponent was discovered by Cartwright, see [Woe00, Section 17.B].

▷ **Exercise 14.33.** *Explain why the strategy of Exercises 1.5, 1.6, 1.7 to prove the exponent $3/2$ in the free group does not necessarily yield the same $3/2$ in $\mathbb{Z}^d * \mathbb{Z}^d$.*

In a work in progress, Grabowski and Virág show that the discreteness of the spectral measure can be completely ruined by a change of variables: in the lamplighter group, by changing the generators, they can set the spectrum to be basically anything, from purely discrete to absolutely continuous.

The locality of the spectral measure suggests that if a graph parameter can be expressed via simple random walk return probabilities on the graph, then it might also be local. We have seen in Section 11.2 that the Uniform Spanning Tree measure of a finite graph is closely related to random walks. This motivates the following result of Russ Lyons:

{t.treeent}

**Theorem 14.13** (Locality of the tree entropy [Lyo05])**.** *For any finite graph $G(V,E)$, let $\tau(G)$ be the number of its spanning trees, and let $\mathbf{h}_{\rm tree}(G) := \frac{\log \tau(G)}{|V(G)|}$ be its* **tree entropy***. If $G_n$ converges in the Benjamini-Schramm sense to the unimodular random rooted graph $(G,\rho)$, then, under mild conditions on the unimodular limit graph (e.g., having bounded degrees suffices),*

$$\lim_{n\to\infty} \mathbf{h}_{\rm tree}(G_n) = \mathbf{E}\Big[\log \deg(\rho) - \sum_{k\geq 1} \frac{p_k^G(\rho,\rho)}{k}\Big], \tag{14.6}$$ {e.treeentlim}

*where $p_k^G(\rho,\rho)$ is the SRW return probability on $G$.*

*Sketch of the proof for a special case.* Let $L_G = D_G - A_G$ be the graph Laplacian matrix: $D_G$ is the diagonal matrix of degrees and $A_G$ is the adjacency matrix. (For a $d$-regular graph, this is just $d$ times our usual Markov Laplacian $I - P$.) The **Matrix-Tree Theorem** says that $\tau(G)$ equals $\det(L_G^{ii})$, where the superscript $ii$ means that the $i$th row and column are erased. By looking at the characteristic polynomial of $L_G$, it is easy to see that this truncated determinant is the same as $\frac{1}{n}\prod_{i=2}^n \kappa_i$, where $|V(G)| = n$ and $0 = \kappa_1 \leq \cdots \leq \kappa_n$ are the eigenvalues of $L_G$.

Assume for easier notation that $|V(G_n)| = n$. A less trivial simplification is that we will assume that each $G_n$ is $d$-regular. Then $\kappa_i = d(1-\lambda_i)$, where $-1 \leq \lambda_n \leq \cdots \leq \lambda_1 = 1$ are the eigenvalues of the Markov operator $P$. Thus

$$\frac{\log \tau(G_n)}{n} = \frac{-\log n}{n} + \frac{n-1}{n}\log d + \frac{1}{n}\sum_{i=2}^n \log(1-\lambda_i)\,. \tag{14.7}$$ {e.treeentn}

Consider the Taylor series $\log(1-\lambda) = -\sum_{k\geq 1}\frac{1}{k}\lambda^k$, for $\lambda$ bounded away from 1. Recall that, for the uniform random root $\rho$ in $G_n$, the invariance of trace w.r.t. the choice of basis implies that $\mathbf{E}\, p_k^{G_n}(\rho,\rho) = \frac{1}{n}\sum_{i=1}^n \lambda_i^k$. Putting these ingredients together,

$$\frac{1}{n}\sum_{i=2}^n \log(1-\lambda_i) = -\sum_{k\geq 1}\frac{1}{k}\left(\mathbf{E}\, p_k^{G_n}(\rho,\rho) - \frac{1}{n}\right). \tag{14.8}$$ {e.treeentm}

We are on the right track towards formula (14.6) — just have to address how to interchange the infinite sum over $k$ and the limit $n \to \infty$.

For lazy SRW in a fixed graph, the distribution after a large $k$ number of steps converges to the stationary one, which is constant $1/n$ in a $d$-regular graph with $n$ vertices. Let us therefore consider the lazy walk in $G_n$, with Markov operator $\tilde{P} = (I+P)/2$ and return probabilities $\tilde{p}_k^{G_n}(\cdot,\cdot)$. Any connected graph is at least 1-dimensional in the sense that any finite subset of the vertices that is not the entire vertex set has at least one boundary edge. Thus, Theorem 8.2 implies that

$$\left|\mathbf{E}\,\tilde{p}_k^{G_n}(\rho,\rho) - \frac{1}{n}\right| \leq \frac{C_d}{\sqrt{k}}, \tag{14.9}$$ {e.pkspeed}

for all $n$. So, if we had $\tilde{p}_k$ instead of $p_k$ on the RHS of (14.8), then we could use the summability of $k^{-3/2}$ to get a control in (14.8) that is uniform in $n$. But how could we relate the lazy SRW to the original walk?

If we add $d$ half-loops at each vertex, so that we get a $2d$-regular graph $\tilde{G}_n$, then SRW on this graph is the same as the lazy SRW on $G_n$, hence we can again write down the identities (14.7) and (14.8), now with $\tilde{G}_n$ and $\tilde{p}_k$. On the other hand, we obviously have $\tau(\tilde{G}_n) = \tau(G_n)$. Thus,

$$\frac{\log \tau(G_n)}{n} = \frac{-\log n}{n} + \frac{n-1}{n} \log(2d) - \sum_{k \geq 1} \frac{1}{k} \left( \mathbf{E}\, \tilde{p}_k^{G_n}(\rho, \rho) - \frac{1}{n} \right). \tag{14.10}$$ {e.treeentlazy

Now (14.9) implies that for any $\epsilon > 0$, if $K$ and $n$ are large enough, then

$$\left| \frac{\log \tau(G_n)}{n} - \log(2d) - \sum_{k=1}^{K} \frac{1}{k} \left( \mathbf{E}\, \tilde{p}_k^{G_n}(\rho, \rho) - \frac{1}{n} \right) \right| < \epsilon\,.$$

Take $n \to \infty$ and then $K \to \infty$. For each fixed $k$, we have $\mathbf{E}\, \tilde{p}_k^{G_n}(\rho,\rho) \to \mathbf{E}\, \tilde{p}_k^{G}(\rho,\rho)$, which yields

$$\lim_{n \to \infty} \frac{\log \tau(G_n)}{n} = \log(2d) - \sum_{k=1}^{\infty} \frac{1}{k} \mathbf{E}\, \tilde{p}_k^{G}(\rho, \rho)\,.$$

Note here that the infinite sum is finite either by taking the limit $n \to \infty$ in (14.9) or by the infinite chain version of the same Theorem 8.2.

This already shows the locality of the tree entropy, but we still would like to prove formula (14.6). This is accomplished by Exercise 14.34 below, the infinite graph version of the identity

$$\log 2 - \sum_{k \geq 1} \frac{1}{k} \left( \mathbf{E}\, \tilde{p}_k^{G_n}(\rho, \rho) - \frac{1}{n} \right) = - \sum_{k \geq 1} \frac{1}{k} \left( \mathbf{E}\, p_k^{G_n}(\rho, \rho) - \frac{1}{n} \right)$$

that we get for any finite graph $G_n$ by comparing (14.7, 14.8) with (14.10). □

{ex.pkqk}

▷ **Exercise 14.34.** *Let $P$ be the transition matrix of any infinite Markov chain. For $\alpha \in [0,1]$, define the lazy transition matrix $Q := \alpha I + (1-\alpha) P$. For a state $x$, let $p_k(x)$ and $q_k(x)$ denote the return probabilities to $x$ after $k$ steps in the two chains. Then $\sum_{k \geq 1} q(x)/k = -\log(1-\alpha) + \sum_{k\geq 1} p(x)/k$. (Hint: For any $z \in (0,1)$, write $\sum_k q_k(x) z^k / k$ as an inner product using the operator $\log(I - zQ)$, then let $z \nearrow 1$.)*

The tree entropy $\lim_n \mathbf{h}_{\mathrm{tree}}(G_n) = \mathbf{h}_{\mathrm{tree}}(G, \rho)$ can actually be calculated sometimes. From the connection between spanning trees and domino tilings in planar bipartite graphs, one can show that $\mathbf{h}_{\mathrm{tree}}(\mathbb{Z}^2) = 4G/\pi \approx 1.166$, where $G := \sum_{k \geq 0} (-1)^k/(2k+1)^2$ is Catalan's constant; see the examples after Theorem 6.1 of [BurtP93]. For the 4-regular tree, from Theorem 14.13 and the tree's Green function (1.9), Lyons deduced $\mathbf{h}_{\mathrm{tree}}(\mathbb{T}_4) = 3 \log(3/2)$ in [Lyo05].

A similarly defined notion of entropy is the **$q$-colouring entropy** $\mathbf{h}_{q\text{-col}}(G) := \frac{\log \mathsf{ch}(G,q)}{|V(G)|}$, where $\mathsf{ch}(G,q)$ is number of proper colourings of $G$ with $q$ colours (i.e., colourings of $V(G)$ such that neighbours never get the same colour). This $\mathsf{ch}(G,q)$ is called the **chromatic polynomial** of $G$, for the following reason:

{ex.chpoly}

▷ **Exercise 14.35.**

- **(a)** *Show that, for any $q \in \mathbb{Z}_+$ and any $e \in E(G)$, we have $\mathsf{ch}(G,q) = \mathsf{ch}(G \setminus e, q) - \mathsf{ch}(G/e, q)$, where $G \setminus e$ is the graph obtained from $G$ by deleting $e$, and $G/e$ is obtained by gluing the endpoints of $e$ and erasing the resulting loops.*
- **(b)** *Deduce that $\mathsf{ch}(G,q)$ is a polynomial in $q$, of the form $q^n + a_{n-1}(G) q^{n-1} + \cdots + a_1(G) q$, where $|V(G)| = n$.*

**(c)** *Show that* $\sum_{S\subseteq V(G)} \mathsf{ch}(G[S],x)\,\mathsf{ch}(G[V\setminus S],y) = \mathsf{ch}(G,x+y)$, *where* $G[S]$ *is the subgraph of* $G$ *induced by* $S$.

The roots of the chromatic polynomial $\mathsf{ch}(G,q) = \prod_{i=1}^{n}(q-\lambda_i)$ are called the **chromatic roots** of $G(V,E)$, and (analogously to the spectral measure) the counting measure on the roots normalized to have total mass 1 is called the **chromatic measure** $\mu_G^{\mathrm{col}}$. We can express the $q$-colouring entropy using this measure:

$$\mathbf{h}_{q\text{-col}}(G) = \frac{1}{n}\log \mathsf{ch}(G,q) = \int_{\mathbb{C}} \log(q-z)\, d\mu_G^{\mathrm{col}}(z)\,, \tag{14.11}$$ {e.qcolorint}

whenever $q \notin \{\lambda_1,\dots,\lambda_n\}$. Note that the 4-colour theorem says that any planar graph $G$ satisfies $\mathsf{ch}(G,q) > 0$ for all integers $q \geq 4$. A certain generalization has been proved by Alan Sokal [Sok01]: there exists an absolute constant $C < 8$, such that if $G$ has maximal degree $d$, then all roots of $\mathsf{ch}(G,q)$ in $\mathbb{C}$ are contained in the ball of radius $Cd$. There is a huge body of work on the chromatic polynomial by Sokal and coauthors (just search the arXiv).

What about locality? It was proved in [BorChKL13] that $\mathbf{h}_{q\text{-col}}(G)$ is local for graphs with degrees bounded by $d$, whenever $q > 2d$. They used the **Dobrushin uniqueness theorem**: with this many colours, the effect of any boundary condition decays exponentially fast with the distance, and hence there is only one Gibbs measure in any infinite volume limit (i.e., on any limiting unimodular random rooted graph). It should not be surprising that this has to do with the locality of how many different colourings are possible, but the proof is somewhat tricky.

Now, in light of (14.11) and the locality of the random walk spectral measure (Exercise 14.26), it is natural to ask about the locality of the chromatic measure. This was addressed by Abért and Hubai [AbH15]; then Csikvári and Frenkel [CsF16] found a simpler argument that generalizes to a wide class of graph polynomials. We start with the definitions needed.

A graph polynomial $f(G,z) = \sum_{k=0}^{n} a_k(G) z^k$ is called **isomorphism invariant** if it depends only on the isomorphism class of $G$. It is called **multiplicative** if $f(G_1 \sqcup G_2, z) = f(G_1,z)\,f(G_2,z)$, where $\sqcup$ denotes disjoint union. It is of **exponential type** if it satisfies the identity of Exercise 14.35 (c). It is of **bounded exponential type** if there is a function $R : \mathbb{N} \longrightarrow \mathbb{R}_{\geq 0}$ not depending on $G$ such that, for any graph $G$ with all degrees at most $d$, any $v \in V(G)$, and any $t \geq 1$, we have

$$\sum_{\substack{v\in S\subseteq V(G)\\ |S|=t}} |a_1(G[S])| \leq R(d)^{t-1}. \tag{14.12}$$ {e.boundedexp}

Besides the chromatic polynomial, here are some further examples that satisfy all these properties (see [CsF16] for proofs and references):

(1) The **Tutte polynomial** of $G(V,E)$ is defined as

$$T(G,x,y) := \sum_{\omega\subseteq E} (x-1)^{k(\omega)-k(E)}(y-1)^{k(\omega)+|\omega|-|V|}\,, \tag{14.13}$$ {e.Tutte}

where $k(\omega)$ is the number of connected components of $(V,\omega)$. This is almost the same as the partition function $Z_{\mathsf{FK}}(p,q)$ of the FK model in (13.7); the exact relation is

$$T(G,x,y) = \frac{y^{|E|}}{(x-1)^{k(E)}(y-1)^{|V|}}\, Z_{\mathsf{FK}}\left(\frac{y-1}{y},\,(x-1)(y-1)\right). \tag{14.14}$$ {e.TutteFK}

A third common version is $F(G,q,v) := \sum_{\omega\subseteq E} q^{k(\omega)} v^{|\omega|}$. The variable $v$ corresponds to $p/(1-p)$ in $\mathsf{FK}(p,q)$. This third form is the best now, since its degree in $q$ is $|V|$. In fact, it satisfies the above properties for any fixed $v$. Note that $\mathsf{ch}(G,q) = F(G,q,-1)$.

(2) The **Laplacian characteristic polynomial** $L(G,z)$ is the characteristic polynomial of the Laplacian matrix $L_G = D_G - A_G$, featured in the proof of Theorem 14.13.

(3) The **(modified) matching polynomial** is $M(G,z) := z^n - m_1(G)z^{n-1} + m_2(G)z^{n-2} - m_3(G)z^{n-3} + \dots$, where $m_k(G)$ is the number matchings of size $k$.

Sokal's theorem above about the chromatic roots of bounded degree graphs generalizes rather easily to graph polynomials of bounded exponential type: Theorem 1.6 of [CsF16] says that if $f(G,z)$ has a bounding function $R$ in (14.12), and $G$ has degrees less than $d$, then the absolute value of any root is less than $cR(d)$, where $c < 7.04$. In other words, the polynomial then has **bounded roots**. Now, the main result of [CsF16], generalizing the case of chromatic polynomials from [AbH15] is the following:

**Theorem 14.14** (Csikvári-Frenkel [CsF16])**.** *Let $f(G,z)$ be an isomorphism-invariant monic multiplicative graph polynomial of exponential type. Assume that it has bounded roots. Let $G_n$ be a sequence that converges in the Benjamini-Schramm sense, and $K \subset \mathbb{C}$ a compact domain that contains all the roots of $f(G_n,z)$. Let $\mu_G^f$ be the uniform distribution on the roots of $f(G,z)$. Then the holomorphic moments $\int_K z^k d\mu_{G_n}^f(z)$ converge for all $k \in \mathbb{N}$. In particular, if we define the $f$-**entropy** or **free energy** at $\zeta \in \mathbb{C} \setminus K$ by*

$$\mathbf{h}_{f,\zeta}(G) := \frac{1}{|V(G)|} \log |f(G,\zeta)| = \int_{\mathbb{C}} |\log(\zeta - z)|\, d\mu_G^f(z)\,, \tag{14.15}$$ {e.fentint}

*then the Taylor series of $\log|z|$ shows that $\mathbf{h}_{f,\zeta}(G_n)$ converges to a harmonic function locally uniformly on $\mathbb{C} \setminus K$. That is, for $\zeta \in \mathbb{C} \setminus K$, the $f$-entropy at $\zeta$ is local.*

As opposed to moments of a compactly supported measure on $\mathbb{R}$, the holomorphic moments do not characterize uniquely a compactly supported measure on $\mathbb{C}$. And, somewhat surprisingly, the chromatic measure itself is in fact not local, as shown by the following exercise:

{ex.chcycle}

▷ **Exercise 14.36.**

**(a)** *Show that the chromatic polynomial of the path on $n$ vertices is $\mathsf{ch}(P_n,q) = q(q-1)^{n-1}$, while the chromatic polynomial of the cycle on $n$ vertices is $\mathsf{ch}(C_n,q) = (q-1)^n + (-1)^n(q-1)$.*

**(b)** *Deduce that the chromatic measure of $C_n$ converges weakly to the uniform distribution on the circle of radius 1 around $z=1$, while the chromatic measure of $P_n$ converges weakly to the point mass at $z=1$.*

**(c)** *** *Is one of the limiting measures more canonical than the other?*

Non-trivially, but it follows from the locality of the matching polynomial root moments that the **matching ratio**, i.e., the maximal size of a disjoint set of edges, divided by the number of vertices is a local parameter. See also [ElL10]. Surprisingly at first sight, the analogous notion with independent subsets of vertices instead of edges behaves very differently. The **independence ratio** of a finite graph $G(V,E)$ is the size of the largest independent set (i.e., a subset of the vertices such that no two of them are neighbours) divided by $|V|$. For instance, the independence ratio of a balanced bipartite graph (i.e., the two parts have the same size) is $1/2$.

{t.indepratio}

**Theorem 14.15** (The independence ratio is not local [Bol81])**.** *For any $d \geq 3$, there exists an $\epsilon > 0$ and a sequence of $d$-regular graphs with girth tending to infinity (i.e., converging locally to $\mathbb{T}_d$) for which the independence ratio is less than $1/2 - \epsilon$. Basically, random $d$-regular graphs will do. Since uniformly random $d$-regular balanced bipartite graphs also converge to $\mathbb{T}_d$ in the local weak sense (see Exercise 14.8), this shows that the independence ratio of $d$-regular graphs is not local.*

What is then the independence ratio of random $d$-regular graphs? This seems to be intimately related to the question of how dense an invariant independent set can be defined on the $d$-regular tree $\mathbb{T}_d$ that is a factor of an i.i.d. process. One direction is clear: given any measurable function of an i.i.d. process on $\mathbb{T}_d$ that produces an independent set, we can approximate that by functions depending only on a bounded neighbourhood, then we can apply the same local functions on any sequence of finite graphs with girth going to infinity. Several people (Balázs Szegedy, Endre Csóka, maybe David Aldous) have independently arrived at the conjecture that on random $d$-regular graphs, the optimal density (in fact, all possible limit densities) can be achieved by such a local construction:

**Conjecture 14.16** (Balázs Szegedy)**.** *The possible values for the densities of independent sets in random $d$-regular graphs coincides with the possible densities of independent sets as i.i.d. factors in $d$-regular trees. (This is part of a much more general conjecture on the so-called* **local-global limits** *that I plan to discuss in a later version; see [HatLSz12] for now.)*

The ideology is that random $d$-regular graphs have no global structure, hence all limit processes can be constructed locally (i.e., as a factor of i.i.d.) on the local limit graph (the regular tree).

It is an important question what kind of processes can be realized as a factor of an i.i.d. process. See [LyNaz11, HatLSz12, Mes11], for instance. And here is another beautiful conjecture:

{c.AbSzeg}

**Conjecture 14.17** (Abért-Szegedy)**.** *Let $G_n$ be a Benjamini-Schramm-convergent sequence of finite graphs, $\omega_n$ an i.i.d. process on $G_n$, and $F(G_n, \omega_n)$ a factor process. Let $\mathbf{h}(F, G_n)$ be the entropy of the resulting measure. (For simplicity, we can assume that there are finitely many possible configurations of $F(G_n, \omega_n)$.) Then $\lim_{n\to\infty} \mathbf{h}(F, G_n)/|G_n|$ exists.*

We have mentioned in Section 13.5 that Ornstein-Weiss have developed a very good entropy theory for amenable groups. An affirmative answer to Conjecture 14.17 would say that there is a good entropy notion also for i.i.d. factor processes on all sofic groups.

Another interesting corollary would be the (mod $p$) version of the Lück Approximation Theorem 14.11: if $A_n$'s are the adjacency matrices of a convergent sequence of finite graphs $G_n$, then $\dim \ker_{\mathbb{F}_p} A_n/|V(G_n)|$ converges. Indeed, this is equal to $1 - \dim \mathrm{Im}_{\mathbb{F}_p} A_n/|V(G_n)|$, and the normalized dimension of the image, times $\log p$, is the normalized entropy of the uniform measure on the image space. And we can easily get this uniform distribution as a factor of an i.i.d. process: assign i.i.d. uniform random labels to the vertices of $G_n$ from $\{0, 1, \dots, p-1\}$, then write on each vertex the (mod $p$) sum of its neighbouring labels.

Let us turn to a locality question that is quite different from spectral measures and entropies: the critical parameter in Bernoulli percolation:

{c.pcloc}

**Conjecture 14.18** (Locality of $p_c$, O. Schramm)**.** *If $G_n$ are infinite transitive graphs locally converging to $G$, with $\sup_n p_c(G_n) < 1$, then $p_c(G_n) \to p_c(G)$.*

▷ **Exercise 14.37.** *Show that the $\sup_n p_c(G_n) < 1$ condition in Conjecture 14.18 is necessary.*

This conjecture appears to be quite hard: the case of $\mathbb{Z}^d$ was proved in [GriMar90], one of the key papers of classical percolation. This was recently extended by [MaTa17] to Cayley graphs of Abelian groups, by making the Grimmett-Marstrand argument more robust. At the other end, for highly non-amenable transitive graphs with high girth, the conjecture was proved in [BenNP11].

We are going to prove now one direction of Conjecture 14.18, which is clearly the easier one, but we are not aware that it has been done anywhere else:

$$p_c(G) \le \liminf_{n\to\infty} p_c(G_n)\,. \tag{14.16}$$ {e.locinf}

Assume that this fails, hence there exists $\epsilon > 0$ such that $p_c(G_n) < p_c(G) - \epsilon$ for infinitely many $n$. For easier notation, we relabel this subsequence to be $\{G_n\}_{n=1}^\infty$. Now, fixing $p = p_c(G) - \epsilon/2$, for any $\delta > 0$ there exists a large radius $R > 0$ such that $\mathbf{P}_p^G[o \longleftrightarrow \partial B_R(o)] < \delta$. If $n$ is large enough, then $B_R^G(o) \simeq B_R^{G_n}(o)$, hence $\mathbf{P}_p^{G_n}[o \longleftrightarrow \partial B_R(o)] < \delta$, as well. On the other hand, $p > p_c(G_n) + \epsilon/2$, and $p_c(G_n) < 1 - \epsilon$, hence Theorem 12.29 on the linear lower bound for $\theta(p_c + \epsilon)$ says that $\mathbf{P}_p^{G_n}[o \longleftrightarrow \infty] > c(\epsilon)$ for all $n$. Taking $\delta > 0$ less than $c(\epsilon)$, then $n$ large enough, we have arrived at a contradiction.

For the other direction, one would need to show that the existence of $p$ with $p_c(G) < p < p_c(G_n) < 1 - \delta$ for infinitely many $n$ is impossible. I imagine there would be two main steps in this (not at all independent of each other). Firstly, the content of Questions 12.34 and 12.35 in Section 12.5 is that doing percolation at $p > p_c(G)$ should be somehow visible for large enough $R$ in the cluster structure restricted to $B_R^G(o)$. In the case of amenable $G$, it is probably better to take a large Følner set, and in that case, the Følner set should be really well-connected internally. Secondly, for large enough $n$, we have $B_R^G(o) \simeq B_R^{G_n}(o)$, so one could try to show that it is impossible to glue these large balls or Følner sets globally in a way that the quite large clusters inside these large sets do not get connected, forming an infinite cluster. Here the condition $p_c(G_n) < 1 - \delta$ is really important: the graph $\mathbb{Z}_n \times \mathbb{Z}$ has large large well-connected pieces at scale $n$, provided that $p > 1/2$ and $n > n_0(p)$, but they do not form globally an infinite cluster. However, the condition $p_c(G_n) < 1 - \delta$ should imply that the "renormalized graph" that has the large balls or Følner sets as vertices, together with some natural neighbor structure, should have some version of $p_c$ that is

(a) either close to zero (e.g., large balls of a regular tree form a regular tree of huge degree),

(b) or is at least still bounded away from 1, but this weaker case should occur only for amenable graphs, where the large sets should be really well-connected internally with high probability (as in renormalization on Abelian graphs), again producing a super-critical situation globally.

A joint corollary to Conjecture 14.18 on locality and Conjecture 12.7 on $p_c < 1$ would be that 1 is an isolated point in each of the sets $\mathcal{P}_d := \{p_c(G) : G \text{ is an infinite } d\text{-regular transitive graph}\}$. Indeed, if we had a sequence of $d$-regular graphs with $p_c(G_n) < 1$ but $p_c(G_n) \to 1$, then by compactness there would be a subsequence $G_{n_k}$ converging locally to some $G$. By locality of $p_c$, we would have $p_c(G) = 1$. By Conjecture 12.7, this $G$ would be a finite extension of $\mathbb{Z}$. Then, by Exercise 14.14, there is a finite $n_k$ such that $G_{n_k} = G$, hence $p_c(G_{n_k}) = 1$, a contradiction. One expects that 1 is also isolated in $\bigcup_{d \in \mathbb{Z}_+} \mathcal{P}_d$, since with large degrees it seems harder to get large $p_c$ values, but I do not see a proof for this stronger version.

Of course, one can also ask about the locality of $p_c(q)$ in the $\mathsf{FK}(p,q)$ random cluster measures. But there is an even more basic question:

**Question 14.19.** *If $G_n$ converges to a non-amenable transitive $G$ in the local weak sense, is it true for all $\mathsf{FK}(p,q)$ random cluster models (especially for the much more accessible $q > 1$ case) that the limit measure from the $G_n$'s is the wired measure on $G$?*

This is easy to see for the $\mathsf{WUSF}$ (the $q = p = 0$ case) using Wilson's algorithm. For the $q = 2$ Ising case, when $G$ is a regular tree, it is proved in [MonMS12] that the Ising measure converges to the symmetric mixture of the $+$ and $-$ extremal measures, which strongly suggests a positive answer. In the amenable case, it is clear that both the free and the wired measures can be achieved by local limits. In the non-amenable case, the question is if the limit could be strictly dominated by the wired measure.

# 15 Some more exotic groups

{s.exotic}

We now present some less classical constructions of groups, which have played a central role in geometric group theory in the last two decades or so, and may provide or have already provided exciting examples for probability on groups.

## 15.1 Self-similar groups of finite automata

{ss.automata}

This section shows some natural ways to produce group actions on rooted trees, which come up independently in complex and symbolic dynamical systems and computer science. A key source of interest in this field is Grigorchuk's construction of groups of intermediate volume growth (1984). The main references are [GriNS00, BartGN03, Nek05].

In Section 13.5 we already encountered the adding machine action of $\mathbb{Z}$: the action of the group on finite and infinite binary sequences was

$$\begin{aligned}(0w)^a &= 1w \\ (1w)^a &= 0w^a\,.\end{aligned} \tag{15.1}$$ {e.adding2}

By representing finite and infinite binary sequences as vertices and rays in a binary tree, we clearly get an action by tree automorphisms. The first picture of Figure 15.1 shows this action, with the Schreier graphs of the first few levels. The second picture is called the "profile" of the action of the generator $a$; the picture should be self-explanatory once one knows that the switches of the subtrees are to be read off from bottom to top, towards the root.

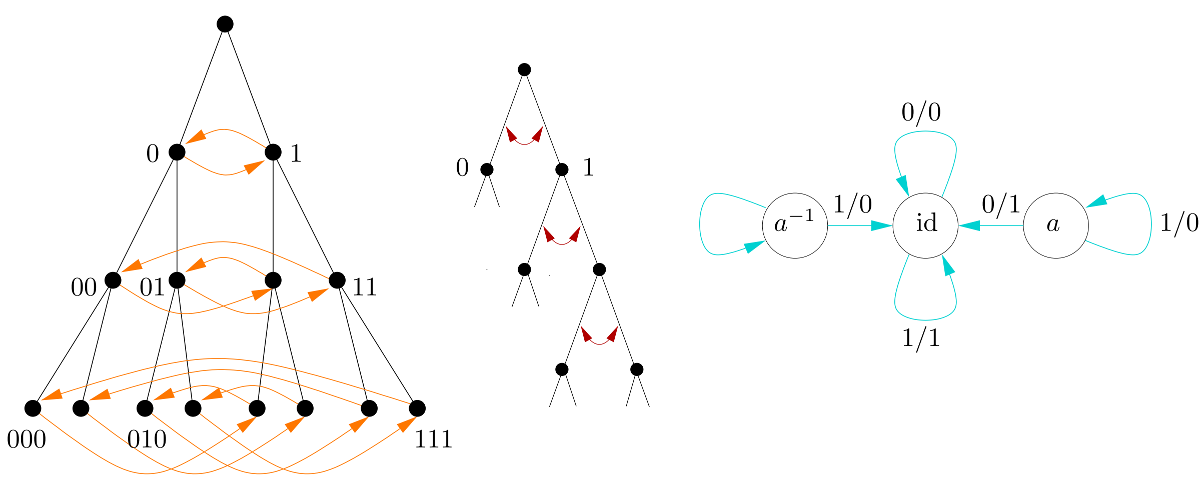



Figure 15.1: The adding machine action of $\mathbb{Z}$: (a) the Schreier graphs on the levels (b) the profile of the generator $a$'s action (c) the Moore diagram of the automaton

{f.adding}

Finally, the third picture of Figure 15.1 is called the **Moore diagram** of the **automaton** generating the group action. This automaton is a directed labeled graph $G(V,E)$, whose vertices (the states) correspond to some tree automorphisms, as follows. Given any initial state $s_0 \in V$, for any infinite binary sequence $x_1x_2\dots$ the automaton produces a new sequence $y_1y_2\dots$: from $s_0$ we follow the arrow whose first label is $x_1$, we output the second label, this will be $y_1$, and then we continue with the target state of the arrow (call it $s_1$) and the next letter $x_2$, and so on. The group generated by the tree automorphisms given by the states $V$ is called the group generated by the automaton. There is of course a version with labels from $\{0, 1, \ldots, b-1\}$ instead of binary sequences, generating an action on the $b$-ary tree $\mathbb{T}_b$.

{d.ss}

**Definition 15.1.** *The action of a group $\Gamma$ on the $b$-ary tree $\mathbb{T}_b$ ($b \geq 1$) is called* **self-similar** *if for any $g \in \Gamma$, any letter $x \in \{0, 1, \ldots, b-1\}$, and any finite or infinite word $w$ on this alphabet, there is a letter $y$ and $h \in \Gamma$ such that $(xw)^g = y(w^h)$. If $S \subseteq \Gamma$ generates $\Gamma$ as a semigroup, and $\forall s \in S$ and word $xw$ there is a letter $y$ and $t \in S$ such that $(xw)^s = y(w^t)$, then $S$ is called a* **self-similar generating set***. Then the group can clearly be generated by an automaton with states $S$. If there is a finite such $S$, then $\Gamma$ is called a* **finite-state self-similar group***.*

For a self-similar action by $\Gamma$, for any $g \in \Gamma$ and finite word $v$ there is a word $u$ of the same length and $h \in \Gamma$ such that $(vw)^g = u(w^h)$ for any word $w$. This $h$ is called the **restriction** $h = g|_v$, and we get an action of $\Gamma$ on the subtree starting at $v$. The action of the full automorphism group of $\mathbb{T}_b$ is of course self-similar, and there is the obvious wreath product decomposition

$$\mathrm{Aut}(\mathbb{T}_b) \simeq \mathrm{Aut}(\mathbb{T}_b) \wr \mathrm{Sym}_b\,, \tag{15.2}$$ {e.treathA}

corresponding to the restriction actions inside the $b$ subtrees at the root and then permuting them. (But, of course, $\mathrm{Aut}(\mathbb{T}_b)$ is not finitely generated.) For a general self-similar action by $\Gamma \leq \mathrm{Aut}(\mathbb{T}_b)$, the isomorphism (15.2) gives an embedding

$$\Gamma \hookrightarrow \Gamma \wr \mathrm{Sym}_b\,. \tag{15.3}$$ {e.treathG}

Using this embedding, we can write the action of any $g \in \Gamma$ recursively as $(g_0, \ldots, g_{b-1})\pi_g$, where $g_i = g|_i$ are the restrictions to the children at the first level, $\pi_g$ is the permutation of these children (together with the subtrees below them), and the order of the multiplication is dictated by having a right action on the tree: first apply the restrictions, then apply $\pi_g$. For instance, the adding machine action (13.16) can be written very concisely as the recursion $a = (1, a)\epsilon$, where $\epsilon$ is the involution switching the two subtrees starting at the first level.

There is also a nice geometric way of arriving at the adding machine action of $\mathbb{Z}$ (and other self-similar actions). Consider $\varphi : x \mapsto 2x$, an expanding automorphism of the Lie group $(\mathbb{R}, +)$. Since $\varphi(\mathbb{Z}) = 2\mathbb{Z} \subseteq \mathbb{Z}$, we can consider $\varphi : \mathbb{R}/\mathbb{Z} \longrightarrow \mathbb{R}/\mathbb{Z}$, a twofold self-covering of the circle $S^1$. Pick a base point $x \in S^1$, and a loop $\gamma : [0, 1] \longrightarrow S^1$ starting and ending at $x$, going around $S^1$ once. Now, $\varphi^{-1}(x)$ consists of two points, call them $x_0$ and $x_1$. Using Proposition 2.11, we can lift $\gamma$ starting from either point, getting $\gamma_0$ and $\gamma_1$, respectively. Clearly, $\gamma_i$ ends at $\gamma_{1-i}$, $i = 0, 1$. Following these $\gamma_i$'s we get a permutation on $\varphi^{-1}(x)$, the transposition $(x_0 x_1)$ in the present case. (For a general, possibly branched, $b$-fold covering $\varphi : X \longrightarrow X$, we should start with one $\gamma$ for each generator of $\pi_1(X)$, and then would get an action of $\pi_1(X)$ on $\varphi^{-1}(x)$, as in Section 2.2.) We can now iterate this procedure, taking the preimage set $\varphi^{-2}(x) = \varphi^{-1}(x_0) \cup \varphi^{-1}(x_1) = \{x_{00}, x_{01}, x_{10}, x_{11}\}$, etc., and we get an action of $\pi_1(S^1) = \mathbb{Z}$ on the entire binary tree. It is easy to see that the lifts $\gamma_0$, $\gamma_1$, $\gamma_{00}$ etc. will be geometric representations of the edges in the Schreier graph of the adding machine action. For a general $b$-fold covering, it is possible that the resulting action on the $b$-ary tree is not faithful, so the actual group of tree-automorphisms that we get will be a factor of $\pi_1(X)$. This is called the **Iterated Monodromy Group** $\mathsf{IMG}(\varphi)$ of the covering map.

A particularly nice case is when $X$ is a Riemannian manifold, and $\varphi : X_1 \longrightarrow X$ is a locally expanding partial $b$-fold self-covering map for some $X_1 \subseteq X$ (we typically get $X_1$ by removing the branch points from $X$ in the case of a branched covering). Then the resulting action of $\Gamma = \mathsf{IMG}(\varphi)$ on $\mathbb{T}_b$ is finite state self-similar, moreover, **contracting**: there is a finite set $\mathcal{N} \subset \Gamma$ such that for every $g \in \Gamma$ there is a $k \in \mathbb{N}$

such that the restriction $g|_v$ is in $\mathcal{N}$ for all vertices $v \in \mathbb{T}_b$ of depth at least $k$. Moreover, if in addition, $X_1 = X$ is a compact Riemannian manifold, like $X = S^1$ in our running example, then the action of $\pi_1(X)$ is faithful, so $\mathsf{IMG}(\varphi) = \pi_1(X)$. See [Nek03] or [Nek05] for proofs.

{ex.nucleus}

▷ **Exercise 15.1.** *Show that, for any contracting self-similar action of some $\Gamma$ on $\mathbb{T}_b$, there is a unique minimal set $\mathcal{N} \subseteq \Gamma$ giving the contraction property. It is called the* **nucleus** *of the self-similar action. Show that $\mathcal{N}$ is a self-similar generating set of $\langle \mathcal{N} \rangle$. Give an example where $\langle \mathcal{N} \rangle \neq \Gamma$.*

At this point, the Reader might feel a bit uneasy: we have been talking about countable groups acting on trees, so where do these continuum objects $\mathbb{R}$ and $S^1$ and Riemannian manifolds come from? Well, their appearance in the story is not accidental: it is possible to reconstruct $X$ and $\varphi$, at least topologically, from the action itself. (Or, in less perfect situations when $\varphi$ is not expanding and the resulting tree-action is not contracting, some $\varphi$-invariant subset of $X$, the Julia set of $\varphi$.) Given any contracting action by $\Gamma$ on $\mathbb{T}_b$, one can define the **limit space** $\mathcal{J}_\Gamma$ as the quotient of the set of left-infinite sequences $\{0, \dots, b-1\}^{-\mathbb{N}}$ by the following **asymptotic equivalence** relation: the sequences $(\dots, x_{-1}, x_0)$ and $(\dots, y_{-1}, y_0)$ are equivalent iff there exists a finite subset $K \subset \Gamma$ such that for all $k \in \mathbb{N}$ there is some $g_k \in K$ with $(x_{-k}, \dots, x_0)^{g_k} = (y_{-k}, \dots, y_0)$ with the action of $\Gamma$ on $\mathbb{T}_b$, i.e., with $x_{-k}$ on the first level, $(x_{-k}, x_{-k+1})$ on the second level, etc. (In particular, this equivalence is very different from two rays in $\partial\mathbb{T}_b = \{0, \dots, b-1\}^{\mathbb{N}}$ being in the same $\Gamma$-orbit.) A similar notion is the **limit solenoid** $\mathcal{S}_\Gamma$, the quotient of $\{0, \dots, b-1\}^{\mathbb{Z}}$ by the equivalence relation that $(\dots, x_{-1}, x_0, x_1, \dots) \sim (\dots, y_{-1}, y_0, y_1, \dots)$ iff there is a finite $K \subset \Gamma$ such that $\forall k \in \mathbb{N}\ \exists g_k \in K$ with $(x_{-k}, x_{-k+1}, \dots)^{g_k} = (y_{-k}, y_{-k+1}, \dots)$ in $\partial\mathbb{T}_b$. On $\mathcal{J}_\Gamma$ we take the topology to be the image of the product topology under the equivalence quotient map, while on $\mathcal{S}_\Gamma$ it will be the image of the topology on $\{0, \dots, b-1\}^{\mathbb{Z}}$ that is product topology on the left tail but discrete on the right; to emphasize the asymmetric topology, we can denote this space by $\mathcal{S}_\Gamma^{\triangleleft}$. Now, given a finite word $w = x_k x_{k+1} \dots x_0$, with $k \in -\mathbb{N}$, we can consider the **tile**

$$T_w := \big\{ \dots x_{k-2} x_{k-1} w : x_{k-i} \in \{0, 1, \dots, b-1\}, i \geq 1 \big\},$$

a subset of $\mathcal{J}_\Gamma$ after the factorization. Similarly, given an infinite word $w = x_k x_{k+1} \dots$, $k \in \mathbb{Z}$, the tile $T_w$ will be a subset of $\mathcal{S}_\Gamma^{\triangleleft}$. Because of the factorization, these tiles are not at all disjoint for different $w$'s with the same starting level $k = k(w)$. However, if the action of $\Gamma$ is nice enough, then their interiors are disjoint, and the intersections of the boundaries are given by the action of $\mathcal{N}$: $T_{w^g} \cap T_{w^h} \neq \emptyset$ iff $g^{-1}h \in \mathcal{N}$. In particular, if $\langle \mathcal{N} \rangle = \Gamma$, then the adjacencies are given by the Schreier graph on that level. In $\mathcal{J}_\Gamma$, as we take the level $k \to -\infty$, the tiles are getting smaller and smaller, and hence the Schreier graphs, drawn on the tiles, approximate the structure of $\mathcal{J}_\Gamma$ more and more. The situation in $\mathcal{S}_\Gamma^{\triangleleft}$ is a bit more complicated: it is a highly disconnected space, so we need to restrict our attention to the **leaves** $\mathcal{L}_{O(w)} := \bigcup_{g \in \Gamma} T_{w^g} \subseteq \mathcal{S}_\Gamma^{\triangleleft}$ corresponding to $\Gamma$-orbits $O(w)$ in $\partial\mathbb{T}_b$. Finally, consider the shift action $\mathsf{s}$ on $\{0, \dots, b-1\}^{-\mathbb{N}}$ that deletes the last (the $0^{\text{th}}$) letter, hence is $b$-to-1, or that moves the origin in $\{0, \dots, b-1\}^{\mathbb{Z}}$ to the left. In both cases, $\mathsf{s}$ preserves the asymptotic equivalence relation, and thus we get the dynamical systems $(\mathcal{J}_G, \mathsf{s})$ and $(\mathcal{S}_\Gamma^{\triangleleft}, \mathsf{s})$.

The upshot (proved by Nekrashevych) is that when the contracting action is obtained from a locally expanding map $\varphi : X_1 \longrightarrow X$, then, under mild additional assumptions, $(\mathcal{J}_G, \mathsf{s})$ is topologically conjugate to $(\mathcal{J}(X, \varphi), \varphi)$ (i.e., there is a homemorphism between the spaces that interchanges the actions), where

$$\mathcal{J}(X, \varphi) := \left\{ \text{accumulation points of } \bigcup_{n=0}^{\infty} \varphi^{-n}(x) \right\} \subseteq X$$

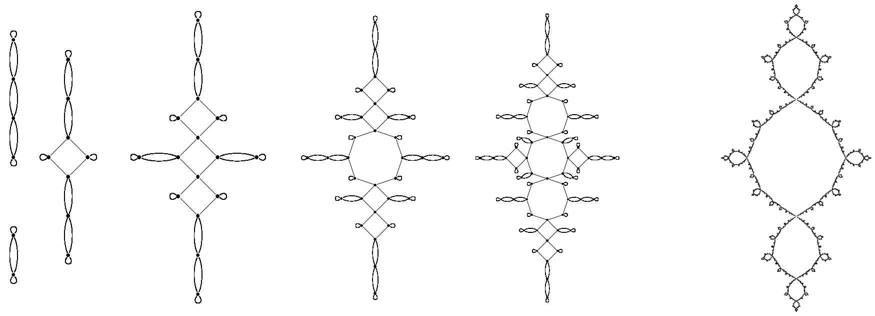


Figure 15.2: Schreier graphs of the Basilica group, and the Julia set in $\mathbb{C}$. Pictures taken from [DDMN10].

is the **Julia set** of $\varphi$, easily shown to be independent of $x \in X$. For instance, if $X_1 = X$ is a compact Riemannian manifold, then $\mathcal{J}(X, \varphi) = X$. Recall that $\pi_1(X) = \mathsf{IMG}(\varphi)$ in this case, so, $\Gamma = \mathsf{IMG}(\varphi)$ and the limit space construction $X = \mathcal{J}_\Gamma$ are true inverses of each other.

▷ **Exercise 15.2.** *For the adding machine action, give a homeomorphism between $\mathcal{J}_\mathbb{Z}$ and $S^1$ that interchanges the actions $x \mapsto 2x$ on $S^1$ and $\mathsf{s}$ on $\mathcal{J}_\mathbb{Z}$. Show that for any $w \in \partial\mathbb{T}_2$, the leaf $\mathcal{L}_{O(w)}$ is homeomorphic to $\mathbb{R}$.*

▷ **Exercise 15.3.** *Show that the limit set $\mathcal{J}_\Gamma$ for the self-similar action*

$$a = (a, 1, 1)(12) \qquad b = (1, b, 1)(02) \qquad c = (1, 1, c)(01)$$

*on the alphabet $\{0, 1, 2\}$ is homeomorphic to the Sierpiński gasket.*

▷ **Exercise 15.4** (Basilica group)**.** * {ex.Basil}

**(a)** *Consider the complex polynomial $\varphi(z) = z^2 - 1$ as a covering map from $X_1 = \hat{\mathbb{C}} \setminus \{1, 0, -1, \infty\}$ to $X = \hat{\mathbb{C}} \setminus \{0, -1, \infty\}$. Show that it is locally expanding in a neighbourhood of its Julia set $\mathcal{J}(X, \varphi)$.*

**(b)** *Show that $\mathsf{IMG}(\varphi)$ is the self-similar group $\mathsf{Bas}$ generated by the finite automaton*

$$a = (1, b)\,, \qquad b = (1, a)\epsilon\,. \tag{15.4}$$ {e.Basil}

By the last exercise, the theorem of Nekrashevych mentioned above applies to the Basilica group $\mathsf{Bas} = \mathsf{IMG}(z^2 - 1)$: the limit space $\mathcal{J}_{\mathsf{Bas}}$ is homemorphic to the Julia set $\mathcal{J}(X, \varphi)$, and in fact, the Julia set could be considered as a scaling limit of the Schreier graphs of $\mathsf{Bas}$ on the levels. (Has it been done, metrically, anywhere?)

Expanding endomorphisms of the real and integer Heisenberg group $H_3(\mathbb{R})$ and $H_3(\mathbb{Z})$ were used in [NekP09] to produce nice contracting self-similar actions, with the tiles and the shift map $\mathsf{s}$ leading to scale-invariant tilings in some Cayley graphs of $H_3(\mathbb{Z})$, as defined in Question 12.33. The same paper used the self-similar actions of several groups of exponential growth to show that they are scale-invariant (see Theorem 4.17): the lamplighter group $\mathbb{Z}_2 \wr \mathbb{Z}$, the solvable Baumslag-Solitar groups $\mathsf{BS}(1, m)$, and the affine groups $\mathbb{Z}^d \rtimes \mathsf{GL}(d, \mathbb{Z})$.

When a group turns out to have a finite state self-similar action, a huge box of great tools opens up — but it is far from obvious if a given group has such an action. The **lamplighter group** $\mathsf{G} = \mathbb{Z}_2 \wr \mathbb{Z}$ is a famous unexpected example, whose self-similarity was first noticed and proved by Grigorchuk and Żuk in [GriŻ01], but there is a much simpler proof, found by [GriNS00], using $\mathbb{Z}_2[[t]]$, which we will present below. The self-similar action of $\mathsf{G}$ was used in [GriŻ01] to show that the spectrum of the simple random walk on the Cayley graph generated by the self-similar generators $a$ and $b$ below is discrete — see Section 14.2, just before Conjecture 14.12, for a bit more on this. The scale-invariance of $\mathsf{G}$ proved in [NekP09] is closely related to the way how the spectrum was computed, even though the way we found it followed an orthogonal direction of thought, namely, that the **Diestel-Leader graph** $\mathsf{DL}(2,2)$ is the Cayley graph of $\mathsf{G}$ with the generators $\langle R, Rs\rangle$ on one hand, and of the index two subgroup $\Gamma_1 = \langle Rs, sR\rangle$ on the other. We now show $\mathsf{G}$ can be generated by a finite automaton.

Let $H$ be the additive subgroup of the group $\mathbb{Z}_2[[t]]$ of formal power series over $\mathbb{Z}_2$ consisting of finite Laurent polynomials of $(1+t)$, and consider the injective endomorphism $\psi(F(t)) := tF(t)$ for $F(t) \in \mathbb{Z}_2[[t]]$. Since $tF(t) = (1+t)F(t) - F(t)$, we have that $\psi(H) \subseteq H$. Observe that $(1+t)^k - 1 \in \psi(H)$ for any $k \in \mathbb{Z}$; this easily implies that $\psi(H)$ is exactly the subgroup of $H$ of power series divisible by $t$, with index $[H : H_1] = 2$. We then let $H_n := \psi^{\circ n}(H)$, a nested sequence of finite index isomorphic subgroups.

Consider now the **right coset tree** $\mathcal{T}$ corresponding to the subgroup sequence $(H_n)_{n\geq 0}$, as defined before (13.21). We have $\mathcal{T} = \mathbb{T}_2$, and the boundary $\partial\mathcal{T}$ is a topological group: the profinite additive group $\mathbb{Z}_2[[t]]$, via the identification

$$\Phi : x_1 x_2 \cdots \mapsto \sum_{i\geq 1} x_i t^{i-1}, \qquad (15.5) \quad \texttt{\{e.Phi\}}$$

where $x = x_1x_2\ldots$ is the shorthand notation for the ray $H = H_0x_0 \supset H_1x_1 \supset H_2x_2 \supset \ldots$ in $\mathcal{T}$.

Let $A$ be the cyclic group $\mathbb{Z}$ acting on $H$ by multiplication by $(1+t)$. Thus the semidirect product $\mathsf{G} = A \ltimes H$ is the group of the following transformations of $\mathbb{Z}_2[[t]]$:

$$F(t) \mapsto (1+t)^m F(t) + \sum_{k\in\mathbb{Z}} f(k)(1+t)^k, \qquad (15.6) \quad \texttt{\{e.trafo\}}$$

where $m \in \mathbb{Z}$ and $f : \mathbb{Z} \longrightarrow \mathbb{Z}_2$ is any function with finitely many non-zero values.

This group $\mathsf{G} = \big(\oplus_{\mathbb{Z}} \mathbb{Z}_2\big) \rtimes \mathbb{Z} = \mathbb{Z}_2 \wr \mathbb{Z}$ is the standard lamplighter group; for each element $(m, f)$, one can think of $m \in \mathbb{Z}$ as the position of the lamplighter, while $f : \mathbb{Z} \longrightarrow \mathbb{Z}_2$ is the configuration of the lamps. We can represent $f$ by the finite set $\operatorname{supp} f \subset \mathbb{Z}$. The usual wreath product generators are $s$ and $R$, representing "switch" and "Right"; we will also use $L = R^{-1}$. So, for example, $Rs = (1, \{1\})$. In terms of the representation (15.6), the action of $s$ is $F(t) \mapsto F(t) + 1$, while the action of $R$ is $F(t) \mapsto (1+t)F(t)$.

The action of $\mathsf{G}$ on the infinite binary tree $\mathcal{T}$ can now be described by the combination of (15.5) and (15.6), and it turns out to be a finite-state self-similar action. Namely, consider the following new generators of the lamplighter group: $a = Rs$, $b := R$. Note that $s = b^{-1}a = a^{-1}b$. Then, the action of these generators on the binary tree $\mathcal{T}$ can be easily checked to be

$$\begin{aligned} (0w)^a &= 1w^b \qquad & (0w)^b &= 0w^b \\ (1w)^a &= 0w^a \qquad & (1w)^b &= 1w^a, \end{aligned} \qquad (15.7) \quad \texttt{\{e.GZ\}}$$

for any finite or infinite $\{0, 1\}$ word $w$. Hence $\{a, b\}$ is a finite self-similar generating set. Another usual notation for this self-similar action, using (15.3), is

$$a = (b, a)\epsilon, \qquad\qquad b = (b, a), \qquad (15.8) \quad \texttt{\{e.LLselfsim\}}$$

We note that in the literature there are a few slightly different versions of (15.8) to describe the lamplighter group. This is partly due to the fact that interchanging the generators $a$ and $b$ induces an automorphism $\iota$ of $\mathsf{G}$, see e.g. [GriŻ01].

One can easily write down a couple of formulas like (15.8), but the resulting group might be very complicated. We will briefly discuss two famous examples, Grigorchuk's group of intermediate growth and the Basilica group.

**Grigorchuk's first group** $\mathbb{G}$ is defined by the following self-similar action on the binary tree:

$$a = \epsilon, \qquad b = (a, c), \qquad c = (a, d), \qquad d = (1, b). \tag{15.9}$$ {e.chuk}

If this looks a bit ad hoc, writing down the profiles of the generators will make it clearer, see Figure 15.3.

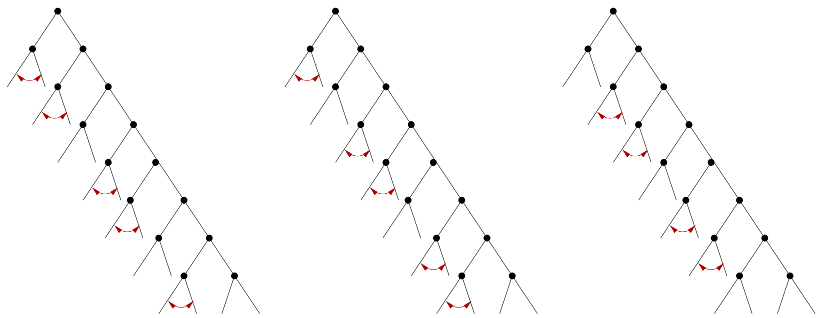


Figure 15.3: The profiles of the generators $b, c, d$ in Grigorchuk's group. {f.chuk}

Now, we have the following easy exercise:

▷ **Exercise 15.5.** *For Grigorchuk's group (15.9), check that the stabilizer $\mathbb{G}_v$ of any vertex in the binary tree is isomorphic to the original group, hence $\mathbb{G}$ has $\mathbb{G} \times \mathbb{G}$ as an index 2 subgroup.*

Moreover, using the third level stabilizers, one can show that there is an expanding virtual isomorphism from the direct product of eight copies of $\mathbb{G}$ to itself, hence, by Lemma 4.21, $\mathbb{G}$ has intermediate growth. See [GriP08] for more details.

Most of our course has been about how algebraic and geometric properties of a group influence the behaviour of SRW on it, though Kleiner's proof of Gromov's theorem was already somewhat in the other direction, probabilistic ideas giving algebraic results, see Section 10.1. But the first example of SRW applied to a group theory problem was by Bartholdi and Virág [BartV05]: they showed that the so-called **Basilica group**, introduced in Exercise 15.4, is amenable by finding a finite generating system on it for which they were able to compute that the speed is zero. At that point this was the furthest known example of an amenable group from abelian groups, namely, it cannot be built from groups of subexponential growth via group extensions.

Continuations of this Basilica work include [BartKN10] and [AmAV09]. The **activity growth** $\mathsf{Act}_G(n)$ of a finite state self-similar group $\Gamma$ is the number of length $n$ words $w$ such that the section $g|_w$ is not the identity for some of the self-similar generators $g$.

▷ **Exercise 15.6.** *Show that any finite state self-similar group $\Gamma$ with bounded activity growth is contracting (as defined just before Exercise 15.1).*

▷ **Exercise 15.7.** *Show that the activity growth $\mathsf{Act}_G(n)$ is either polynomial or exponential. (Hint: we get exponential growth iff the Moore diagram contains at least two directed cycles that can be reached from each other.)*

▷ **Exercise 15.8.** *Construct finite groups with exponential activity growth.*

Sidki [Sid00] showed that a polynomial activity self-similar group cannot contain a free subgroup $F_2$. Are they always amenable? [BartKN10] showed this for bounded activity groups, while [AmAV09] for linear activity groups. For at most quadratic activity, the Poisson boundary is conjectured to be trivial (proved for at most linear growth), but not for larger activity. This is quite analogous to the case of lamplighter groups $\mathbb{Z}_2 \wr \mathbb{Z}^d$. Furthermore, it is not known if all contracting groups are amenable.

### 15.2 Constructing monsters using hyperbolicity

{ss.monsters}

### 15.3 Thompson's group $F$

{ss.ThompF}

A very famous example of a group whose amenability is not known is the following. See [CanFP96] for some background, and [Cal09] for more recent stories.

{d.Thompson}

**Definition 15.2.** *Consider the set $F$ of orientation preserving piecewise linear homeomorphisms of $[0,1]$ to itself whose graphs satisfy the conditions that*

*a) All slopes are of the form $2^a$, $a \in \mathbb{Z}$.*

*b) All break points have first coordinate of the form $\frac{k}{2^n}$, $k, n \in \mathbb{N}$.*

*Clearly, $F$ is a group with composition of maps as a multiplication operation and this group is called* **Thompson's group $F$.**

**Question 15.3.** *Determine whether Thompson's group $F$ is amenable or not.*

Kaimanovich has proved that SRW has positive speed on Thompson's group $F$ for some generating set, hence this probabilistic direction of attack is not available here.

## 16 Quasi-isometric rigidity and embeddings

{s.qirigid}

It is a huge project posed by Gromov (1981) to classify groups up to quasi-isometries. One step towards such a classification would be to describe all self-quasi-isometries of a given group. Certain groups, e.g., fundamental groups of compact hyperbolic manifolds of dimension $n \geq 3$) are quite rigid: all quasi-isometries come in some sense from group automorphisms. A similar, more classical, result is:

{t.Mostow}

**Theorem 16.1** (Mostow rigidity 1968)**.** *If two complete finite volume hyperbolic manifolds $M, N$ have $\pi_1(M) \simeq \pi_1(N)$, then they are isometric, moreover, the group isomorphism is induced by an isometry of $\mathbb{H}^n$.*

Let us give here the rough strategy of the proof. The group isomorphism induces a quasi-isometry of $\mathbb{H}^n$, which then induces a quasi-conformal map on the ideal boundary $S^{n-1}$. This turns out (because of what?) to be a Möbius map, i.e., it comes from an isometry of $\mathbb{H}^n$.

An application of the Mostow rigidity theorem that is interesting from a probabilistic point of view, related to Theorem 11.1, was by Thurston, who proved that any finite triangulated planar graph has an essentially unique circle packing representation. However, there is also a simple elementary proof, due to Oded Schramm, using a maximal principle argument [Wik10a].

On the other end, the quasi-isometry group of $\mathbb{Z}$ is huge and not known.

Here is a completely probabilistic problem of the same flavor. See [Pel07] for details.

**Question 16.2** (Balázs Szegedy)**.** *Take two independent Poisson point processes on $\mathbb{R}$. Are they quasi-isometric to each other a.s.?*

This is motivated by the probably even harder question of Miklós Abért (2003): are two independent infinite clusters of $\mathsf{Ber}(p)$ percolation on the same transitive graph quasi-isometric almost surely?

Gromov conjectured that if two groups are quasi-isometric to each other, then they are also bi-Lipschitz equivalent, i.e., there is a bijective quasi-isometry. The following proof of a special case I learnt from Gábor Elek:

▷ **Exercise 16.1.** * *Using wobbling paradoxical decompositions, show that if two non-amenable groups are quasi-isometric to each other, then they are also bi-Lipschitz equivalent.*

However, the conjecture turned out to be false for solvable groups [Dym10]. But it is very much open for nilpotent groups, a favourite question of mine:

**Question 16.3.** *Are quasi-isometric nilpotent groups also bi-Lipschitz equivalent?*

I think the answer is yes, Bruce Kleiner thinks it's no. It is known to be yes if one of the groups is $\mathbb{Z}^d$, see [Sha04].

For a long while it was not known if all transitive graphs are quasi-isometric to some Cayley graph. For instance, transitive graphs of polynomial growth are always quasi-isometric to a nilpotent group. This is an interesting question from the percolation point of view, since we expect most properties to be quasi-isometry invariant, while a key tool, the Mass Transport Principle (12.9), works only for unimodular ones (including Cayley graphs). So, an affirmative answer would mean that percolation theory is somehow on the wrong track. Fortunately, Eskin-Fisher-Whyte proved as a byproduct of their work [EsFW07] on the quasi-isometric rigidity of the lamplighter groups $\mathsf{F}\wr\mathbb{Z}$, where $\mathsf{F}$ is a finite Abelian group, that the non-unimodular Diestel-Leader graphs $\mathsf{DL}(\mathsf{k},\ell)$ with $k \neq \ell$, see [Woe05] for their definition, are counterexamples.

**Question 16.4.** *Is every unimodular transitive graph quasi-isometric to a Cayley graph?*

The Eskin-Fisher-Whyte proof introduces something called "coarse metric differentiation", a technique similar to the one used by Cheeger and Kleiner to prove that the Heisenberg group does not have a Lipschitz embedding into $L^1$, see [ChKN09], and by Lee and Raghavendra to show that there are finite planar graphs needing at least a Lipschitz constant $2-\epsilon$ [LeeRa10]. It is conjectured that a universal constant suffices for all planar graphs.

In general, it is a huge subject what finite and infinite metric spaces embed into what $L^p$ space with how much metric distortion. One motivation is from theoretical computer science: analyzing large data sets (equipped with a natural metric, like the number of disagreements in two DNA sequences) is much easier if the data set is a subset of some nice space. We have also used nice harmonic embeddings into $L^2$ to gain algebraic information (in Kleiner's proof of Gromov's theorem) and to analyze random walks (in the Erschler-Lee-Peres results). There are a lot of connections between random walks and embeddings, see [NaoP08]. The target case of $L^2$ is easier, $L^1$ is more mysterious.

▷ **Exercise 16.2.** *Show that any finite subset of $L^2$ embeds isometrically into $L^1$.*

# References


[AbH15] M. Abért and T. Hubai. Benjamini-Schramm convergence and the distribution of chromatic roots for sparse graphs. *Combinatorica* **35** (2015), 127–151. `arXiv:1201.3861 [math.CO]`

[AbN07] M. Abért and N. Nikolov. Rank gradient, cost of groups and the rank versus Heegaard genus problem. *J. Eur. Math. Soc.*, to appear. [arXiv:math.GR/0701361]

[Ada92] S. Adams. Følner independence and the amenable Ising model. *Ergod. Th. & Dynam. Sys.* **12** (1992), 633–657.

[AiB87] M. Aizenman and D. Barsky. Sharpness of the phase transition in percolation models. *Comm. Math. Phys.* **108** (1987), 489–526.

[AiBNW99] M. Aizenman, A. Burchard, C. M. Newman, and D. B. Wilson. Scaling limits for minimal and random spanning trees in two dimensions. *Random Structures Algorithms* **15** (1999), 319–367. [arXiv:math.PR/9809145]

[AiDCS15] M. Aizenman, H. Duminil-Copin, and V. Sidoravicius. Random currents and continuity of Ising model's spontaneous magnetization. *Comm. Math. Phys.* **334** (2015), 719–742. arXiv:1311.1937 [math-ph]

[AiKN87] M. Aizenman, H. Kesten, and C. M. Newman. Uniqueness of the infinite cluster and continuity of connectivity functions for short- and long- range precolation. *Comm. Math. Phys.* **111** (1987), 505–532.

[AiN84] M. Aizenman and C. M. Newman. Tree graph inequalities and critical behavior in percolation models. *J. Statist. Phys.* **36** (1984), 107–143.

[AiW06] M. Aizenman and S. Warzel. The canopy graph and level statistics for random operators on trees. *Math. Phys. Anal. Geom.* **9** (2006), 291–333. [arXiv:math-ph/0607021]

[AjKSz82] M. Ajtai, J. Komlós and E. Szemerédi. Largest random component of a k-cube. *Combinatorica* **2** (1982), 1–7.

[AjKSz87] M. Ajtai, J. Komlós, and E. Szemerédi. Deterministic simulation in LOGSPACE. In *Proc. of the 19th Ann. ACM Symp. on Theory of Computing*, pp. 132–140, 1987.

[Ald91] D. Aldous. The continuum random tree II: an overview. In *Stochastic Analysis (ed. M.T. Barlow and N.H. Bingham)*, pages 23–70. Cambridge University Press, 1991. http://www.stat.berkeley.edu/~aldous/Papers/me55.pdf

[Ald97] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.* **25** (1997), 812–854.
http://www.stat.berkeley.edu/~aldous/Papers/me73.ps.Z

[AldL07] D. Aldous and R. Lyons. Processes on unimodular random networks. *Electron. J. Probab.* **12** (2007), Paper 54, 1454–1508.
http://128.208.128.142/~ejpecp/viewarticle.php?id=1754

[AldS04] D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on Discrete Structures*, vol. 110 of Encyclopaedia Math. Sci., pages 1–72. Springer, Berlin, 2004. http://www.stat.berkeley.edu/~aldous/Papers/me101.pdf

[Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica* **6** (1986), 83–96.

[AloBS04] N. Alon, I. Benjamini and A. Stacey. Percolation on finite graphs and isoperimetric inequalities. *Ann. Probab.* **33** (2004), 1727–1745. [arXiv:math.PR/0207112]

[AloM85] N. Alon and V. D. Milman. $\lambda_1$, isoperimetric inequalities for graphs, and superconcentrators. *J. Combin. Theory Ser. B* **38** (1985), no. 1, 73–88.

[AloS00] N. Alon and J. H. Spencer. *The Probabilistic Method.* 2nd edition. Wiley, 2000.

[AmAV09] G. Amir, O. Angel and B. Virág. Amenability of linear-activity automaton groups. *Preprint*, arXiv:0905.2007 [math.GR].

[Ang03] O. Angel. Growth and percolation on the uniform infinite planar triangulation. *Geom. Funct. Anal.* **13** (2003), no. 5, 935–974. [arXiv:math.PR/0208123]

[AngB07] O. Angel and I. Benjamini. A phase transition for the metric distortion of percolation on the hypercube. *Combinatorica* **27** (2007), 645–658. [arXiv:math.PR/0306355]

[AngSz] O. Angel and B. Szegedy. On the recurrence of the weak limits of discrete structures. *In preparation.*

[AntP96] P. Antal and A. Pisztora. On the chemical distance in supercritical Bernoulli percolation. *Ann. Probab.* **24** (1996), 1036–1048.

[AntV08] T. Antunović and I. Veselić. Sharpness of the phase transition and exponential decay of the subcritical cluster size for percolation on quasi-transitive graphs. *J. Statist. Phys.* **130** (2008), 983–1009. arXiv:0707.1089 [math.PR]

[Arr83] R. Arratia. Site recurrence for annihilating random walks on $\mathbb{Z}^d$. *Ann. Probab.* **11** (1983), 706–713.

[ArBLRSV05] G. N. Arzhantseva, J. Burillo, M. Lustig, L. Reeves, H. Short and E. Ventura. Uniform non-amenability. *Adv. Math.* **197** (2005), 499–522. http://www.mat.univie.ac.at/~arjantseva/Abs/folner.pdf

[Aus13] T. Austin. Rational group ring elements with kernels having irrational dimension. *Proc. London Math. Soc.* **107** (2013), 1424–1448. arXiv:0909.2360v2 [math.GR]

[BabSz92] L. Babai and M. Szegedy. Local expansion of symmetrical graphs. *Combin. Probab. & Computing* **1** (1992), 1–11.

[BabB99] E. Babson and I. Benjamini. Cut sets and normed cohomology, with applications to percolation. *Proc. Amer. Math. Soc.* **127** (1999), 589–597.

[BaliBo12] P. Balister and B. Bollobás. Projections, entropy and sumsets. *Combinatorica* **32** (2012), 125–141. arXiv:0711.1151v1 [math.CO]

[BalBDCM10] J. Balogh, B. Bollobás, H. Duminil-Copin, R. Morris. The sharp threshold for bootstrap percolation in all dimensions. *Trans. Amer. Math. Soc.*, to appear. arXiv:1010.3326 [math.PR]

[BalBM10] J. Balogh, B. Bollobás and R. Morris. Bootstrap percolation in high dimensions. *Combinatorics, Probability & Computing* **19** (2010), 643–692. arXiv:0907.3097 [math.PR]

[BalP07] J. Balogh and B. Pittel. Bootstrap percolation on random regular graphs. *Random Struc. & Alg.* **30** (2007), 257–286. `http://www.math.uiuc.edu/~jobal/cikk/bootfig.pdf`

[BalPP06] J. Balogh, Y. Peres and G. Pete. Bootstrap percolation on infinite trees and non-amenable groups, *Combin. Probab. & Computing* **15** (2006), no. 5, 715–730. `[arXiv:math.PR/0311125]`

[BarGN91] D. J. Barsky, G. R. Grimmett, and C. M. Newman. Percolation in half-spaces: equality of critical densities and continuity of the percolation probability. *Probab. Theory Related Fields* **90** (1991), 111–148.

[BartE11] L. Bartholdi and A. Erschler. Poisson-Furstenberg boundary and growth of groups. *Preprint*, `arXiv:1107.5499 [math.GR]`

[BartGN03] L. Bartholdi, R. Grigorchuk and V. Nekrashevych. From fractal groups to fractal sets. *Fractals in Graz (P. Grabner and W. Woess, eds.)*, Trends in Mathematics, Birkhäuser Verlag, Basel, 2003, pp. 25–118. `[arXiv:math.GR/0202001v4]`

[BartKN10] L. Bartholdi, V. Kaimanovich and V. Nekrashevych. On amenability of automata groups. *Duke Math. J.* **154** (2010), no. 3, 575–598. `arXiv:0802.2837 [math.GR]`

[BartV05] L. Bartholdi and B. Virág. Amenability via random walks. *Duke Math. J.* **130** (2005), 39–56. `[arXiv:math.GR/0305262]`

[BasS17] D. Basu and A. Sapozhnikov. Kesten's incipient infinite cluster and quasi-multiplicativity of crossing probabilities. *Elect. Comm. Probab.*

[Bea83] A. F. Beardon. *The geometry of discrete groups.* Graduate Texts in Mathematics, Vol. 91. Springer, New York, 1983.

[BefDC10] V. Beffara and H. Duminil-Copin. The self-dual point of the two-dimensional random-cluster model is critical for $q \geq 1$. *Probab. Theory Related Fields*, to appear. `arXiv:1006.5073 [math.PR]`

[BekdHV08] B. Bekka, P. de la Harpe and A. Valette. *Kazhdan's property (T).* Cambridge University Press, 2008. `http://perso.univ-rennes1.fr/bachir.bekka/KazhdanTotal.pdf`

[BekV97] M. Bekka and A. Valette. Group cohomology, harmonic functions and the first $L^2$-Betti number. *Potential Anal.* **6** (1997), no. 4, 313–326.

[Bel03] I. Belegradek. On co-Hopfian nilpotent groups. *Bull. London Math. Soc.* **35** (2003) 805–811.

[Ben91] I. Benjamini. Instability of the Liouville property for quasi-isometric graphs and manifolds of polynomial volume growth. *J. Theoret. Probab.* **4** (1991), no. 3, 631–637.

[Ben08] I. Benjamini. Discussion at an American Institute of Mathematics workshop, Palo Alto, May 2008.

[BenC10] I. Benjamini and N. Curien. Ergodic theory on stationary random graphs. *Preprint*, `arXiv:1011.2526 [math.PR]`

[BenDCKY] I. Benjamini, H. Duminil-Copin, G. Kozma and A. Yadin. Minimal harmonic functions I, upper bounds. *In preparation.*

[BenKPS04] I. Benjamini, H. Kesten, Y. Peres and O. Schramm. Geometry of the uniform spanning forest: transitions in dimensions 4, 8, 12, .... *Ann. of Math. (2)* **160** (2004), 465–491. `[arXiv:math.PR]0107140`

[BenK10] I. Benjamini and G. Kozma. Nonamenable Liouville graphs. *Preprint*, `arXiv:1010.3365 [math.MG]`

[BLPS99a] I. Benjamini, R. Lyons, Y. Peres and O. Schramm. Group-invariant percolation on graphs. *Geom. Funct. Anal.* **9** (1999), 29–66.

[BLPS99b] I. Benjamini, R. Lyons, Y. Peres and O. Schramm. Critical percolation on any nonamenable group has no infinite clusters. *Ann. Probab.* **27** (1999), 1347–1356.

[BLPS01] I. Benjamini, R. Lyons, Y. Peres and O. Schramm. Uniform spanning forests. *Ann. Probab.* **29** (2001), 1–65.

[BenLS99] I. Benjamini, R. Lyons and O. Schramm. Percolation perturbations in potential theory and random walks, *In: Random walks and discrete potential theory (Cortona, 1997)*, Sympos. Math. XXXIX, M. Picardello and W. Woess (eds.), Cambridge Univ. Press, Cambridge, 1999, pp. 56–84. `[arXiv:math.PR/9804010]`

[BenNP11] I. Benjamini, A. Nachmias and Y. Peres. Is the critical percolation probability local? *Probab. Theory Related Fields* **149** (2011), 261–269. `arXiv:0901.4616 [math.PR]`

[BenS96a] I. Benjamini and O. Schramm. Harmonic functions on planar and almost planar graphs and manifolds, via circle packings. *Invent. Math.* **126** (1996), 565–587.

[BenS96b] I. Benjamini and O. Schramm. Random walks and harmonic functions on infinite planar graphs using square tilings. *Ann. Probab.* **24** (1996), 1219–1238.

[BenS96c] I. Benjamini and O. Schramm. Percolation beyond $\mathbb{Z}^d$, many questions and a few answers. *Elect. Commun. Probab.* **1** (1996), 71–82. `http://www.math.washington.edu/~ejpecp/ECP/viewarticle.php?id=1561&layout=abstract`

[BenS01] I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.* **6** (2001), no. 23, 13 pp. (electronic). `[arXiv:math.PR/0011019]`

[vdBerS99] J. van den Berg and J. Steif. On the existence and non-existence of finitary codings for a class of random fields. *Ann. Probab.* **27** (1999), 1501–1522. `http://www.math.chalmers.se/~steif/p27.pdf`

[BerKMP05] N. Berger, C. Kenyon, E. Mossel and Y. Peres. Glauber dynamics on trees and hyperbolic graphs. *Probab. Theory Related Fields* **131** (2005), no. 3, 311–340. `[arXiv:math.PR/0308284]`

[Bergm68] G. Bergman. On groups acting on locally finite graphs. *Ann. Math.* **88** (1968), 335–340.

[BhES12] S. Bhamidi, S. N. Evans and A. Sen. Spectra of large random trees. *J. Theoret. Probab.* **25** (2012), 613–654. `arXiv:0903.3589 [math.PR]`.

[Bil86] P. Billingsley. *Probability and measure.* Second edition. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[BlRZ95] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Stat. Phys.* **79** (1995), 473–482.

[BobT06] S. G. Bobkov and P. Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *J. Theoret. Probab.* **19** (2006), no. 2, 289–336. `http://people.math.gatech.edu/~tetali/PUBLIS/BT_ENT.pdf`

[Bol81] B. Bollobás. The independence ratio of regular graphs. *Proc. Amer. Math. Soc.* **83** (1981), 433–436. `http://www.jstor.org/stable/2043545`

[Bol01] B. Bollobás. *Random graphs. 2nd edition.* Cambridge University Press, 2001.

[BolL91] B. Bollobás and I. Leader. Edge-isoperimetric inequalities in the grid. *Combinatorica* **11** (1991), 299–314.

[BolR06] B. Bollobás and O. Riordan. Sharp thresholds and percolation in the plane. *Random Structures & Algorithms* **29** (2006), 524–548. `[arXiv:math.PR/0412510]`

[BolT87] B. Bollobás and A. Thomason. Threshold functions. *Combinatorica* **7** (1987), 35–38.

[BorBL09] J. Borcea, P. Brändén and T. M. Liggett. Negative dependence and the geometry of polynomials. *J. Amer. Math. Soc.* **22** (2009), 521–567. `arXiv:0707.2340 [math.PR]`

[BorChKL13] C. Borgs, J. Chayes, J. Kahn and L. Lovász. Left and right convergence of graphs with bounded degree. *Random Structures & Algorithms* **42** (2013), 1–28. `arXiv:1002.0115 [math.CO]`

[BouKKKL92] J. Bourgain, J. Kahn, G. Kalai, Y. Katznelson and N. Linial. The influence of variables in product spaces. *Israel J. Math.* **77** (1992), 55–64.

[BurgM00] M. Burger and S. Mozes. Lattices in products of trees. *Publ. Math. I.H.E.S.* **92** (2000), 151–194.

[BurtK89] R. M. Burton and M. Keane. Density and uniqueness in percolation. *Comm. Math. Phys.* **121** (1989), 501–505.

[BurtP93] R. M. Burton and R. Pemantle. Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances. *Ann. Probab.* **21** (1993), 1329–1371. `[arXiv:math.PR/0404048]`

[Cal09] D. Calegari. Amenability of Thompson's group $F$? *Geometry and the imagination*, weblog. `http://lamington.wordpress.com/2009/07/06/amenability-of-thompsons-group-f/`

[CanFP96] J. W. Cannon, W. J. Floyd, and W. R. Parry. Introductory notes on Richard Thompson's groups. *Enseign. Math. (2)*, **42** (1996), No. 3-4, 215–256.

[CapM06] P. Caputo and F. Martinelli. Phase ordering after a deep quench: the stochastic Ising and hard core gas models on a tree. *Probab. Theory Relat. Fields* **136** (2006), 37–80. `[arXiv:math.PR/0412450]`

[Car85] T. K. Carne. A transmutation formula for Markov chains. *Bull. Sci. Math. (2)* **109** (1985), 399–405.

[ChCh86] J. T. Chayes and L. Chayes. Inequality for the infinite-cluster density in Bernoulli percolation. *Phys. Rev. Lett.* **56** (1986), no. 16, 1619–1622.

[Che70] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *In: R.C. Gunning,editor, Problems in Analysis*, pages 195–199, Princeton Univ. Press, Princeton, NJ, 1970. A symposium in honor of Salomon Bochner, Princeton University, 1–3 April 1969.

[ChKN09] J. Cheeger, B. Kleiner and A. Naor. Compression bounds for Lipschitz maps from the Heisenberg group to $L_1$. `arXiv:0910.2026 [math.MG]`

[ChPP04] D. Chen and Y. Peres, with an appendix by G. Pete. Anchored expansion, percolation and speed. *Ann. Probab.* **32** (2004), 2978–2995. `[arXiv:math.PR/0303321]`

[ChoD60] G. Choquet and J. Deny. Sur l'équation de convolution $\mu = \mu * \sigma$. *C.R. Acad. Sci. Paris Sér. I* **250** (1960), 799–801.

[ChuL06] F. Chung and L. Lu. Concentration inequalities and martingale inequalities — a survey. *Internet Mathematics* **3** (2006), 79–127. `http://www.math.ucsd.edu/~fan/wp/concen.pdf`

[ColM97] T. H. Colding and W. P. Minicozzi II. Harmonic functions on manifolds. *Ann. of Math. (2)* **146** (1997), no. 3, 725–747.

[Cou00] T. Coulhon. Random walks and geometry on infinite graphs. *Lecture notes on analysis on metric spaces, Trento, C.I.R.M., 1999.* L. Ambrosio, F. Serra Cassano, éd., Scuola Normale Superiore di Pisa, 2000, pp. 5–30. `http://www.u-cergy.fr/rech/pages/coulhon/trento.ps`

[CouGP01] T. Coulhon, A. Grigor'yan and C. Pittet. A geometric approach to on-diagonal heat kernel lower bounds on groups. *Ann. Inst. Fourier (Grenoble)* **51** (2001), 1763–1827.

[CouSC93] T. Coulhon and L. Saloff-Coste. Isopérimétrie pour les groupes et les variétés. *Rev. Mat. Iberoamericana* **9** (1993), 293–314.

[CovT06] T. M. Cover and J. A. Thomas. *Elements of information theory. Second edition.* John Wiley and Sons, New York, 2006.

[CsF16] P. Csikvári, P. E. Frenkel. Benjamini–Schramm continuity of root moments of graph polynomials. *Europ. J. Combin.* **52** (2016), 302–320. `arXiv:1204.0463 [math.CO]`

[DDMN10] D. D'Angeli, A. Donno, M. Matter and T. Nagnibeda. Schreier graphs of the Basilica group. *Journal of Modern Dynamics* **4** (2010), 167–205. `arXiv:0911.2915 [math.GR]`

[Dei07] P. Deift. Universality for mathematical and physical systems. In *International Congress of Mathematicians (Madrid, 2006), Vol I.*, pp. 125–152. Eur. Math. Soc., Zürich, 2007. `http://www.icm2006.org/proceedings/Vol_I/11.pdf`

[Del99] T. Delmotte. Parabolic Harnack inequality and estimates of Markov chains on graphs. *Revista Matematica Iberoamericana* **15** (1999), no. 1, 181–232.

[DemZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications. Second Edition.* Springer, New York, 1998.

[Der76] Y. Derriennic. Lois "zéro ou deux" pour les processus de Markov. Applications aux marches aléatoires. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **12** (1976), no. 2, 111–129.

[DiaF90] P. Diaconis and J. A. Fill. Strong stationary times via a new form of duality. *Ann. Probab.* **18** (1990), 1483–1522.

[DicS02] W. Dicks and T. Schick. The spectral measure of certain elements of the complex group ring of a wreath product. *Geom. Dedicata*, **93** (2002), 121–137.

[Die00] R. Diestel *Graph Theory. 2nd Edition.* Graduate Texts in Mathematics, vol. 173. Springer, New York, 2000.

[DieL01] R. Diestel and I. Leader. A conjecture concerning a limit of non-Cayley graphs. *J. Alg. Combin.* **14** (2001), 17–25.

[DinKLP11] J. Ding, J. H. Kim, E. Lubetzky, and Y. Peres. Anatomy of a young giant component in the random graph. *Random Struct. Alg.* **39** (2011), 139–178.

[Dod84] J. Dodziuk. Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Amer. Math. Soc.* **284** (1984), 787–794.

[Doo59] J. L. Doob. Discrete potential theory and boundaries. *J. Math. Mech.* **8** (1959), 433–458; erratum 993.

[Doy88] P. Doyle. Electric currents in infinite networks (preliminary version). *Unpublished manuscript*, available at `[arXiv:math.PR/0703899]`.

[DrKa09] C. Druţu and M. Kapovich: *Lectures on geometric group theory.* `http://www.math.ucdavis.edu/~kapovich/EPR/ggt.pdf`

[DuC14] H. Duminil-Copin. *Parafermionic observables and their applications to planar statistical physics models.* Sociedade Brasileira de Matemática, 2014.

[DuCGP13] H. Duminil-Copin, C. Garban and G. Pete. The near-critical planar FK-Ising model. *Comm. Math. Phys.*, to appear. `arXiv:1111.0144v3 [math.PR]`

[DuCS11] H. Duminil-Copin and S. Smirnov. *Conformal invariance of lattice models.* Lecture notes for the Clay Mathematical Institute Summer School in Buzios, 2010. `arXiv:1109.1549 [math.PR]`

[Dur10] R. Durrett. *Probability: theory and examples.* Fourth edition. Cambridge University Press, 2010.

[Dym10] T. Dymarz. Bilipschitz equivalence is not equivalent to quasi-isometric equivalence for finitely generated groups. *Duke Math. J.* **154** (2010), 509–526. `arXiv:0904.3764 [math.GR]`

[EdS88] R. G. Edwards and A. D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. D (3)* **38** (1988), no. 6, 2009–2012.

[Ele07] G. Elek. The combinatorial cost. *l'Enseignement Math.* **53** (2007), 225–236. `[arXiv:math.GR/0608474]`

[Ele15] G. Elek. Full groups and soficity. *Proc. Amer. Math. Soc.* **143** (2015), 1943–1950. `arXiv:1211.0621 [math.GR]`

[ElL10] G. Elek and G. Lippner. Borel oracles. An analytical approach to constant-time algorithms. *Proc. Amer. Math. Soc.* **138** (2010), 2939–2947. `arXiv:0907.1805 [math.CO]`

[ElSz06] G. Elek and E. Szabó. On sofic groups. *J. of Group Theory* **9** (2006), 161–171. `[arXiv:math.GR/0305352]`

[ElTS05] G. Elek and V. T. Sós. Paradoxical decompositions and growth conditions. *Combin., Probab. & Comput.* **14** (2005), 81–105.

[vEnt87] A. C. D. van Enter. Proof of Straley's argument for bootstrap percolation. *J. Stat. Phys.* **48** (1987), 943–945.

[ErdR60] P. Erdős and A. Rényi. On the evolution of random graphs. *Mat. Kutató Int. Közl.* (1960) **5**, 17–60. `http://www.renyi.hu/~p_erdos/1960-10.pdf`

[Ers04a] A. Erschler. Liouville property for groups and manifolds. *Invent. Math.* **155** (2004), 55–80.

[Ers04b] A. Erschler. Boundary behavior for groups of subexponential growth. *Ann. of Math. (2)* **160** (2004), no. 3, 1183–1210.

[EsFW07] A. Eskin, D. Fisher and K.Whyte. Quasi-isometries and rigidity of solvable groups. *Pure and Applied Mathematics Quarterly* **3** (2007), 927–947. `[arXiv:math.GR/0511647]`

[Eva98] L. C. Evans. *Partial differential equations.* Graduate Studies in Mathematics, 19. American Mathematical Society, Providence, RI, 1998.

[EvKPS00] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Prob.* **10** (2000), 410–433.
`http://stat-www.berkeley.edu/~peres/mine/broad.ps`

[FeM75] J. Feldman and C. C. Moore. Ergodic equivalence relations, cohomology, and von Neumann algebras. *Bull. Amer. Math. Soc.* **81** (1975), 921–924.

[Fel68] W. Feller. *Introduction to probability theory and its applications, Vol. I. Third edition.* John Wiley and Sons, 1968.

[Føl55] E. Følner. On groups with full Banach mean values. *Math. Scand.* **3** (1955), 243–254.

[ForK72] C. M. Fortuin and P. W. Kasteleyn. On the random-cluster model. I. Introduction and relation to other models. *Physica* **57** (1972), 536–564.

[FriB99] E. Friedgut, with an Appendix by J. Bourgain. Sharp thresholds of graphs properties, and the $k$-sat problem. *J. Amer. Math. Soc.* **12** (1999), 1017–1054. `http://www.ma.huji.ac.il/~ehudf/docs/thre.ps` and `http://www.ma.huji.ac.il/~ehudf/docs/bo.ps`

[FrHTV18] J. Frisch, Y. Hartman, O. Tamuz, and P. Vahidi-Ferdowsi. Choquet-Deny groups and the infinite conjugacy class property. *Preprint*, `arXiv:1802.00751 [math.GR]`

[GaGa81] O. Gabber and Z. Galil. Explicit constructions of linear-sized superconcentrators. *J. Comput. System Sci.* **22** (1981), no. 3, 407–420.

[Gab02] D. Gaboriau. On orbit equivalence of measure preserving actions. *Rigidity in dynamics and geometry (Cambridge, 2000)*, pp. 167–186, Springer, Berlin, 2002. `http://www.umpa.ens-lyon.fr/~gaboriau/Travaux-Publi/Cambridge/Cambridge.html`

[Gab05] D. Gaboriau. Invariant percolation and harmonic Dirichlet functions. *Geom. Funct. Anal.* **15** (2005), no. 5, 1004–1051. `[arXiv:math.PR/0405458]`

[Gab10] D. Gaboriau. Orbit equivalence and measured group theory. *In: Proceedings of the 2010 Hyderabad ICM*, to appear. `arXiv:1009.0132v1 [math.GR]`

[GabL09] D. Gaboriau and R. Lyons. A measurable-group-theoretic solution to von Neumann's problem. *Invent. Math.* **177** (2009), no. 3, 533–540. `arXiv:0711.1643 [math.GR]`

[GácsL09] P. Gács and L. Lovász. *Complexity of algorithms.* Lecture notes, 1997-2009. `http://www.cs.elte.hu/~lovasz/complexity.pdf`

[GarPS10a] C. Garban, G. Pete, and O. Schramm. The Fourier spectrum of critical percolation. *Acta Math.* **205** (2010), no. 1, 19–104. `arXiv:0803.3750 [math.PR]`

[GarPS10b] C. Garban, G. Pete and O. Schramm. The scaling limit of the Minimal Spanning Tree — a preliminary report. *XVIth International Congress on Mathematical Physics (Prague 2009)*, ed. P. Exner, pp. 475–480. World Scientific, Singapore, 2010. `arXiv:0909.3138 [math.PR]`

[GarPS13a] C. Garban, G. Pete and O. Schramm. Pivotal, cluster and interface measures for critical planar percolation. *Journ. Amer. Math. Soc.* **26** (2013), 939–1024. `arXiv:1008.1378 [math.PR]`

[GarPS13b] C. Garban, G. Pete and O. Schramm. The scaling limits of near-critical and dynamical percolation. `arXiv:1305.5526 [math.PR]`

[Geo03] H-O. Georgii. Probabilistic aspects of entropy. *In: Entropy*, A. Greven, G. Keller, G. Warnecke (editors), pp. 37–54, Princeton University Press, 2003. `http://www.mathematik.uni-muenchen.de/~georgii/papers/Dresden.pdf`

[GeHM01] H-O. Georgii, O. Häggström and C. Maes. The random geometry of equilibrium phases. *Phase transitions and critical phenomena*, Vol. 18, Academic Press, San Diego, CA, pp. 1–142. `[arXiv:math.PR/9905031v1]`

[GlW97] E. Glasner and B. Weiss. Kazhdan's property T and the geometry of the collection of invariant measures. *Geom. Funct. Anal.* **7** (1997), 917–935.

[GoV97] E. I. Gordon and A. M. Vershik. Groups that are locally embeddable in the class of finite groups. *Algebra i Analiz* **9** (1997), no. 1, 71–97. English translation: *St. Petersburg Math. J.* **9** (1997), no. 1, 49–67. `http://www.pdmi.ras.ru/~vershik/gordon.ps`

[Got73] W. Gottschalk. Some general dynamical notions. *In: Recent Advances in Topological Dynamics*, Lecture Notes in Mathematics 318, pp. 120–125. Springer, Berlin, 1973.

[Grab10] L. Grabowski. On Turing machines, dynamical systems and the Atiyah problem. *Preprint*, `arXiv:1004.2030 [math.GR]`.

[GraG06] B. T. Graham and G. R. Grimmett. Influence and sharp threshold theorems for monotonic measures. *Ann. Probab.* **34** (2006), 1726–1745.

[GraG96] J. Gravner and D. Griffeath. First passage times for threshold growth dynamics on $\mathbb{Z}^2$. *Ann. Probab.* **24** (1996), 1752–1778.

[GHS70] R. B. Griffiths, C. A. Hurst and S. Sherman. Concavity of magnetization of an Ising ferromagnet in a positive external field. *J. Math. Phys.* **11** (1970), 790–795.

[Gri83] R. I. Grigorchuk. On the Milnor problem of group growth. *Soviet Math. Dokl.* **28** (1983), no. 1, 23–26.

[GriNS00] R. I. Grigorchuk, V. V. Nekrashevich and V. I. Sushchanskii. Automata, dynamical systems and groups. *Proceedings of the Steklov Institute of Mathematics* **231** (2000), 128–203. http://www.math.tamu.edu/~grigorch/publications/PSIM128.PS

[GriP08] R. Grigorchuk and I. Pak. Groups of intermediate growth, an introduction. *Enseign. Math. (2)* **54** (2008), no. 3-4, 251–272. http://www.math.tamu.edu/~grigorch/publications/grigorchuk_pak_intermediate_growth.pdf

[GriŻ01] R. Grigorchuk and A. Żuk. The lamplighter group as a group generated by a 2-state automaton and its spectrum. *Geom. Dedicata* **87** (2001), no. 1-3, 209–244.

[Gri95] G. Grimmett. The stochastic random-cluster process and the uniqueness of random-cluster measures. *Ann. Probab.* **23** (1995), 1461–1510.

[Gri99] G. Grimmett. *Percolation. Second edition.* Grundlehren der Mathematischen Wissenschaften, 321. Springer-Verlag, Berlin, 1999.

[Gri06] G. Grimmett. *The random-cluster model.* Grundlehren der mathematischen Wissenschaften, 333, Springer-Verlag, Berlin, 2006.

[Gri10] G. Grimmett. *Probability on Graphs.* IMS Textbook Series, Vol. 1. Cambridge University Press, 2010. http://www.statslab.cam.ac.uk/~grg/books/pgs.html

[GriMan11] G. R. Grimmett and I. Manolescu. Inhomogeneous bond percolation on square, triangular, and hexagonal lattices. *Preprint*, arXiv:1105.5535 [math.PR].

[GriMar90] G. R. Grimmett and J. M. Marstrand. The supercritical phase of percolation is well behaved. *Proc. Roy. Soc. London, Ser. A* **430** (1990), 439–457.

[Gro81] M. Gromov. Groups of polynomial growth and expanding maps. Appendix by J. Tits. *Publ. Math. I.H.E.S.* **53** (1981), 53–73, 74–78.

[Gro93] M. Gromov. Asymptotic invariants of infinite groups. *Geometric group theory, Vol. 2 (Sussex, 1991)*, pages 1–295. Cambridge Univ. Press, Cambridge, 1993.

[Gro99] M. Gromov. Endomorphisms of symbolic algebraic varieties. *J. European Math. Soc.* **1** (1999), 109–197.

[GuGN13] O. Gurel-Gurevich and A. Nachmias. Recurrence of planar graph limits. *Ann. Math.* **177** (2013), no. 2, 761–781. arXiv:1206.0707 [math.PR]

[Häg95] O. Häggström. Random-cluster measures and uniform spanning trees. *Stochastic Process. Appl.* **59** (1995), 267–275.

[HäM09] O. Häggström and P. Mester. Some two-dimensional finite energy percolation processes. *Electr. Comm. Probab.* **14** (2009), 42–54. `arXiv:1011.2872 [math.PR]`

[HäPS99] O. Häggström, Y. Peres and R. H. Schonmann. Percolation on transitive graphs as a coalescent process: Relentless merging followed by simultaneous uniqueness. *Perplexing Problems in Probability, Festschrift in honor of Harry Kesten*, ed. M. Bramson and R. Durrett, pages 69–90. Birkhäuser, Boston, 1999.

[HäPSt97] O. Häggström, Y. Peres, and J. E. Steif. Dynamical percolation. *Ann. Inst. H. Poincaré Probab. Statist.* **33** (1997), 497–528.

[Ham57] J. M. Hammersley. Percolation processes. Lower bounds for the critical probability. *Ann. Math. Stati.* **28** (1957), 790–795.

[HamMP12] A. Hammond, E. Mossel and G. Pete. Exit time tails from pairwise decorrelation in hidden Markov chains, with applications to dynamical percolation. *Elect. J. Probab.* **17** (2012), Article no. 68, 1–16. `arXiv:1111.6618 [math.PR]`.

[HamPS15] A. Hammond, G. Pete and O. Schramm. Local time on the exceptional set of dynamical percolation, and the Incipient Infinite Cluster. *Ann. Probab.* **43** (2015), 2949-3005. `arXiv:1208.3826 [math.PR]`

[HarvdHS03] T. Hara, R. van der Hofstad and G. Slade. Critical two-point functions and the lace expansion for spread-out high-dimensional percolation and related models. *Ann. Probab.* **31** (2003), no. 1, 349–408. `http://www.math.ubc.ca/~slade/xspace_AOP128.pdf`

[dlHar00] P. de la Harpe. *Topics in geometric group theory.* Chicago Lectures in Mathematics, University of Chicago Press, 2000.

[Har60] T. E. Harris. A lower bound for the critical probability in a certain percolation process. *Proc. Cambridge Philos. Soc.* **56** (1960), 13–20.

[Hart76] S. Hart. A note on the edges of the $n$-cube. *Discrete Math.* **14** (1976), 157–163.

[HatLSz12] H. Hatami, L. Lovász and B. Szegedy. Limits of local-global convergent graph sequences. *Preprint*, `arXiv:1205.4356 [math.CO]`.

[HeS95] Z. X. He and O. Schramm. Hyperbolic and parabolic packings. *Discrete Comput. Geom.* **14** (1995), 123–149.

[HeSC93] W. Hebisch and L. Saloff-Coste. Gaussian estimates for Markov chains and random walks on groups. *Ann. Probab.* **21** (1993), 673–709. `http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aop/1176989263`

[vdHof13] R. van der Hofstad. *Random graphs and complex networks.* Book in preparation. `http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf`

[dHoS97] F. den Hollander and J. Steif. On K-automorphisms, Bernoulli shifts and Markov random fields. *Ergodic Th. & Dynam. Sys.* **17** (1997), 405–415.

[Hol74] R. Holley. Remarks on the FKG inequalities. *Comm. Math. Phys.* **36** (1974), 227–231.

[Holr07] A. E. Holroyd: Astonishing cellular automata (expository article). *Bulletin du Centre de Recherches Mathematiques* (2007), Vol 13, No 1, 10–13. `http://www.math.ubc.ca/~holroyd/papers/cell.pdf`

[HooLW06] S. Hoory, N. Linial and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society* **43** (2006), No. 4, 439–561. `http://www.cs.huji.ac.il/~nati/PAPERS/expander_survey.pdf`

[How00] C. D. Howard. Zero-temperature Ising spin dynamics on the homogeneous tree of degree three. *J. Appl. Prob.* **37** (2000), 736–747.

[HowN03] C. D. Howard and C. M. Newman. The percolation transition for the zero-temperature stochastic Ising model on the hexagonal lattice. *J. Stat. Physics*, **111** (2003), 57–72.

[IKT09] A. Ioana, A. S. Kechris and T. Tsankov. Subequivalence relations and positive-definite functions. *Groups Geom. Dyn.* **3** (2009), no. 4, 579–625. `arXiv:0806.0430 [math.DS]`

[JaR09] T. S. Jackson and N. Read. Theory of minimum spanning trees I: Mean-field theory and strongly disordered spin-glass model. `arXiv:0902.3651v3 [cond-mat.stat-mech]`

[JaŁR00] S. Janson, T. Łuczak and A. Ruciński. *Random graphs.* Wiley-Interscience, New York, 2000.

[Jár03] A. Járai. Incipient infinite percolation clusters in 2D. *Ann. Probab.* **31** (2003) no. 1, 444–485.

[JoS99] J. Jonasson and J. E. Steif. Amenability and phase transition in the Ising model. *J. Theoret. Probab.* **12** (1999), 549–559. `http://www.math.chalmers.se/~steif/p25.pdf`

[KahKL88] J. Kahn, G. Kalai and N. Linial. The influence of variables on boolean functions. *29th Annual Symposium on Foundations of Computer Science*, 68–80, 1988. `http://www.ma.huji.ac.il/~kalai/kkl.ps`

[Kai80] V. Kaimanovich. Spectral measure of transition operator and harmonic functions connected with random walks on discrete groups. *Zap. Nauch. Sem. LOMI* **97** (1980), 102–109. In Russian.

[Kai00] V. Kaimanovich. The Poisson formula for groups with hyperbolic properties. *Ann. of Math. (2)* **152** (2000), no. 3, 659–692.

[KaiV83] V. A. Kaimanovich and A. M. Vershik. Random walks on discrete groups: boundary and entropy. *Ann. Probab.* **11** (1983), 457–490.

[KalS06] G. Kalai and S. Safra. Threshold phenomena and influence: perspectives from mathematics, computer science, and economics. In *Computational complexity and statistical physics*, St. Fe Inst. Stud. Sci. Complex., pages 25–60. Oxford Univ. Press, New York, 2006. `http://www.ma.huji.ac.il/~kalai/ML.pdf`

[Kap14] M. Kapovich. Energy of harmonic functions and Gromov's proof of the Stallings' theorem. *Georgian Math. J.* **21** (2014), 281–296. `arXiv:0707.4231 [math.GR]`

[Kaz67] D. A. Kazhdan. Connection of the dual space of a group with the structure of its closed subgroups. *Functional Analysis and its Applications* **1** (1967), 63–65.

[KecM04] A. S. Kechris and B. D. Miller. *Topics in orbit equivalence.* Springer, 2004.

[Kes59] H. Kesten. Full Banach mean values on countable groups. *Math. Scand.* **7** (1959), 146–156.

[Kes80] H. Kesten. The critical probability of bond percolation on the square lattice equals 1/2. *Comm. Math. Phys.* **74** (1980), no. 1, 41–59.

[Kes82] H. Kesten. *Percolation theory for mathematicians.* Birkhäuser, Boston, 1982.

[Kes86] H. Kesten. The incipient infinite cluster in two-dimensional percolation. *Probab. Th. Rel. Fields* **73** (1986), 369–394.

[Kes87] H. Kesten. Scaling relations for 2D-percolation. *Comm. Math. Phys.* **109** (1987), 109–156.

[Kir07] W. Kirsch. *An invitation to random Schrödinger operators.* Lecture notes, `arXiv:0709.3707 [math-ph]`.

[Kle10] B. Kleiner. A new proof of Gromov's theorem on groups of polynomial growth. *J. Amer. Math. Soc.* **23** (2010), 815–829. `arXiv:0710.4593 [math.GR]`

[KorS97] N. J. Korevaar and R. M. Schoen. Global existence theorems for harmonic maps to non-locally compact spaces. *Comm. Anal. Geom.* **5** (1997), no. 2, 333–387.

[Koz07] G. Kozma. Percolation, perimetry, planarity. *Rev. Math. Iberoam.* **23** (2007), no. 2, 671–676. `[arXiv:math.PR/0509235]`

[Koz11] G. Kozma. Percolation on a product of two trees. *Ann. Probab.* **39** (2011), 1864–1895. Special issue dedicated to the memory of Oded Schramm. `arXiv:1003.5240 [math.PR]`

[KozN09] G. Kozma and A. Nachmias. The Alexander-Orbach conjecture holds in high dimensions. *Invent. Math.* **178** (2009), no. 3, 635–654. `arXiv:0806.1442 [math.PR]`

[KozN11] G. Kozma and A. Nachmias. Arm exponents in high dimensional percolation. *J. Amer. Math. Soc.* **24** (2011), 375–409. `arXiv:arXiv:0911.0871v1 [math.PR]`.

[KrSu09] M. Krivelevich and B. Sudakov. Minors in expanding graphs. *Geom. Funct. Anal.* **19** (2009) 294–331. `arXiv:0707.0133 [math.CO]`

[KrSu13] M. Krivelevich and B. Sudakov. The phase transition in random graphs: a simple proof. *Random Structures Algorithms* **43** (2013), 131–138. `arXiv:1201.6529 [math.CO]`

[Lac16] H. Lacoin. Mixing time and cutoff for the adjacent transposition shuffle and the simple exclusion. *Ann. Probab.* **44** (2016) 1426–1487. `arXiv:1309.3873 [math.PR]`

[LaRu69] O. Lanford and D. Ruelle. Observables at infinity and states with short range correlations in statistical mechanics. *Comm. Math. Phys.* **13** (1969), 194–215.

[Leb77] J. Lebowitz. Coexistence of phases in Ising ferromagnets. *J. Stat. Phys.* **16** (1977), 463–476.

[Led01] M. Ledoux. *The concentration of measure phenomenon.* Mathematical surveys and monographs, Vol. 89. American Mathematical Society, Providence, 2001.

[LeeP13] J. Lee and Y. Peres. Harmonic maps on amenable groups and a diffusive lower bound for random walks. *Ann. Probab.*, **41** (2013), 3392–3419. `arXiv:0911.0274 [math.PR]`

[LeeRa10] J. Lee and P. Raghavendra. Coarse differentiation and multi-flows in planar graphs. *Discrete Comput Geom.* **43** (2010), 346–362. `arXiv:0804.1573v4 [math.MG]`

[LeNW07] F. Lehner, M. Neuhauser and W. Woess. On the spectrum of lamplighter groups and percolation clusters. *Math. Ann.*, to appear. `arXiv:0712.3135v2 [math.FA]`

[LevP09] L. Levine and Y. Peres. Strong spherical asymptotics for rotor-router aggregation and the divisible sandpile. *Potential Anal.* **30** (2009), 1–27. `arXiv:0704.0688 [math.PR]`

[LevPW17] D. A. Levin, Y. Peres and E. L. Wilmer. *Markov chains and mixing times. Second edition.* With a chapter by J. G. Propp and D. B. Wilson. American Mathematical Society, Providence, RI, 2017. `http://www.uoregon.edu/~dlevin/MARKOV/`

[LipT79] R. J. Lipton and R. E. Tarjan. A separator theorem for planar graphs. *SIAM J. on Applied Math.* **36** (1979), 177–189.

[LipT80] R. J. Lipton and R. E. Tarjan. Applications of a planar separator theorem. *SIAM J. Comput.* **9** (1980), 615–627.

[Los87] V. Losert. On the structure of groups with polynomial growth. *Math. Z.* **195** (1987), 109–117.

[Lov12] L. Lovász. *Large networks and graph limits.* American Mathematical Society, Providence, Rhode Island, 2012. `http://www.cs.elte.hu/~lovasz/bookxx/hombook-almost.final.pdf`

[LovKa99] L. Lovász and R. Kannan. Faster mixing via average conductance. *Proceedings of the 27th Annual ACM Symposium on theory of computing*, 1999.

[Lub94] A. Lubotzky. *Discrete groups, expanding graphs and invariant measures.* Progress in Mathematics, 125. Birkhäuser Verlag, Basel, 1994. With an appendix by J. D. Rogawski.

[Łucz90] T. Łuczak. Component behavior near the critical point of the random graph process. *Random Structures Algorithms* **1** (1990), no. 3, 287–310.

[Lüc94] W. Lück. Approximating $L^2$-invariants by their finite-dimensional analogues. *Geom. Funct. Anal.* **4** (1994), no. 4, 455–481.

[Lyo83] T. Lyons. A simple criterion for transience of a reversible Markov chain. *Ann. Probab.* **11** (1983), 393–402.

[Lyo87] T. Lyons. Instability of the Liouville property for quasi-isometric Riemannian manifolds and reversible Markov chains. *J. Differential Geom.* **26** (1987), no. 1, 33–66.

[Lyo90] R. Lyons. Random walks and percolation on trees. *Ann. Probab.* **18** (1990), 931–958. `http://mypage.iu.edu/~rdlyons/pdf/rwpt.pdf`

[Lyo92] R. Lyons. Random walks, capacity, and percolation on trees. *Ann. Probab.* **20** (1992), 2043–2088. `http://mypage.iu.edu/~rdlyons/ps/cap11.ps`

[Lyo00] R. Lyons. Phase transitions on nonamenable graphs. *J. Math. Phys.* **41** (2000), 1099–1126. `[arXiv:math.PR/9908177]`

[Lyo05] R. Lyons. Asymptotic enumeration of spanning trees. *Combin. Probab. & Comput.* **14** (2005), 491–522. `[arXiv:math.CO/0212165]`

[Lyo09] R. Lyons. Random complexes and $\ell^2$-Betti numbers. *J. Top. Analysis* **1** (2009), 153–175. `arXiv:0811.2933 [math.PR]`

[Lyo14] R. Lyons. Factors of IID on trees. *Combin. Probab. Comput.*, to appear. `arXiv:1401.4197 [math.PR]`

[LyNaz11] R. Lyons and F. Nazarov. Perfect matchings as IID factors on non-amenable groups. *Europ. J. Combin.* **32** (2011), 1115–1125. `arXiv:0911.0092v2 [math.PR]`

[LyPer16] R. Lyons and Y. Peres. *Probability on trees and networks*. Cambridge University Press, 2016. Available at `http://mypage.iu.edu/~rdlyons`.

[LyPS03] R. Lyons, Y. Peres and O. Schramm. Markov chain intersections and the loop-erased walk. *Ann. Inst. H. Poincaré, Probab. Statist.* **39** (2003), 779–791. `[arXiv:math.PR/0107055]`

[LyPS06] R. Lyons, Y. Peres and O. Schramm. Minimal spanning forests. *Ann. Probab.* **34** (2006), no. 5, 1665–1692. `[arXiv:math.PR/0412263v5]`

[LySch99] R. Lyons and O. Schramm. Indistinguishability of percolation clusters. *Ann. Probab.* **27** (1999), 1809–1836. `[arXiv:math.PR/9811170v2]`

[MaTa17] S. Martineau and V. Tassion. Locality of percolation for Abelian Cayley graphs. *Ann. Probab.* **45** (2017), 1247–1277. `arXiv:1312.1946 [math.PR]`.

[McDi98] C. McDiarmid. Concentration. *In: Probabilistic methods for algorithmic discrete mathematics*, pp. 195–248. Algorithms Combinatorics 16, Springer, Berlin, 1998.

[Men86] M. Menshikov. Coincidence of critical points in percolation problems. *Sov. Math. Dokl.* **33** (1986), 856–859.

[Mes13] P. Mester. Invariant monotone coupling need not exist. *Ann. Probab.* **41** (2013), 1180–1190. `arXiv:1011.2283 [math.PR]`

[Mes11] P. Mester. A factor of i.i.d. with uniform marginals and infinite clusters spanned by equal labels. *Preprint*, `arXiv:1111.3067 [math.PR]`.

[Mil68a] J. Milnor. A note on curvature and fundamental group. *J. Diff. Geom.* **2** (1968), 1–7.

[Mil68b] J. Milnor. Growth of finitely generated solvable groups. *J. Diff. Geom.* **2** (1968), 447–449.

[Moh88] B. Mohar. Isoperimetric inequalities, growth, and the spectrum of graphs. *Linear Algebra Appl.* **103** (1988), 119–131.

[MohW89] B. Mohar and W. Woess. A survey on spectra of infinite graphs. *Bull. London Math. Soc.* **21** (1989), no. 3, 209–234.

[Mok95] N. Mok. Harmonic forms with values in locally constant Hilbert bundles. *Proceedings of the Conference in Honor of Jean-Pierre Kahane (Orsay, 1993). J. Fourier Anal. Appl.* (1995), Special Issue, 433–453.

[MonMS12] A. Montanari, E. Mossel and A. Sly. The weak limit of Ising models on locally tree-like graphs. *Probab. Theory Related Fields* **152** (2012), 31–51. `arXiv:0912.0719 [math.PR]`

[MorP05] B. Morris and Y. Peres. Evolving sets, mixing and heat kernel bounds. *Prob. Th. Rel. Fields* **133** (2005), no. 2, 245–266. `[arXiv:math.PR/0305349]`

[Mor10] R. Morris. Zero-temperature Glauber dynamics on $\mathbb{Z}^d$. *Prob. Th. Rel. Fields* (2010), `arXiv:0809.0353v2 [math-ph]`

[MöPe10] P. Mörters and Y. Peres. *Brownian motion.* Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 30. Cambridge University Press, 2010. `http://people.bath.ac.uk/maspm/book.pdf`.

[MuPa01] R. Muchnik and I. Pak. Percolation on Grigorchuk groups. *Comm. Algebra* **29** (2001), 661–671. `http://www.math.ucla.edu/~pak/papers/perc.ps`

[NaPe10] A. Nachmias and Y. Peres. The critical random graph, with martingales. *Israel J. Math.* **176** (2010), 29–43. `[arXiv:math.PR/0512201]`

[NaNS00] S. Nanda, C. M. Newman and D. Stein. Dynamics of Ising spin systems at zero temperature. In *On Dobrushin's way (From Probability Theory to Statistical Mechanics)*, eds. R. Minlos, S. Shlosman and Y. Suhov, *Am. Math. Soc. Transl.* (2) **198** (2000), 183–194.

[NaoP08] A. Naor and Y. Peres. Embeddings of discrete groups and the speed of random walks. *International Mathematics Research Notices* (2008), article ID rnn076, 34 pages. `arXiv:0708.0853 [math.MG]`

[NasW61] C. S. J. A. Nash-Williams. Edge-disjoint spanning trees of finite graphs. *J. London Math. Soc.* **36** (1961), 445–450.

[Nek03] V. V. Nekrashevych. Iterated monodromy groups. `[arXiv:math.DS/0312306]`

[Nek05] V. V. Nekrashevych. *Self-similar groups.* Mathematical Surveys and Monographs, Vol. 117, Amer. Math. Soc., Providence, RI, 2005.

[NekP09] V. Nekrashevych and G. Pete. Scale-invariant groups. *Groups, Geometry and Dynamics*, to appear. `arXiv:0811.0220v4 [math.GR]`

[NewS96] C. M. Newman and D. L. Stein. Ground-state structure in a highly disordered spin-glass model. *J. Statist. Phys.* **82** (1996), 1113–1132.

[Orn74] D. S. Ornstein. *Ergodic theory, randomness and dynamical systems.* Yale University Press, New Haven, 1974.

[OW74] D. S. Ornstein and B. Weiss. Finitely determined implies very weak Bernoulli. *Israel J. Math.* **17** (1974), 94–104.

[OW75] D. S. Ornstein and B. Weiss. Every transformation is bilaterally deterministic. *Israel J. Math.* **21** (1975), 154–158.

[OW87] D. S. Ornstein and B. Weiss. Entropy and isomorphism theorems for actions of amenable groups. *J. Analyse Math.* **48** (1987), 1–141.

[Osi02] D.V. Osin. Weakly amenable groups. *Computational and statistical group theory*, Contemporary Mathematics, vol. 298, American Mathematical Society, Providence, RI, 2002, pp. 105–113. `http://www.sci.ccny.cuny.edu/~osin/papers/wa.pdf`

[PaSN00] I. Pak and T. Smirnova-Nagnibeda. On non-uniqueness of percolation on nonamenable Cayley graphs *Comptes Rendus de l'Academie des Sciences Series I Mathematics* **330** (2000), No. 6, 495–500. `http://www.math.ucla.edu/~pak/papers/psn.pdf`

[Pel07] R. Peled. On rough isometries of Poisson processes on the line. *Annals of Applied Probability*, to appear. `arXiv:0709.2383v2 [math.PR]`

[Per04] Y. Peres. *Course notes for Probability on Trees and Networks*, UC Berkeley, Fall 2004. Edited by Asaf Nachmias. `http://stat-www.berkeley.edu/~peres/notes1.pdf`

[Per09] Y. Peres. Connectivity probability in critical percolation: An unpublished gem from Oded. *Lecture at the Oded Schramm Memorial Conference*, August 30-31, 2009, Microsoft Research. Viewable with Internet Explorer and Windows Media Player at `http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/17504/lecture.htm`.

[PerPS06] Y. Peres, G. Pete and A. Scolnicov. Critical percolation on certain non-unimodular graphs. *New York J. of Math.* **12** (2006), 1–18. `[arXiv:math.PR/0501532]`

[PerR04a] Y. Peres and D. Revelle. Mixing times for random walks on finite lamplighter groups. *Electron. J. Probab.* **26** (2004), 825–845. `[arXiv:math.PR/0404190]`

[PerR04b] Y. Peres and D. Revelle. Scaling limits of the uniform spanning tree and loop-erased random walk on finite graphs. *Preprint* (2004), `[arXiv:math.PR/0410430]`

[PerW10] Y. Peres, with D. Wilson. *Game Theory: Alive*. Book in preparation, available at `http://dbwilson.com/games/`.

[Pes08] V. G. Pestov. Hyperlinear and sofic groups: a brief guide. *Bulletin of Symbolic Logic* **14** (2008), 449–480. `arXiv:0804.3968v8 [math.GR]`

[Pet08] G. Pete. A note on percolation on $\mathbb{Z}^d$: isoperimetric profile via exponential cluster repulsion. *Elect. Comm. Probab.* **13** (2008), 377–392. `[arXiv:math.PR/0702474v4]`

[Pet09] G. Pete. How to prove tightness for the size of strange random sets? Talk based on [GarPS10a]. Oded Schramm Memorial Conference, Microsoft Research, 2009. `http://math.bme.hu/~gabor/OdedMemorial.pdf`

[Pey08] R. Peyre. A probabilistic approach to Carne's bound. *Potential Analysis* **29** (2008), 17–36. `http://www.normalesup.org/~rpeyre/pro/research/carne.pdf`

[Pin73] M. Pinsker. On the complexity of a concentrator. *7th International Teletraffic Conference*, Stockholm, 1973. 318/1-4.

[Pip77] N. Pippenger. Superconcentrators. *SIAM J. Comput.* **6** (1977), no. 2, 298–304.

[Rau04] A. Raugi. A general Choquet-Deny theorem for nilpotent groups. *Ann. Inst. H. Poincaré Probab. Statist.* **40** (2004), no. 6, 677–683.

[Ros74] J. Rosenblatt. Invariant measures and growth conditions. *Trans. Amer. Math. Soc.* **193** (1974), 33–53.

[Rud73] W. Rudin. *Functional Analysis.* McGraw-Hill, 1973.

[SaC97] L. Saloff-Coste. Lectures on finite Markov chains. *Lectures on probability theory and statistics (Saint-Flour, 1996)*, 301–413. LNM 1665, Springer, Berlin, 1997.

[Sap07] M. V. Sapir. Some group theory problems. *Internat. J. of Algebra and Comput.* **17** (2007), 1189–1214. `arXiv:0704.2899v1 [math.GR]`

[Saw97] S. A. Sawyer. Martin boundaries and random walks. *Harmonic functions on trees and buildings (New York, 1995)*, ed. A. Koranyi, pp. 17–44. Contemporary Mathematics 206, American Mathematical Society, Providence, RI, 1997. `http://www.math.wustl.edu/~sawyer/hmhandouts/martbrwf.pdf`

[Scho92] R. H. Schonmann. On the behavior of some cellular automata related to bootstrap percolation. *Ann. Probab.* **20** (1992), no. 1, 174–193.

[Scho01] R. H. Schonmann. Multiplicity of phase transitions and mean-field criticality on highly non-amenable graphs. *Comm. Math. Phys.* **219** (2001), no. 2, 271–322.

[Scho02] R. H. Schonmann. Mean-field criticality for percolation on planar non-amenable graphs. *Comm. Math. Phys.* **225** (2002), no. 3, 453–463.

[Schr07] O. Schramm. Conformally invariant scaling limits: an overview and a collection of problems. In *International Congress of Mathematicians (Madrid, 2006), Vol. I*, pages 513–543. Eur. Math. Soc., Zürich, 2007. `[arXiv:math.PR/0602151]`

[Schr08] O. Schramm. Hyperfinite graph limits. *Elect. Res. Announce. Math. Sci.* **15** (2008), 17–23. `arXiv:0711.3808 [math.PR]`

[SchrSt10] O. Schramm and J. E. Steif. Quantitative noise sensitivity and exceptional times for percolation. *Ann. Math.* **171** (2010), 619–672. `[arXiv:math.PR/0504586]`

[Sha04] Y. Shalom. Harmonic analysis, cohomology, and the large-scale geometry of amenable groups *Acta Math.* **192** (2004), 119–185.

[ShaT10] Y. Shalom and T. Tao. A finitary version of Gromov's polynomial growth theorem. *Geom. Funct. Anal.* **20** (2010), 1502–1547. `arXiv:0910.4148 [math.GR]`.

[She06] S. Sheffield. Uniqueness of maximal entropy measure on essential spanning forests. *Ann. Probab.* **34** (2006), 857–864. `[arXiv:math.PR/0406513]`

[Sid00] S. Sidki. Automorphisms of one-rooted trees: growth, circuit structure, and acyclicity. *J. Math. Sci. (New York)* **100** (2000) no. 1, 1925–1943.

[Sla02] G. Slade: Scaling limits and super-Brownian motion. *Notices of Amer. Math. Soc.* **49** (2002), no. 9, 1056–1067. `http://www.ams.org/notices/200209/index.html`

[Sla06] G. Slade. *The lace expansion and its applications.* Lecture Notes in Mathematics 1879, xiv + 232 pages. Springer, Berlin, 2006.

[Sly10] A. Sly. Computational transition at the uniqueness threshold. Extended abstract to appear in the *Proceedings of FOCS 2010*. *Preprint* at `arXiv:1005.5584 [cs.CC]`.

[Sok01] A. D. Sokal. Bounds on the complex zeros of (di)chromatic polynomials and Potts-model partition functions. *Combinatorics, Probability and Computing* **10** (2001), 41–77. `[arXiv:cond-mat/9904146]`

[Špa09] I. Špakulova. Critical percolation of virtually free groups and other tree-like graphs. *Ann. Probab.* **37** (2009), no. 6, 2262–2296. `[arXiv:math.PR/0801.4153]`

[Spi75] F. Spitzer. Markov random fields on an infinite tree *Ann. Probab.* **3** (1975), 387–398. `http://www.jstor.org/stable/2959462`

[Sta68] J. Stallings. On torsion-free groups with infinitely many ends. *Ann. of Math.* **88** (1968), 312–334.

[SwW87] R. H. Swendsen and J. S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett.* **58** (1987), no. 2, 86–88.

[Tal94] M. Talagrand. On Russo's approximate zero-one law. *Ann. Probab.* **22** (1994), 1576–1587.

[Tal96] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. de l'I.H.E.S.* **81** (1996), 73–205.

[Tao10] T. Tao. A proof of Gromov's theorem. *What's new* blog entry, February 2010, `http://terrytao.wordpress.com/2010/02/18/a-proof-of-gromovs-theorem/`

[Thm08] A. Thom. Sofic groups and diophantine approximation. *Comm. Pure Appl. Math.* **LXI** (2008), 1155–1171. `[arXiv:math.FA/0701294]`

[Ths83] C. Thomassen. Girth in graphs. *J. Combin. Theory Ser. B* **35** (1983), 129–141.

[Ths92] C. Thomassen. Isoperimetric inequalities and transient random walks on graphs. *Ann. Probab.* **20** (1992), 1592–1600.

[Tim06a] Á. Timár. Neighboring clusters in Bernoulli percolation. *Ann. Probab.* **34** (2006), no. 6. 2332–2343. `[arXiv:math.PR/0702873]`

[Tim06b] Á. Timár. Percolation on nonunimodular transitive graphs. *Ann. Probab.* **34** (2006), no. 6, 2344–2364. `[arXiv:math.PR/0702875]`

[Tim07] Á. Timár. Cutsets in infinite graphs. *Combin. Probab. & Comput.* **16** (2007), no. 1, 159–166. `arXiv:0711.1711 [math.CO]`

[Tim12] Á. Timár. Approximating Cayley diagrams versus Cayley graphs. *Combin., Probab. & Comput.* **21** (2012), 635–641. `arXiv:1103.4968 [math.CO]`

[Tit72] J. Tits. Free subgroups in linear groups. *J. Algebra* **20** (1972), 250–270.

[Tro85] V. I. Trofimov. Graphs with polynomial growth. *Math. USSR Sbornik* **51** (1985), 405–417.

[Var85a] N. Th. Varopoulos. Isoperimetric inequalities and Markov chains. *J. Funct. Anal.* **63** (1985), 215–239.

[Var85b] N. Th. Varopoulos. Long range estimates for Markov chains. *Bull. Sci. Math. (2)* **109** (1985), no. 3, 225–252.

[VarSCC92] N. Th. Varopoulos, L. Saloff-Coste and T. Coulhon. *Analysis and geometry on groups.* Cambridge University Press, Cambridge, 1992.

[Ver00] A. M. Vershik. Dynamic theory of growth in groups: Entropy, boundaries, examples. *Russ. Math. Surveys* **55** (2000), 667–733.

[Vir02] B. Virág. Fast graphs for the random walker. *Probab. Theory Related Fields* **124** (2002), 50–74. `[arXiv:math.PR/0102200]`

[Wag93] S. Wagon. *The Banach-Tarski paradox.* With a foreword by Jan Mycielski. Corrected reprint of the 1985 original. Cambridge University Press, Cambridge, 1993.

[Wei00] B. Weiss. Sofic groups and dynamical systems. Ergodic theory and harmonic analysis (Mumbai, 1999). *Sankhyā: The Indian J. of Stat., Ser. A* **62** (2000), no. 3, 350–359.

[Wer04] W. Werner. *Random planar curves and Schramm-Loewner evolutions.* 2002 Saint-Flour Summer School, Lect. Notes in Math. 1840, pp. 107–195. Springer, 2004. `[arXiv:math.PR/0303354]`

[Wer07] W. Werner. *Lectures on two-dimensional critical percolation*, IAS Park City Graduate Summer School, 2007. `arXiv:0710.0856v2 [math.PR]`

[Wik10a] English Wikipedia. Circle packing theorem. Version of May 7, 2010. `http://en.wikipedia.org/wiki/Circle_packing_theorem#A_uniqueness_statement`

[Wik10b] English Wikipedia. FKG inequality. Version of May 7, 2010. `http://en.wikipedia.org/wiki/FKG_inequality`

[Wil96] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing*, pages 296–303, New York, ACM, 1996.

[Wil04] D. B. Wilson. Mixing times of Lozenge tiling and card shuffling Markov chains. *Ann. Appl. Probab.* **14** (2004), 274–325. `[arXiv:math.PR/0102193]`

[Wil98] J. S. Wilson. *Profinite groups.* London Mathematical Society Monographs. New Series, 19. The Clarendon Press, Oxford University Press, New York, 1998.

[Wil09] H. Wilton. *392C Geometric Group Theory*, taught at the University of Texas at Austin, Spring 2009. `http://392c.wordpress.com/2009/01/`

[Woe00] W. Woess. *Random walks on infinite graphs and groups.* Cambridge Tracts in Mathematics Vol. 138, Cambridge University Press, 2000.

[Woe05] W. Woess. Lamplighters, Diestel-Leader graphs, random walks, and harmonic functions. *Combin. Probab. & Comput.* **14** (2005), 415–433.

[Wol68] J. Wolf. Growth of finitely generated solvable groups and curvature of Riemannian manifolds. *J. Diff. Geom.* **2** (1968), 421–446.

[Żuk00] A. Żuk. On an isoperimetric inequality for infinite finitely generated groups. *Topology* **39** (2000), no. 5, 947–956.