Budapest University of Technology and Economics
Institute of Mathematics
Department of Stochastics

MSc Thesis

Data-Driven Analysis of Fractality and Other Characteristics of Complex Networks

Marcell Nagy

Supervisors:
Roland Molontay
PhD Student, Department of Stochastics
Budapest University of Technology and Economics

Prof. Károly Simon
Head of Department of Stochastics
Budapest University of Technology and Economics

2018

Contents

1 Introduction

Complex networks have been in the focus of research for decades, especially since the millennium owing to the fact that the prompt evolution of information technology made the comprehensive exploration of real networks possible. The study of networks pervades all of science, such as Biology (e.g. neuroscience networks), Chemistry (e.g. protein interaction networks), Physics, Information Technology (e.g. WWW and Internet), Economics (e.g. interbank payment flows) and Social Sciences (e.g. collaboration and social networks).

Despite the fact that networks can originate from different domains, they share a few common characteristics such as scale-free and small-world property [1, 2, 3], high clustering [4, 5] and sparseness [6], i.e. they differ in many ways form the completely random graphs introduced by Erdős and Rényi [7]. Scale-free property means, that the degree distribution follows a power-law, small-world behaviour refers to the fact that the diameter is relatively small compared to the size of the network i.e. the average distance between vertices scales logarithmically with the number of nodes. High clustering means there is a high probability that "the friend of my friend is also my friend" [2], in topological terms this means that there is a heightened density of triangles (cycles of length three or closed triplets) in the network [8]; while sparseness means that there are much smaller number of edges than the maximum possible number of links within the same network [9].

Although many real networks have been claimed to be scale-free, some statistical and theoretical research have argued against its ubiquity [10, 11, 12, 13]. The reasons behind this controversy are non-rigorous methods for power-law fitting since it may be confused with log-normal, exponential or truncated power-law distributions, furthermore reliance on insufficient domain specific datasets and the ambiguity regarding the notion of scale-free property [12].

Analysis of variety of real networks showed that there are other essential frequently emerging properties such as self-similarity and fractality, for example protein interaction networks are typically fractal [14]. The latter one is in the focus of this thesis and in a nutshell it means that there is a power-law relation between the minimum number of boxes needed to cover the entire network and the size of the boxes. In this work we will investigate the origins of fractality and

its connection to other graph metrics. My BSc thesis [15] is also devoted to fractal networks, specifically to the relationship of fractality and assortativity, and I showed through a network model, that hubs can be connected in fractal networks i.e. they can show assortative mixing behaviour, which is in contrast to the existing results and claims [14, 16, 17], furthermore there are a few works which support my observation [18, 19, 20].

Modelling real-networks is of great interest, since it may help to understand the underlying mechanisms and principles governing the evolution of networks. Moreover, such models are mathematically tractable and allow for rigorous analysis. Throughout the years several network models have been proposed to gain better understanding of real-world networks, the [21] paper gives an extensive overview of such network models, however without attempting to be comprehensive the most influential models are for example the scale-free Barabási–Albert model [22], the small-world Watts–Strogatz model [2], Newman and Park's Community Structure model [4] and Geographical or Spatial models [23], each of them was motivated by some of the aforementioned observed characteristics of real-networks.

In order to characterize the topology and capture the structure of networks, numerous graph metrics have been introduced, the Network Science book of Barabási Albert and the Characterization of complex networks: A survey of measurements article of L.F. Costa et al give a comprehensive overview of the graph metrics and measurements [9, 23]. Naturally, there is significant redundancy among these measures, unfortunately, it is still unclear which non-redundant selection of measurements describes every aspects of networks. There is a great deal of effort to study the correlation between these metrics together with identify a non-redundant subset of them [24, 25, 26], as well as to construct such models, which better explain real networks according to these measures and the distribution of node-level features.

The main purpose of Section 2 is to understand the fractality with the help of mathematical network models. To this end we investigate several network models based on simulations implemented in Wolfram Mathematica. Not only do we study network models from the literature, but also introduce new models based on our discoveries regarding fractality. We analyze how the fractality of the model generated graphs affects other graphs metrics such as the mean graph distances

and assortativity. We also relieve the the contradiction of two articles [14] and [18], which both presented fractal models to support their conflicting observations and statements. Furthermore we highlight a group of real-world fractal networks that are rather uninvestigated, and propose a novel model which mimics the properties of these networks and mixes it with other aforementioned common characteristics such as small-world and scale-free properties.

Furthermore, we thoroughly investigate the box-covering algorithm and its alternatives, and by our own implemented program codes, we show that how these different algorithms perform on different graphs considering both efficiency and running time. We also highlight that, due to the NP-hard nature of the box-covering algorithms, unfortunately there is a trade-off between the accuracy and running time, meaning that we cannot simultaneously minimize the running time and gain optimal results, but we present recent promising research results that uses novel techniques to estimate the fractal dimension of networks.

As the title of this work suggests, our approach is mostly empirical, i.e. data-based, thus we will use both descriptive statistics and statistical learning techniques in order to analyze the relationship of the metrics and their effect on fractality. However, we also associate our empirical observations to theoretical models of the literature an we introduce new models based on our findings

Machine learning is an interdisciplinary field combining the methods of statistics, computer science and information technology, evolved from pattern recognition and computational learning theory in artificial intelligence, which uses statistical techniques to study and construct algorithms usually to learn from and to make predictions on data [27]. Machine learning tasks are typically divided into two main categories; supervised learning and unsupervised learning. The prediction problem belongs to the so-called supervised learning task (see Section 3.3), on the other hand unsupervised learning involves problems such as clustering, anomaly detection, dimension reduction and feature selection.

The concept of Data science (formerly Data mining) does not have a generally accepted definition but it can be described as a generic term for data understanding, data visualization, data preparation (i.e. cleansing and transforming), machine learning and modelling, result validation and deploying. The applications of data science are present in all aspects of our everyday life, furthermore there

are more and more applications in scientific research from different disciplines such as high energy physics[1], astrophysics[2], healthcare and biology[3]. To bring examples from everyday life, all the search engines use machine learning algorithms to deliver the best result for our searched query. Once we searched for something online, the next few days every digital commercial will be related to it, this is called targeted advertising. One of the most sophisticated application of data science is the recommendation system, which suggests us similar products, songs[4], videos[5] and TV shows[6], based on our past behaviour and taste. Without attempting to be exhaustive machine learning algorithms are applied in face and speech recognition, fraud and risk detection, and seeing into the near-future days, self-driving cars and human-like robots also apply data science techniques.

There is an emerging discipline of data science, called Educational Data Mining[7] (EDM), that uses and develops data science methods to answer educational research questions such as identifying the key factors of dropout, understand the study behaviours of students or recommend appropriate courses and training sessions. Relying on the database of Budapest University of Technology and Economics, the author and the first supervisor of the present paper, employed and evaluated several machine learning models to identify students at-risk, to predict student dropout and to determine the affecting factors of the students' performance, for more details see [28].

The purpose of Section 3 is to study the relationship of numerous network metrics and how the network characteristics effect fractality on different network domains using both models and real networks. To this end we use statistical methods such as explanatory analysis, correlation analysis, and machine learning techniques e.g. predictive analytics. This study relies on a large dataset, containing rich set of information of 584 real networks from different domains. The

[1]See `https://sites.google.com/site/trackmlparticle/`

[2]See `https://ieeexplore.ieee.org/document/6382200/`

[3]See `https://www.techemergence.com/machine-learning-medical-diagnostics-4-current-applications/`

[4]See `https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe`

[5]See `https://dl.acm.org/citation.cfm?id=1864770`

[6]See `https://medium.com/netflix-techblog/tagged/data-science`

[7]See `http://educationaldatamining.org/`

real networks for the dataset are from online network data-bases, such as Network Repository [29], Index of Complex Networks (ICON) [30], NeuroData's Graph Data-base [31], The Koblenz Network Collection [32] and Interaction Web Database (IWDB) [33]. After evaluating the statistical and machine learning analyzes, we compare the results obtained on real and model networks to test the descriptive/explanatory ability of the models. Furthermore, we aspired to find the mutual and different behaviors of real and model networks.

The most closely related work is that of Garcia-Robledo et al. [26] who followed a data-driven approach to study the correlation of different properties of evolving Internet networks, furthermore, applied clustering techniques (k-means and Ward's method) to find and validate a non-redundant set of metrics. Bounova and de Weck gave a great overview of network topology metrics and their normalization and used correlation and principal component analysis on a dataset consisting of both random graph models and 200-300 real networks [24]. Jamakovic and Uhlig investigated the correlation of metrics and visual comparison of distribution node-level features using 13 real networks [25]. Filkov et al. similarly used both models and a collection of 113 real networks to find a set of metrics, which enables comprehensive comparison between any two networks [34]. Grando and Lamb trained machine learning algorithms on a huge dataset derived from network models to estimate centrality measures [35].

In this paper, we focus on a more general and comprehensive research of real networks i.e. we collected brain networks, food webs, social networks (collaboration, Facebook friendship and retweet networks), protein interaction networks, and synthetic networks such as the collection of DIMACS [36] and sparse matrices from SuiteSparse [37]. Furthermore, our goal is to find the attributes, which affects the fractality of the networks, and then construct a new fractal network model, inspired by the newly found properties and relationships. To the best of our knowledge, this is the first work that uses Machine Learning techniques in order to estimate the fractal dimension of networks.

1.1 Definitions and notations

In this subsection we introduce the most important definitions and fix the notations used throughout this paper. Since networks can be modeled by graphs, the notions of network theory root in graph theory. Note that in this work the graph and network words are interchangeable, but usually when we talk about networks, we focus on the real, physical properties, on the other hand in the case of graphs, the bare mathematical characteristics are under consideration. Regarding to the fact that network theory is a fresh field in the intersection of graph theory and computer science, and researched by scientists from different disciplines, the definitions are not always mathematically rigorous. Here we rely on [9, 38] and [14].

Definition 1. ***(Graph)*** *A simple (undirected) graph is an ordered pair* $G = (V, E)$*, where* V *is the set of vertices or nodes, together with a set* E *of edges or links, which are two-element subsets of* V*. The size of the graph is the number of its nodes, and it is usually denoted by* n*.*

Note that there are directed and weighted graphs as well, and the following definitions can be generalized or modified to those cases, what's more real networks often modelled with directed and weighted graphs, but in this work we only consider the simplified versions of those networks, hence here we only state definitions for simple graphs.

Definition 2. ***(Path)*** *A path is a sequence of edges which connect a sequence of vertices i.e. the target of the previous edge is the source of the next edge. Formally: a path is a sequence of vertices* $P = (v_1, v_1, \ldots, v_n) \in V \times \ldots \times V$ *such that* v_i *is adjacent to* v_{i+1} *for* $1 \leq i \leq n$*. Such a path* P *is called a path of length* $n-1$ *from* v_1 *to* v_n *i.e. the number of its links. A path is geodesic or* ***shortest path*** *if its endpoints cannot be connected by shorter paths.*

Definition 3. ***(Distance)*** *The distance* $d(u, v)$ *between the vertices* u *and* v *is the length (number of edges) of the shortest path connecting them. Note that the vertex set (of an undirected graph) and the distance function* d *form a metric space, if and only if the graph is connected.*

Definition 4. ***(Vertex eccentricity)*** *The vertex eccentricity $\epsilon(v)$ of a vertex v (in a connected graph G) is the maximum graph distance between v and any other vertex u of G.*

Definition 5. ***(Radius)*** *The radius r of a graph is the minimum eccentricity of any vertex, i.e.*

$$r = \min_{v \in V} \epsilon(v)$$

Definition 6. ***(Diameter)*** *The diameter $\mathrm{Diam}(G)$ of a graph G is the maximum eccentricity of any vertex in the graph, i.e.*

$$\mathrm{Diam}(G) = \max_{v \in V(G)} \epsilon(v).$$

In other words the diameter of a graph is the length of the greatest shortest path.

Definition 7. ***(k-neighbourhood)*** *The k-neighbourhood Γ_v^k of the vertex v is the set of vertices u whose distance from v is not greater than k.*

Definition 8. ***(Proportionality)*** *Given two variables x and y, we say that y is directly proportional to y if there is always a constant ratio between them, i.e. if there is a non-zero constant c such that $y = c \cdot x$. The c constant is called the coefficient of proportionality or proportionality constant. In this paper we denote this relation by $y \propto x$ or by $x \sim y$.*

Definition 9. ***(Small-world property)*** *A network is said to be small-world, if the "typical" distance L (i.e. the average length of short paths) between any two nodes grows proportionally to the logarithm of the size of the network i.e. $L \propto \log |V|$. Note that scale-free networks are ultra-small worlds [39], i.e. due to hubs, the shortest paths become significantly smaller and scale as $L \propto \log \log |V|$*

In graph theory and network analysis, indicators of centrality identify the most important and influential nodes within a graph. There are numerous centrality metrics, though here we only focus on the most frequently used ones, since these metrics are usually highly-correlated.

Definition 10. ***(Betweenness centrality)*** *The betweenness centrality of a node v is given by the expression:*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Definition 11. ***(Edge betweenness centrality)*** *The edge betweenness centrality of an edge is the number of shortest paths between pairs of vertices that run along it. In other words this is analogous to the previously defined σ_{st}, but here we consider an edge instead of a node.*

Definition 12. ***(Eigenvector centrality)*** *For a (connected undirected) graph, the vector of eigenvector centralities $\mathbf{c}$ satisfies the eigenvector equation $\mathbf{A} \cdot \mathbf{c} = \lambda_1 \mathbf{c}$, where λ_1 is the largest eigenvalue of the graph's adjacency matrix $\mathbf{A}$. In other words, for a connected undirected graph, the vector of eigenvector centralities is given by the (suitably normalized) eigenvector of corresponding to its largest eigenvalue Note that eigenvector centrality is a normalized special case of Katz centrality with $\alpha = 1/\lambda_1$ and $\beta = 0$. A related centrality is PageRank centrality.*

Definition 13. ***(Degree distribution)*** *The degree $d(v)$ of a vertex v in a graph is its number of incident edges. The degree distribution P is the probability distribution of these degrees over the whole network, i.e. $P(k)$ is the probability that the degree of a randomly chosen vertex is equal to k.*
The correlations between degrees in a graph are frequently measured by the the joint probability distribution $P(k_1, k_2)$: the probability that a node with degree k_1 is connected to another node of degree k_2 [40].

Definition 14. ***(Scale-free property)*** *A scale-free network is a connected graph with the property that the $P(k)$ degree distribution follows power-law distribution, i.e.*

$$P(k) \sim k^{-\gamma},$$

where $\gamma \geq 1$ and typically falls in the range $2 < \gamma < 3$ [41].

Definition 15. ***(Connected graph)*** *A connected graph is one in which each pair of vertices forms the endpoints of a path, i.e. there is a path from any point to any other point in the graph. A graph that is not connected is said to be disconnected.*

Definition 16. ***(Vertex and edge connectivity)*** *The vertex connectivity of a graph is the minimum number of nodes whose deletion disconnects it. Similarly, edge connectivity is the minimum number of edges whose deletion from a graph disconnects it.*

Definition 17. ***(Link efficiency)*** *The link efficiency measures how tightly connected the graph is in relation to its number of edges. Let L denote the average of all shortest paths length. For a simple unweighted graph G, the link efficiency $E(G)$ of G is given by:*

$$E(G) = 1 - \frac{L}{|E|}$$

Definition 18. ***(Graph density)*** *Graph density D is the ratio of the number of edges divided by the number of edges of a complete graph with the same number of vertices, i.e:*

$$D = \frac{|E|}{\frac{1}{2}|V|(|V|-1)}.$$

A dense graph is a graph in which the number of edges is close to the maximal number of edges, i.e D is close to 1. The opposite, a graph with only a few edges, is a sparse graph, when D is close to 0.

Definition 19. ***(Variance)*** *The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = \mathbb{E}[X]$:*

$$\operatorname{Var}(X) = \sigma_X^2 = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}[X^2] - \mu^2.$$

The standard deviation σ_X of X is the square root of the variance of X.

Definition 20. ***(Covariance)*** *The covariance between two jointly distributed real-valued random variables X and Y (with finite second moments) is defined as the expected product of their deviations from their individual expected values:*

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[\big(X - \mathbb{E}\left[X\right]\big)\big(Y - \mathbb{E}\left[Y\right]\big)\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Definition 21. ***(Pearson's correlation coefficient)*** *The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as*

$$\rho_{X,Y} = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Definition 22. ***(Assortativity coefficient)*** *The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes [40]. The assortativity coefficient is given by*

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2},$$

where the term q_k is the distribution of the remaining degree and j and k indicates the remaining degrees, this captures the number of edges leaving the node, other than the one that connects the pair, i.e. the degree of the node minus one. Furthermore, e_{jk} refers to the joint probability distribution of the remaining degrees of the two vertices, thus e_{jk} is symmetric on an undirected graph, and follows the sum rule $\sum_{jk} e_{jk} = 1$, and $\sum_j e_{jk} = q_k$, i.e. q_k is the marginal distribution of e_{jk}. Finally, σ_q^2 denotes the variance of the q_k distribution, i.e. $\sigma_q^2 = \sum_k k^2 q_k - \left(\sum_k k q_k\right)^2$

Definition 23. ***(R degree correlation ratio)*** *For a given G graph, the $R(k_1, k_2)$ degree correlation ratio is defined as follows:*

$$R(k_1,k_2) = \frac{P(k_1,k_2)}{P_r(k_1,k_2)},$$

where $P(k_1, k_2)$ and $P_r(k_1, k_2)$ are joint degree distributions of G (see Definition 13), and of a random graph, obtained by randomly (uniformly) swapping the links of G without modifying the original degree distribution respectively, i.e. $P_r(k_1, k_2)$ is the joint degree distribution of the so-called Configuration model [42], which has the same distribution as G.

Definition 24. ***(Global clustering coefficient)*** *The global clustering coefficient C of the graph G is the fraction of paths of length two (triplets) in G that are closed over all paths of length two (closed triplets) in G.*

Definition 25. ***(Local clustering coefficient)*** *The local clustering coefficient of the vertex v is the fraction of pairs of neighbors of v that are connected over all pairs of neighbors of v. Formally:*

$$C_{loc}(v) = \frac{|\{(s,t) \text{ edges} : s,t \in \Gamma_v^1 \text{ and } (s,t) \in E|\}}{d(v)(d(v)-1)}$$

1.1.1 Probability theory definitions

The following definitions and theorems are necessary for the proofs of Theorem 7 and 8. We assume that the reader is familiar with the basic concepts of probability and measure theory, here we follow Williams' book [43].

Definition 26. ***(Filtered space)*** *A filtered space is $(\omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$, where $(\omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\{\mathcal{F}_n\}_{n=0}^{\infty}$ is a filtration. This means:*

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}$$

is an increasing sequence of sub σ-algebras of $\mathcal{F}$.

When we say simply "process" in this work, we mean discrete time stochastic process, i.e. a sequence of random variables.

Definition 27. ***(Adapted process)*** *We say that the process $M = \{M_n\}_{n=0}^{\infty}$ is adapted to the filtration $\{\mathcal{F}_n\}$ if $\forall n \in \mathbb{N} M_n \in \mathcal{F}_n$, i.e. M_n is an $\mathcal{F}_n$-measurable function.*

Definition 28. ***(Martingale)*** *Let $M = \{M_n\}_{n=0}^{\infty}$ be an adaptive process to the filtration $\{\mathcal{F}_n\}$. We say that M is a martingale if*

(i) $\mathbb{E}(|M_n|) < +\infty$, $\forall n$

(ii) $\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) = M_{n-1}$ *almost surely for* $n \geq 1$

Furthermore, we say that M is supermartingale if we substitute (ii) *with*

$$\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) \leq M_{n-1},$$

almost surely as $n \geq 1$*, and finally* M *is submartingale is we substitute* (ii) *with*

$$\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) \geq M_{n-1},$$

almost surely as $n \geq 1$*.*

Definition 29. ***(Bounded martingale)*** *Let* $M = (M_n)$ *be a martingale. We say that* $M_n \in L^k$*, i.e.* M_n *is bounded in* L^k*, for some* $k \geq 1$ *if*

$$\sup_n \mathbb{E}(|M_n|^k) < +\infty \tag{1}$$

Theorem 1. ***(Doob's Forward Convergence Theorem)*** *Let* $X = (X_n)$ *be an* L^1 *bounded supermartingale. Then*

$$X_\infty = \lim_{n\to\infty} X_n$$

exists amd $X_\infty < \infty$ *almost surely.*

Theorem 2. ***(Doob–Meyer decomposition)*** *Given a filtered probability space* $(\omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$*. Let* $X = (X_n)$ *be an adapted process with* $X_n \in L^1$ *for all* n*. Then* X *has a Doob–Meyer (sometimes called Doob decomposition):*

$$X = X_0 + M + A, \tag{2}$$

where $M = (M_n)$ *is a martingale with* $M_0 = 0$*,* $A = (A_n)$ *is previsible (that is* $A_n \in \mathcal{F}_{n-1}$*), with* $A_0 = 0$*. (*A_n *is called compensator of* X_n*). The decomposition is unique mod zero, i.e. if* $X = X - 0 + \tilde{M} + \tilde{A}$ *is another decomposition, then*

$$\mathbb{P}(M_n = \tilde{M}_n, A_n = \tilde{A}_n, \forall n) = 1.$$

Theorem 3. X *is a submartingale if and only if* A *in its Doob decomposition is an increasing process, that is*

$$\mathbb{P}(A_n \leq A_{n+1}) = 1. \tag{3}$$

Definition 30. ***(Moment generating function)*** *The moment generation function of a random variable X is defined as*

$$M_X(t) = \mathbb{E}\left(e^{tX}\right), \quad t \in \mathbb{R},$$

wherever this expectation exists.

Definition 31. ***(Factorial moment)*** *For a natural number r, the rth factorial moment of a random variable X is*

$$\mathbb{E}\left((X)_r\right) = \mathbb{E}\left(X(X-1)(X-2)\dots(X-r+1)\right),$$

where

$$(x)_r = x(x-1)(x-2)\dots(x-r+1) = \frac{x!}{(x-r)!}$$

Theorem 4. ***(Markov's inequality)*** *If X is a nonnegative random variable, and $a > 0$, then*

$$\mathbb{P}(X \geq) \leq \frac{\mathbb{E}(X)}{a} \tag{4}$$

This is useful for getting exponential upper-bounds as follows

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}. \tag{5}$$

Eq. (5) *is usually referred to as exponential Chebysev or exponential Markov inequality.*

2 Fractal networks

In this section, we give an overview of the concept of fractal networks largely relying on [44], furthermore, without attempting to be comprehensive, we give an survey of the most related works about fractal networks.

Since the 20th century fractal structures have been in the focus of research, and became one of the most influential results of mathematics, due to the fact that fractal phenomena and structures are present in several disciplines such as physics [45], chemistry [46], cosmology [47], and even in stock market movements [48]. Furthermore, due to the spectacle of fractals, they can even be found in a field of algorithmic art, called fractal art [49]. The concept of fractality and self-similarity has been introduced by Benoit Mandelbrot [50], furthermore, the concept of fractality was generalized to natural phenomena such as shapes of leaves [51], coastlines [52], snowflakes [53] and clouds [54].

In recent years, fractal networks have been studied intensively, and gained great attention from researchers belonging to different fields. The first works date back to the '80-'90's [55], although most of the research have been done in the last few years. There has been a substantial amount of work done by C. Song, S. Havlin, H. Makse, L. Gallos and H. Rozenfeld [14, 16, 56, 57, 58], which can be considered as the foundation stones of the study of self-similar and fractal complex networks. These works inspired and influenced several scientists, resulting in a growing research interest in this field. In [16] they introduced the basic concepts and relations, also presented some approaches to explore the origins of fractality, furthermore, showed that many real-networks show fractal nature such as World-Wide-Web, cellular (protein interaction) and actor collaboration networks. In [56] they presented, compared and studied a number of box-covering algorithms in detail.

As we have already mentioned in Section 1, in my BSc thesis, we studied two conflicting papers; in the first one [14] the aforecited authors claimed that the key principle that gives rise to fractal architecture of networks is a strong repulsion between hubs, furthermore introduced the Dynamical Growth model (DG model) which supports their observation (see 2.4.3); in the other one L. Kuang et al. [18] showed that that hubs can be connected in fractal networks by modifying the

DG model, called Hub Attraction dynamical growth model (HADG model, see 2.4.4), although their log-log plot of $N_B(l_B)$ is bowed downwards which can reflect a log-normal distribution, instead of a power-law [59]. In [15] we introduced a second variant of the DG model, called Repulsion Based DG model (see 2.4.5), which strength lies in the fact that it constructs fractal graph with any parameter setting, while the repulsion behavior varies between group of nodes with similar degrees i.e. it shows that in fractal networks hubs can be connected until it does not reduces the average distances significantly.

In recent works S.R. de la Torre, J. Kalda et al. [60] analyzed fractal and multifractal properties of Estonia's payment network, which is the first study that analyzes multifractality of a complex network of payments. In [61] Z.J. Zeng, C. Xie et al. investigated the fractal property of stock market network using edge-covering technique which is an alternative of the box-covering method introduced in [62]. C. Yuan et al. [63] extensively investigated the properties of wireless networks (2G and 3G), and among other results they showed that these networks are scale-free, small-world and fractal. Similarly, S. Deng et al. [64] estimated the fractal dimension of metro network of large cities and Y. Deng et al. in [65] compared different box-covering algorithms by evaluating them on fractal real networks.

2.1 Box-covering algorithm

In fractal geometry, the box-covering algorithm is one possible way to estimate the fractal dimension of a fractal. Let the fractal be a set S in the n-dimensional Euclidean space $\mathbf{R}^n$ or more generally in a metric space (M, d). Now, imagine that the S fractal is lying on an evenly spaced n-dimensional grid, and the hypercubes of this grid are called boxes. To calculate the fractal dimension of S, we have to apply the box-covering or box-counting algorithm, i.e. count the minimum number of boxes that are required to cover the entire set S. The dimension is calculated by observing how this number changes as we make the grid finer and finer i.e. how this number scales with the size of the boxes [66]. The Figure 1 illustrates the process of box-covering algorithm on a fern leaf. Note that in the 1960's Mandelbrot calculated the fractal-dimension of different coastlines in [52], which

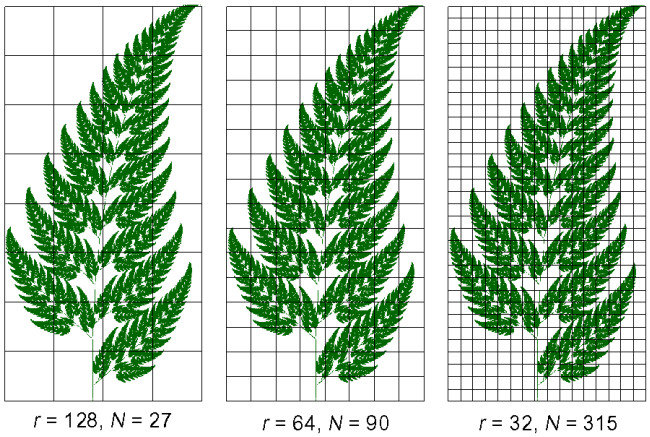


Figure 1: Estimating the box-counting dimension of a fern leaf generated with iterated function systems [67]

.

paper became one of the most influential work in the history of fractal geometry.

Suppose that N_ε is the minimum number of boxes of size ε needed to cover the set. Then the S fractal's box-counting dimension is defined as follows:

$$\dim_{\text{box}}(S) = \lim_{\varepsilon \to 0} \frac{\log N_\varepsilon(S)}{-\log \varepsilon}$$

The box-counting dimension was originally introduced by Hemann Minkowski and Georges Bouligand, thus it is also known as Minkowski–Bouligand dimension.

Estimating the fractal dimension of a network is analogous to the geometric case, owing to the fact that the box-counting algorithm can be easily generalized to networks, because as we mentioned in Definition 3, the vertex set of an undirected graph and the graph distance function form a metric space, and the box-covering algorithm is well-defined in a metric-space. The method works as follows [56]: For a given network G, we partition the nodes into boxes of size l_B. A box is a subgraph of G with diameter smaller than l_B, i.e. a set of nodes, where all distances between any two nodes within the box is less than l_B. The minimum number of boxes of size l_B required to cover the entire network G is denoted by $N_B(l_B)$. Clearly, if $l_B = 1$, then $N_B(1)$ equals to the number of vertices, i.e. the size of the network, while provided that l_B is greater than the diameter of the network, then obviously only one box is needed. Thus, in order to identify the fractal dimension of a network G,

we only have to apply the box-covering algorithm Diam(G) times, starting with $l_B = 1$, and then increasing the size of the boxes up to the diameter of G.

In accordance with geometric fractals, the box-covering dimension d_B can be defined by:

$$N_B(l_B) \sim l_B^{-d_B}, \tag{6}$$

whether this d_B exists and finite, we say that a networks fractal, otherwise it is non-fractal. Thus, in fractal networks the minimum number of boxes scales as a power law with the size of the boxes. Hence, the relationship between the logarithm of N_B and the logarithm of l_B is linear i.e. if a network is fractal, then the log-log plot of the $N_B(l_B)$ function is a straight line, with slope d_B, otherwise the network is non-fractal. Rearranging the formula (6), the fractal dimension can be expressed as:

$$d_B(G) \sim \frac{\log N_B^G(l_B)}{-\log l_B}, \tag{7}$$

where $l_B > 1$.

The box-covering problem is an optimization problem, with an input pair (G, l_B), and the task is to find a box-covering, which uses the fewest boxes. Unfortunately, this problem belongs to the family of NP-hard problems, since it can be mapped on to the vertex coloring problem [56], which is one of the most famous NP-hard problems of graph theory. This means that an algorithm that could give the exact optimal number of boxes in relatively short amount of time does not exist. The polynomial-time reduction is detailed in [44].

Theorem 5. *The box-covering is an NP-hard problem.*

However, there are several approximating algorithms, the most frequently used one is the Compact Box Burning (CBB) algorithm, and in this work we also implemented and use thus algorithm. For more detailed information please refer to [56, 15]. Although, it is an approximating algorithm, it is still time-consuming, hence M. Kitsak, S. Havlin et al. introduced a lower-approximating, simplified version of the CBB algorithm, which has a trade-off between accuracy and time consumption i.e. the estimated N_B of the simplified CBB algorithm is always less then the true minimum number of boxes that are needed to cover the whole network, although we parallelized and implemented this algorithm, and after running

some experiments, we observed that it does not influence the scaling parameter. However, the estimation of this algorithm can be improved by computing $N_B(l_B)$ many times for a given l_B, and then select the maximum of all computation.

2.1.1 Alternatives to box-covering algorithm

Other novel approaches have been proposed to approximate fractal dimension. For example Daijun W. et al [68] used the information dimension introduced by Rényi [69] to capture the fractal property of complex networks, they applied their method to both large and small real-world networks, and in some cases this new measurement gives significantly better results than the traditional box-covering algorithm, because the obtained datapoints are less noisy, and the bias of the fitted distribution is smaller.

Another promising method, proposed by Haixin Z. et al. [70], uses fuzzy set theory to approximate the d_B fractal dimension, furthermore, the complexity of the algorithm is reduced significantly, i.e. it is efficient and less time consuming than the original CBB algorithm. The main idea behind the fuzzy method, is that for a given box, the membership to the box of an arbitrary node is not a true or false (i.e. member or not member variable), but a real number between 0 and 1 which depends on the size of the box and the graph distance between the chosen node and the center of the box. Then an l_B-sized box's covering capability is the sum of these fuzzy numbers, which is proportional to the presumable number of nodes inside an l_B-sized box. More precisely, notice that the Equtation (7) can be rewritten as:

$$d_B \sim \frac{\log\left(N_B(l_B)^{-1}\right)}{\log\left(l_B\right)}, \tag{8}$$

where $N_B(l_B)^{-1}$ is the reciprocal of $N_B(l_B)$, and the clever idea of the authors of [70] was that it can be viewed as the covering ability (CA) of the boxes, i.e. the expected ratio of the size of the network that an l_B-sized box can cover. That is to say, the less boxes are needed to cover the network, the more percentage of nodes of the network can be covered by a box, and similarly the more boxes are needed, the less amount of nodes could be covered by a single box. Thus the ultimate goal of this process is to identify the covering ability of the boxes, and it is calculated as follows:

Identical balls of radius l_B are constructed around every v_i, $1 \leq i \leq N$ vertex of the graph $G = (V, E)$. Then the covering ability of such a ball around the node v is given by:

$$N_v(l_B)^{-1} = \frac{1}{N-1} \sum_{\substack{v_i \neq v \\ d(v_i,v) \leq l_B}} \exp\left(-\frac{d(v_i,v)^2}{l_B^2}\right) = \tag{9}$$

$$= \frac{1}{N-1} \sum_{v_i \neq v} \delta_{v_i,v}(l_B) A_{v_i,v}(l_B), \tag{10}$$

where

$$\delta_{v_i,v}(l_B) = \begin{cases} 1, & d(v_i,v) \leq l_B \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

is the selecting function, that represents whether the node v_i could be covered by the v centered ball, and

$$A_{v_i,v}(l_B) = \exp\left(-\frac{d(v_i,v)^2}{l_B^2}\right) \tag{12}$$

is the fuzzy membership function, with value ranges from 0 to 1, motivated by the fact that in real life situations boundaries between sets or classes are often overlapping or blurred.

Thus the "expected" covering ability of a ball of radius l_B is calculated by the following equation:

$$N_B(l_B)^{-1} = \frac{1}{N} \sum_{v_j \in V} N_{v_j}(l_B)^{-1} = \tag{13}$$

$$= \frac{1}{N(N-1)} \sum_{\substack{v_i,v_j \in V \\ i \neq j}} \delta_{v_i,v_j}(l_B) A_{v_i,v_j}(l_B) \tag{14}$$

The following script is our implementation of this algorithm. Note that, in order to gain faster running time, we modified their algorithm at some places. For example, in their algorithm, the shortest path between node v_i and v_j are calculated multiple times, but here we start our algorithm with calculating the matrix of

graph distances, and later in the cycles, we just have to read the values out from this matrix. Instead of the nested `for` cycles, we used the Wolfram Language's `ParallelTable` and `Table` function, owing to the fact that the outer cycle is parallelizable. Further improvements are obtained by allowing the distance $d(v_i, d_j)$ to be 0 in the sum in Eq. (9), and that is why we subtract 1 from the total. Moreover Wolfram Mathematica calculates everything symbolically by default, and the `N` function gives the numerical value of the expression in its argument, which also accelerates the algorithm, and without this simple function the running time of the algorithm would scale exponentially with the diameter of the input graph.

```
fuzzy[G_] := Module[{distanceMatrix, Nn, invNb, L},
  distanceMatrix = GraphDistanceMatrix[G];
  Nn = VertexCount[G];
  L = Ceiling[GraphDiameter[G]/2];
  invNb =
   ParallelTable[
    Total[Table[
      N[Total[Exp[-(Select[distanceMatrix[[i]], # <= lb &])^2 /
            lb^2 ]] - 1], {i, 1, Nn}]], {lb, 1, L}];
  invNb/(Nn*(Nn - 1))]
```

Listing 1: Our parallelized implementation of the fuzzy algorithm from [70].

In Section 2.3 we compare these algorithms both by their running time and the by accuracy of the obtained d_B fractal dimension, and the results are detailed in Tables 1 and 2.

2.2 Fitting power-laws in empirical data

The problem of identification of fractal networks in practice, relies on the correct detection of power law distribution in the $N_B(l_B)$ empirical data. Unfortunately, the characterization of power laws is complicated, since the tail of the distribution is usually unreliable due to the large fluctuations, furthermore the identification of the range, where the power law relation holds is difficult [10]. Another serious problem is that scientists often leave out of consideration the fact that power laws can be easily confused with other distributions. In several influential articles the validation of power law distributions is carried out visually by comparing log-log plots, or calculating errors of least-square fitting, which are woefully inadequate methods, that produce inaccurate estimates of fractal dimension d_B. A. Clauset,

M.E.J. Newman et al. presented in [10] a statistical framework, that combines maximum-likelihood fitting and likelihood ratios, for discerning and quantifying power law behaviour in empirical data. Their software implementation is also available online[8] in different programming languages.

Generally, power-law distributions are of two basic forms: continuous and discrete distribution. Let x denote the quantity, whose distribution we are interested in. With a slight abuse of notation, the continuous power law distribution can be described by $f(x)$ probability density function as follows [10]:

$$f(x)\,\mathrm{d}x = \mathbb{P}(x \leq X \leq x + \mathrm{d}x) = Cx^{-\alpha}\,\mathrm{d}x, \tag{15}$$

and in the discrete case:

$$p(x) = \mathbb{P}(X = x) = Cx^{-\alpha}, \tag{16}$$

where C is a normalization constant and X is the observed value. Since, these densities diverge as $x \to 0$, thus the power-law relation cannot hold for small values of x. As a solution Newman et al. introduced the $0 < x_{\min}$ lower bound to the power law behaviour, i.e. for which every $x \geq x_{\min}$ the equation (15) or (16) holds. Then, provided $\alpha > 1$ and solving

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = \int_{x_{\min}}^{\infty} Cx^{-\alpha}\,\mathrm{d}x = 1 \tag{17}$$

for C, we obtain that

$$f(x|\alpha) = f(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha}, \tag{18}$$

i.e. X follows Pareto distribution. Similarly, in the discrete case, after calculating the normalizing constant we find that

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}, \tag{19}$$

[8]See `http://tuvalu.santafe.edu/~aaronc/powerlaws/`.

where

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} \frac{1}{(n + x_{\min})^{\alpha}} \tag{20}$$

is the Hurwitz zeta function, i.e. X follows Zipf's law or Zipfian distribution.

Unfortunately, the estimation of the scaling parameter α is often done by performing a least-squares linear regression on the logarithm of the data points, and then extracting the slope of the gained line. It can not be overemphasized that this procedure leads to significant errors, even under relatively common conditions [10]. We will see, that estimating α correctly requires the value of the lower bound $x_{\min}$ of the power-law behaviour of the data. For now, let us assume, that the value of $x_{\min}$ is known. The method of maximum likelihood provably gives an accurate estimation of the scaling parameter [71]. Given that our data are drawn from the continuous distribution described in (18), the maximum likelihood estimation of α can be easily calculated as follows: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote the data vector containing the n observation for which $x_i \geq x_{\min}$. The likelihood function $\mathcal{L}$ of the data $\mathbf{x}$ is given by

$$\mathcal{L}(\alpha, \mathbf{x}) = \prod_{i=1}^{n} f(x_i|\alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \tag{21}$$

The goal of the maximum likelihood parameter estimation is to maximize $\mathcal{L}$ function in α, since the data are the most likely to have been generated by the $\hat{\alpha}$ that maximizes $\mathcal{L}$, i.e. $\hat{\alpha} = \arg\max_{\alpha>1} \mathcal{L}(\alpha, \mathbf{x})$ if a maximum exists. Since it is more convenient to work with sums instead of products, commonly we maximize the logarithm of the likelihood function called log-likelihood function, which has its maximum in the same place, denoted by ℓ. Thus

$$\begin{aligned} \ell(\alpha, \mathbf{x}) = \ln \mathcal{L}(\alpha, \mathbf{x}) &= \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} = \\ &= \sum_{i=1}^{n} \ln \left(\frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \right) = \\ &= n \ln(\alpha - 1) - n \ln(x_{\min}) - \alpha \sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{\min}} \right). \end{aligned} \tag{22}$$

Now, we can easily obtain the maximum likelihood estimate (MLE) for the scaling parameter α by solving $\frac{\partial \ell}{\partial \alpha} = 0$ for α:

$$\frac{\partial \ell(\alpha, \mathbf{x})}{\partial \alpha} = n\frac{1}{\alpha - 1} - \sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right) = 0 \tag{23}$$

From the second equation of (23) we have

$$\frac{1}{\alpha - 1} = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right) \tag{24}$$

$$\alpha - 1 = \left(\frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right)\right)^{-1} \tag{25}$$

$$\hat{\alpha} = 1 + n\left(\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right)\right)^{-1}. \tag{26}$$

Note that, this estimator is asymptotically normal [72] and consistent [73], since the obtained formula (26) is equivalent to the Hill estimator [74] for which these properties are proven.

The MLE for the discrete case is not as straightforward as in the continuous case. Following a similar argument to the continuous variable case the log-likelihood function is as follows:

$$\ell(\alpha, \mathbf{x}) = \ln \prod_{i=1}^{n} \frac{x_i^{-\alpha}}{\zeta(\alpha, x_{\min})} = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i \tag{27}$$

Again, we can obtain the ML estimate $\hat{\alpha}$, by solving $\frac{\partial \ell}{\partial \alpha} = 0$ for α:

$$\frac{-n}{\zeta(\alpha, x_{\min})}\frac{\partial}{\partial \alpha}\zeta(\alpha, x_{\min}) - \sum_{i=1}^{n} \ln x_i = 0 \tag{28}$$

Hence, one can find $\hat{\alpha}$ as a solution of:

$$\frac{\zeta'_{\hat{\alpha}}(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i, \tag{29}$$

where the prime denotes differentiation with respect to the first argument i.e. $\zeta'_\alpha = \frac{\partial \zeta}{\partial \alpha}$. Despite the fact that an exact closed-form solution to $\hat{\alpha}$ does not exist, in [10] they have proposed an approximating solution, by treating the sample from discrete type power law distribution as if they were drawn from the continuous one and then rounded to the closest integer. The details of the derivation are given in [10], the result is below:

$$\hat{\alpha} \approx 1 + n \left(\sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{\min} - \frac{1}{2}} \right) \right)^{-1} . \tag{30}$$

Up to now, we supposed that lower bound parameter $x_{\min}$ is known and notice that the estimation of the scaling parameter α is valid only if $x_{\min}$ is accurate. Thus if we want an accurate estimation of α first we need an accurate estimation of $x_{\min}$. For example, if we choose the value of $x_{\min}$ too small, then we will get a biased estimation of α, due to the fact that we are trying to fit power law distribution to presumably non-power-law data. On the other hand, if the value of $x_{\min}$ is too large, then we are probably throwing away valid data points, which increases both the bias because of the smaller sample size and the statistical error on the scaling parameter α. Figure 2 well-illustrates the problem of choosing too low or too large lower-bound parameter.

A possible approach for estimating $x_{\min}$, proposed by Clauset et al. [75], uses a simple idea: we choose the value of $\hat{x}_{\min}$, that makes the empirical distribution of the observed data and the best-fit power law model as similar as possible above $\hat{x}_{\min}$, i.e. for example which minimizes the Kolmogorov–Smirnov (KS) statistic [76], i.e. the maximum distance between the two cumulative distribution functions (CDF), but there are numerous measures for quantifying the distance between probability distributions. Note that this technique will not give too large or too low value for $\hat{x}_{\min}$, since if $\hat{x}_{\min}$ is higher than the true value $x_{\min}$, then reduced size of the dataset results imperfect empirical distribution and thus poorer match because of statistical fluctuation. Conversely, if $\hat{x}_{\min}$ is smaller than the true value $x_{\min}$, then the empirical and the fitted distribution will be fundamentally different from each other. The method of the estimation of $\hat{x}_{\min}$ is as follows:

Using the previous notations, let $\mathbf{x} = (x_1, \ldots, x_n)$ be the measured data, fur-

thermore, let $S_{x_{\min}}(x)$ denote the empirical CDF for the observations with value at least $x_{\min}$, and $F_{x_{\min}}(x)$ is the CDF of the power-law model that best fits the data in the region $x \geq x_{\min}$, i.e.

$$S_{x_{\min}}(x) = \frac{1}{n_{x_{\min}}} \sum_{i=1}^{n} \mathbb{1}[x_i \leq x], \tag{31}$$

moreover

$$F_{x_{\min}}(x) = \int_{x_{\min}}^{x} \frac{\hat{\alpha}-1}{x_{\min}} \left(\frac{y}{x_{\min}}\right)^{-\hat{\alpha}} \mathrm{d}y = 1 - \left(\frac{x}{x_{\min}}\right)^{1-\hat{\alpha}}, \tag{32}$$

where $n_{x_{\min}}$ normalization constant is the number of data points for which $x_i \geq x_{\min}$, and $\hat{\alpha} > 1$ is the maximum likelihood estimate of α given $x_{\min}$, defined in (26). Now, the Kolmogorov–Smirnov statistic $D_{x_{\min}}$ for $S_{x_{\min}}(x)$ and $F_{x_{\min}}(x)$ is defined as:

$$D_{x_{\min}} = \max_{x \geq x_{\min}} \left| S_{x_{\min}}(x) - F_{x_{\min}}(x) \right|. \tag{33}$$

The estimate $\hat{x}_{\min}$ is then the value of $x_{\min}$ that minimizes $D_{x_{\min}}$, i.e.

$$\hat{x}_{\min} = \arg\min_{x_{\min}} D_{x_{\min}} = \arg\min_{x_{\min}} \max_{x \geq x_{\min}} \left| S_{x_{\min}}(x) - F_{x_{\min}}(x) \right|. \tag{34}$$

Note that commonly, $\hat{x}_{\min}$ is either estimated visually by plotting $\hat{\alpha}$ as a function of $\hat{x}_{\min}$ and choosing a point beyond the value of $\hat{\alpha}$ appears comparatively stable, or beyond which the PDF or CDF of the distribution becomes roughly straight on a log-log scale [10]. For example Figure 2 suggests that the true value of $x_{\min}$ is 50, since beyond this point the function $\hat{\alpha}(x_{\min})$ behaves stably. After the identification of the distribution parameters, one can validate the obtained distribution with goodness-of-fit test, which is a hypothesis test for determining whether the observed data are drawn from the fitted distribution or not. The output of the goodness-of-fit test is a p-value, which quantifies the plausibility of the hypothesis. In our setting the $\mathcal{H}_0$ null hypothesis states that the data follow non-power-law distribution and the alternative $\mathcal{H}_1$ hypothesis is the case when the data are drawn from power-law distribution.

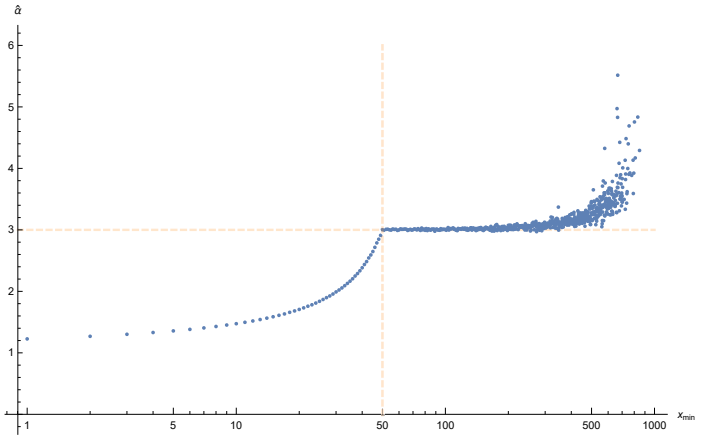


Figure 2: Visual estimation of $x_{\min}$. The points are the mean of the estimated scaling parameter $\hat{\alpha}$ for 100 samples drawn from distribution (32), with parameters $x_{\min} = 50$ and $\alpha = 3$.

2.3 Measurement of fractality

To the best of our knowledge, there is no literature specialized in the power-law fitting of empirical $N_B(l_B)$ data, in spite of the fact, that it is an even more difficult problem, since for a given graph the number of data points is equal to the diameter of the graph. Hence the small-world, specially the ultra-small-world property of real networks results insufficiently small number of data points for reliable hypothesis testing, that is why the commonly used validating technique is the visual comparison. However one could use much larger networks, but that requires powerful computing capability, especially for the box-covering algorithm due to the problem's NP-hard nature (besides impressive computing capacity, huge memory size and maybe even other techniques that are used for handling Big Data are needed). Furthermore, it is easy to see, that when the box size l_B is large, i.e. when it is close to the diameter of the graph, the N_B minimum number of boxes required to cover the network does not follow power law. What is more, in some real networks the presence of fractality is a local property and not global, i.e. power-law holds only for $l_{B,\min} \leq l_B \leq l_{B,\max}$, and $N_B(l_B)$ follows other distribution, typically exponential, for $l_B \geq l_{B,\max}$, thus the size of the relevant data set for power-law fitting is even smaller than the already relatively small sample size. For example this phenomena is well illustrated in the N_B vs l_B plot of Figure 3 and

Figure 4. Furthermore, the goodness-of-fit test is inapplicable, since such small sample size will not result reliable or useful p-values.

Due to the earlier mentioned inconveniences regarding the power-law validation of N_B, in this work we are more interested in the extent of fractality of the graphs, rather than the accurate, exact fractal dimensions of the networks. To this end, the previously mentioned parameter estimation techniques are applicable, since as Song et al. defined the concept of fractality, we distinguish the fractal and non-fractal networks by how fast the decay of the function N_B. However, these parameter estimation techniques may not always return the true fractal dimension of the networks [10], they can measure rate of the decay, thus they can differentiate a fractal and a non-fractal network from each other. In Section 3.1, we will recap the techniques of measuring in more details, when we introduce how we estimated the fractality of real networks.

Owing to the fact that in real-world networks, the pure properties are rare [12], and motivated by the observation of the distribution of N_B of many real-world networks, here we suggest a more precise description or characterization of fractal networks.

In [57] H.D. Rozenfeld et al. suggest that scale-free networks[9] be categorized into three groups:

(i) pure fractal,

(ii) pure small-world,

(iii) mixture between fractal and small-world.

A pure fractal satisfies the fractal scaling equation (6) for all l_B. From Definition 9 a pure small world network satisfies $N \sim e^L$, where N is the size of the network and L is the average distances within the network. This property implies that Eq. (6) never holds, instead $N_B(l_B)$ follows exponential decay with l_B, i.e. $N_B(l_B) \sim e^{-d_e l_B}$. Such networks are also called transfractal, moreover there is a Hierarchical graph-sequence model, introduced by K. Simon and J. Komjáthy [77],

[9]Regarding the fact that the related definitions and concepts are not mathematically precise, this vague categorization of Rozenfeld et al. can be applied on both scale-free and non-scale-free networks as well.

which exponential d_e decay rate (or also referred to as transfractal dimension or modified box dimension) is calculated analytically in [78].

In the case of a mixture between fractal and small-world the fractal scaling Eq. (6) is only satisfied up to a $l_{B,\max}$ cut-off value of l_B, beyond which the fractality breaks down and the small-world property emerges [57] i.e. the small-world property appears in the plot of $N_B(l_B)$ as an exponential cut-off for large l_B, but N_B follows power-law when $l_B \ll D$, that is why they call it locally fractal but globally small world. Thus the key component to appropriately identify a fractal or somewhat fractal network is the exponentially truncated power-law or power-law with exponential cut-off.

Hence, our suggestion is, that the identification of the networks in terms of fractal and small-world property should not only be done by determining the d_B fractal dimension parameter, but besides d_B other two parameters should be involved. Following the idea of categorization of Rozenfeld et al. we suggest to expand the concept of fractal network with a three-parameter-identification technique as follows:
Let us consider a G graph with diameter D, such that the $N_B(l_B)$ is of the form:

$$N_b(l_B) \sim \begin{cases} l_B^{-d_B}, & \text{for } 1 \leq l_B \leq l_{B,\max} \\ \exp(-l_B \cdot d_e) & \text{for } l_{B,\max} < l_B \leq D, \end{cases} \tag{35}$$

then, if G is a mixture between fractal and small-world, then G can be described by the triplet $(d_B, \frac{l_{B,\max}}{D}, d_e)$, where $1 < d_B < \infty$ is the fractal dimension, $1 < l_{B,\max} < D$ is the cut-off value of fractality, and $d_e \neq 0$ is the exponential decay rate. On the pure endpoints of this spectrum if G is pure fractal, then $l_{B,\max} = D$ and $d_e = 0$, on the other hand if G is pure small-world then $l_{B,\max} = 1$ and $d_B = \infty$. Note that the formation and transition of a mixture graph can be easily understand with our simple model detailed later in Section 2.4.6.

2.3.1 Evaluating covering algorithms

As we have already mentioned in Section 2.1.1, we parallelized and implemented the fuzzy algorithm and while it is indeed much faster than the already mentioned CBB algorithm and the obtained data is noiseless, the authors of [70] proposed

linear regression to the logarithm of the data points to calculate the fractal dimension, but when we tested this approach on a two-dimensional grid graph we did not get significantly more accurate results than by the traditional method.

The previously detailed ML estimation of scaling parameter, and the Wolfram Mathematica's built-in parameter estimator function gives slightly better results in case of N_B datapoints obtained by the fuzzy algorithm, although with the optimal $x_{\min}$ parameter the traditional method outperforms the fuzzy method, furthermore the latter one turned out to be extremely sensitive to the $x_{\min}$ parameter, which is because for small values of l_B it behaves "normally" but when the l_B is large, i.e. when the covering capability of a box is close to the size of the network, the $N_B(l_B)$ starts to bow down, thus in this case usage of $x_{\max}$ is more appropriate instead of $x_{\min}$, thus the performance of the linear regression can be slightly improved by taking out of consideration the tail of N_B. The concrete results of the different estimators are detailed in Table 1. Thus, this fuzzy method can only be used to quantify the extent of the fractality of a network, without obtaining the exact fractal dimension, but the speed of the algorithm is impressively faster than the CBB algorithm. Table 2 compares the different algorithms with each other in running time.

Note that, if we only consider the first three values of the estimated $\left(N_B\left(l_B\right)\right)^{-1}$ values, then the fuzzy algorithm with the linear regression gives $d_k = 1.98$, that is a better estimation of the theoretical value of the dimension than the one that we obtain by using more datapoints, which is because as the size of the boxes increases, the size of the overlapping areas increases as well. Hence it gives more accurate results if these fuzzy sets are not overlapping. On the other hand, as we have already mentioned, the datapoints of the fuzzy algorithm are noiseless, thus the most accurate results with this method can be achieved, if we only consider the first few points. This is also reduces the running time of the algorithm.

2.4 Fractal networks and models

In the fundamental work [16], Song et al. showed that several real-world networks from different domains (social, biological, technological) have fractal structure, such as a part of the WWW composed of 325,729 web pages, that are connected

Table 1: The performance of the different parameter estimators. The table contains the estimated d_B of a two-dimensional grid graph with 50×50 vertices, i.e. the true value of d_B equals to 2.

Method	Linear Regression	Wolfram Math.'s builtin estimator	MLE	MLE (optimal $x_{\min}$)
CBB	1.62	1.58	1.43	2
Fuzzy	1.69	1.69	1.5	1.84

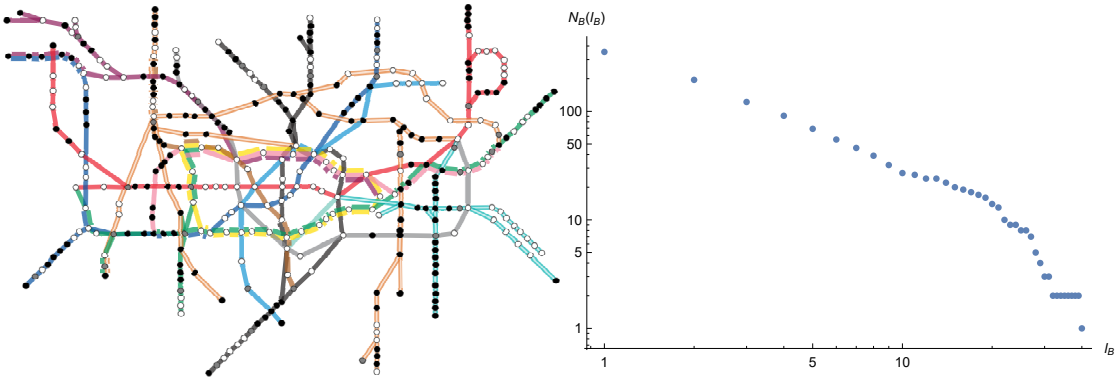
Table 2: Performance of the different box covering algorithms. The table contains the number of seconds in real time that have elapsed during the computation. The columns correspond to different graphs on which the algorithms were run.

Algorithm	Number of nodes and diameter		
	$N = 2500$ $D = 98$	$N = 3126$ $D = 11$	$N = 626$ $D = 161$
Original CBB	30.6	8.77	5.17
Simplified CBB	10.3	1.89	0.98
Parallelized Fuzzy algorithm	5.6	1.57	0.61
Parallelized Simplified CBB	0.21	0.39	0.07

if there exists a URL link from one site to another, a collaboration network of 392,340 actors, where actors linked if they were cast together in at least one movie, and even the network of protein–protein interaction actions found in *Escherichia coli* (also known as *E. coli*) and *Homo sapiens*, where two proteins are linked, if there is a physical binding between them. In more recent works C. Yuan et al. [63] extensively investigated and showed the fractal behaviour of wireless cellular networks (2G and 3G) of two cities, where the nodes are base stations, that are connected if the Pearson correlation of the measured traffic (during a few days) is greater than a circumspectly chosen threshold. Furthermore, S Deng et al. [64] measured the fractal dimension of numerous metro network of large cities. We also calculated the N_B of the metro network of London, the results are shown in Figure 3.

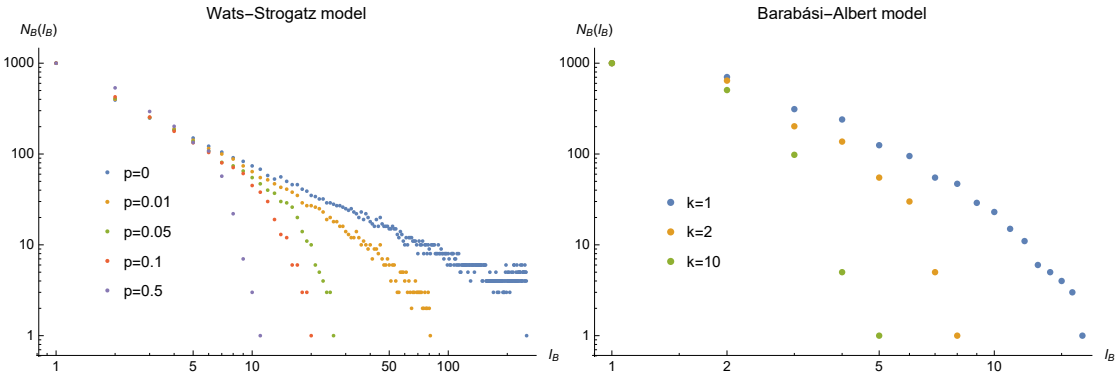
Despite the fact that several real-world networks show fractal behaviour, the aforementioned common scale-free network models (see Section 1) fail to exhibit fractal scaling, or rather it is not typical since only the nearly deterministic Watts–

[10] `http://reference.wolfram.com/language/example/LondonUnderground.html`



(a) (b)

Figure 3: Fractality of London's metro network. The subfigure 3a is the graph of the metro map of London. The figure 3b shows the log-log plot of N_B vs l_B, calculated on the metro graph. The graph is from the Wolfram Data Repository[10].



(a) (b)

Figure 4: The fractal, mixture and non-fractal states of the well-known Watts–Strogatz and Barabási–Albert models. The subfigure 4a WS model's transition from pure fractal to pure small-world as the p rewiring parameter increases, and the subfigure 4b suggests that the BA model with $k = 1$ parameter generates a mixture graph, but clearly the small-world property dominates.

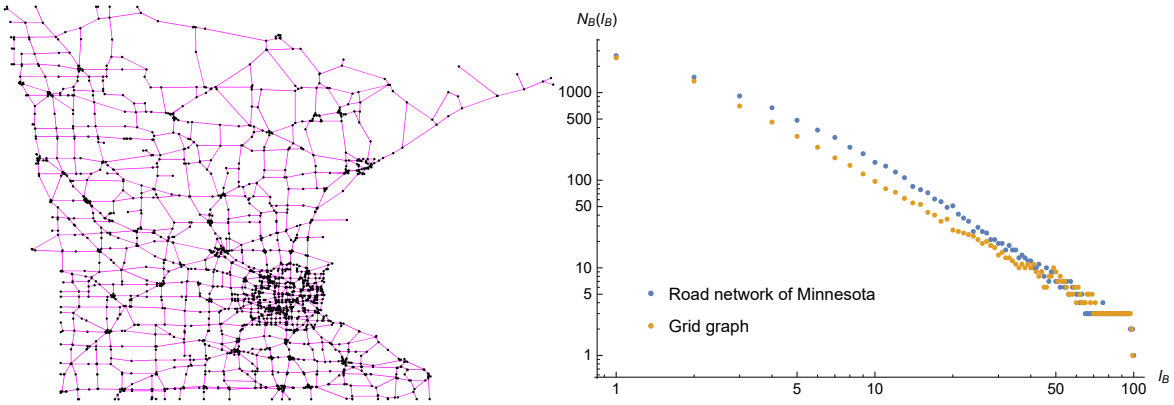
Strogatz (WS) model [2] (more precisely when the edge rewiring probability parameter of the model is equal to or extremely close to zero) generates a trivial fractal graph, and the Barabási–Albert (BA) model with $k = 1$ new edges added in each step, generates a rather mixture fractal tree. The fractality of WS and the BA model is illustrated in Figure 4, furthermore, it also shows that the proposed three-parameter-identification is indeed relevant. By trivial fractal graph we mean those graphs, for which the fractal scaling relation (6) trivially holds, for example it is easy to see that a simple path graph, cycle graph and n-dimensional grid or lattice or gridlike graphs are trivial fractal graphs and their fractal dimension d_B equals to the dimension of the Euclidean space in which these graphs can be embedded.

Typical examples for gridlike real-world networks are the infrastructure networks such as road, metro, water supply, electrical grid and wireless cellular networks of large cities [79, 64, 63] and the 3D structure model of blood vessels and trabecular bones [80]. For example Fig. 5 shows the fractality of the road network of Minnesota, where the subfigure 5a represents the graph of the road network, and the subfigure 5b is the log-log plot of the N_B vs l_B of this graph and of a grid graph. Similarly, Fig. 3 shows the fractality of the metro network of London. The Fig. 5b suggests that the road network's fractal dimension approximately equals to the fractal dimension of the planar grid, which is equal to two. Furthermore, this log-log plot also well-illustrates main problems of empirical power law distributions, such as the large fluctuation of the tail distribution, and the misbehaviour of the distribution for small l_B.

In the following subsections we investigate several mathematical network models and their fractality. First we investigate the well-known Barabási–Albert and Watts–Strogatz models, which fractality, to the best of our knowledge, have never been investigated. Then we detail models, which were directly introduced in order to mimic fractal networks, and to understand the origins of fractality.

2.4.1 Watts–Strogatz model

Tha Watts–Strogatz model (WS), proposed by Duncan J. Watts and Steven Strogatz [2], was motivated by the small-world and highly clustered property of the



(a) (b)

Figure 5: Fractality of Minnesota's road network. The subfigure 5a represents the graph of the road network of Minnesota, USA. The subfigure 5b is the log-log plot of N_B vs. l_B of the road network of Minnesota (blue), and of a two-dimensional grid graph of nearly the same size (orange). The graph is from [37].

real-world networks. The algorithm of the model is as follows:

1. **Initialization**: We start with a regular lattice ring (also called as circulant graph) of N nodes, i.e. a cycle, where every node is connected with its $2K$ nearest neighbours. Formally, if the nodes are labelled $v_1, v_2, \ldots, v_N$, then the there is a link between v_i and v_j if and only if

$$|i - j| \mod (N - K) \leq K.$$

2. **Rewiring the edges**: Each edge is rewired identically with probability p by changing one of the endpoints of the edge, making sure that no self-loop or multiple edge is created. Formally for every $1 \leq i \leq N$, every (v_i, v_j) edge is replaced by (v_i, v_k), with probability p, such that $k \neq i$ and $k \neq j$, and k is chosen uniformly from the set of allowed values.

The WS model illustrates that the pure fractality and pure small-world properties cannot be present simultaneously, but for small p values the model generates mixtures between small-world and fractal.

2.4.2 Barabási–Albert model

The Barabási–Albert (BA) model, introduced by Albert-László Barabási and Réka Albert [22], was inspired by the scale-free property of real networks. The novel concept of the model is the growth and the preferential attachment mechanism. Growth means, that in contrast to the Erdős–Rényi and Watts–Strogatz random graphs, the number of nodes in the BA network increases over time. The preferential attachment mechanism (also referred to as "the rich get richer" or "Yule process") means that the newcomer nodes are more likely to connect to nodes with higher degree, i.e. the more connected a node is, the more likely it receives new links. This phenomena is well discernible in social networks, where a newcomer to a community is more likely to be acquainted with one of the more "visible" or socially active persons. However, originally the idea of the preferential attachment was motivated by the network of the World Wide Web, i.e. the authors of [22] assumed that pages connects preferentially to well-known sites, rather than pages that barely anyone knows. The algorithm of the model is as follows:

1. **Initial condition**: The model starts with a small network of m_0 nodes.

2. **Growth**: At each iteration step, a newcomer node v is connected to $u_1, \ldots, u_m$, $m \leq m_0$ existing nodes, with probability that is proportional to the degree of the u_i nodes, i.e. the p_i probability, that v is connected to the node u_i is

$$p_i = \frac{\deg(u_i)}{\sum_j \deg(v_j)},$$

   where the sum is made over all already existing v_j nodes, which is eventually twice the current number of edges of the network.

Note that this definition of the Barabási–Albert model is rather heuristic and mathematically non-rigorous, for example at $t = 0$ there are no degrees, hence the probabilities are ill-defined, but Bollobás et al. [81] with the help of graph sequences, introduced a mathematically precise version of the model.

This model has been thoroughly investigated, and while for $m > 2$ it generates non-fractal networks which is consistent with the fact that the WWW is non-fractal as well, it is unclear whether for $m \leq 2$ the generated graphs are mixtures

of fractal and non-fractal.

In the following theorems first we will investigate the BA model's the degree distribution and its maximal degree heuristically, then we will give a rigorous proof of the scaling of the maximal degree by [82].

Theorem 6. *[83] The Barabási–Albert model generates scale-free graphs, and the* γ *exponent in the degree distribution equals to* 3*, i.e.* $P(k) \sim k^{-3}$.

Proof. Let us label the nodes of the network by their arrival time, i.e. vertex v_i arrived at time i, furthermore let $d_i(t)$ be the degree of the node v_i at time t. Since the nodes are added one at a time, and the newcomer nodes gain m neighbours, at time t the number of nodes and the number of edges are $t+m_0$ and mt respectively. When a new vertex is added to the network at time t, the probability that it is connected to the old node v_i ($i < t$) is m times the degree of v_i divided by the sum of the degrees i.e

$$\mathbb{P}(\text{At time } t \text{ the newcomer links to } v_i) = m\frac{d_i(t)}{2 \cdot m \cdot t} = \frac{d_i(t)}{2t} \tag{36}$$

Heuristically, if we assume that t and $d_i(t)$ is continuous, then the probability in (36) can be interpreted as the rate of change of the degree of v_i in time [83], i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}d_i(t) = \frac{d_i(t)}{2t}. \tag{37}$$

By solving the simple differential equation in (37), we obtain:

$$d_i(t) = c \cdot (2t)^{\frac{1}{2}} \tag{38}$$

furthermore, we know that $d_i(i)$, the degree of v_i at time i is equal to m, hence the boundary condition to (37) is $d_i(i) = m$, thus $c = m(2i)^{-\frac{1}{2}}$. Substituting it in Eq. (38) we obtain:

$$d_i(t) = m\left(\frac{t}{i}\right)^{\frac{1}{2}} \mathbb{1}[t \geq i]. \tag{39}$$

This can be used in order to calculate γ analytically. Given that $0 < i \leq t$, the

cumulative distribution of $d_i(t)$ is

$$\mathbb{P}\left(d_i\left(t\right) \leq k\right) = \mathbb{P}\left(m\left(\frac{t}{i}\right)^{\frac{1}{2}} \leq k\right) = \mathbb{P}\left(\frac{t}{i} \leq \left(\frac{k}{m}\right)^2\right) \tag{40}$$

$$= \mathbb{P}\left(i > \frac{t \cdot m^2}{k^2}\right) \tag{41}$$

Notice that at time t the arrival time of a node v_i is distributed uniformly on the interval $[0, m_0 + t]$, i.e. it has uniform distribution with density $\frac{1}{m_0+t}$. Thus, substituting this into Eq. (41), we conclude that

$$\mathbb{P}\left(i > \frac{t \cdot m^2}{k^2}\right) = 1 - \mathbb{P}\left(i \leq \frac{t \cdot m^2}{k^2}\right) = 1 - \frac{t \cdot m^2}{k^2} \cdot \frac{1}{m_0 + t} \tag{42}$$

Hence, the probability density function can be obtained by

$$\mathbb{P}\left(k\right) = \frac{\mathrm{d}}{\mathrm{d}k}\mathbb{P}\left(d_i\left(t\right) \leq k\right) = 2\frac{m^2 t}{m_0 + t} \cdot \frac{1}{k^3}, \tag{43}$$

since the node v_i was arbitrary, thus we have that $\mathbb{P}\left(k\right) \sim k^{-3}$ □

Note that this proof is heavily relies on the heuristic argument in Eq. 37, but with the help of the famous Azuma–Hoefding inequality Bollobás et al. showed in [81] how to calculate the scaling parameter γ rigorously on the precisely defined BA model.

In order to prove the scaling of the maximum degree precisely, we need to consider a precise modification of the BA model: For the sake of simplicity, let us start from two nodes connected by an edge. Then at every step a new vertex is added to the graph, and it is connected to the old nodes, with probabilities proportional to the degree of the other vertices, and independently of each other, i.e. the number of new edges in a step is not a fixed parameter but a random variable.

Similarly to the proof of Theorem 6, let us number the vertices according to the order of their creation, hence the vertex set of the model after n iteration is $\{0, 1, \ldots, n\}$. Let $X_{n,k}$ denote the number of vertices of degree k and $Y_{n,k}$ be the

number of vertices of degree at least k, after n steps. Since, after n steps we have $n+1$ nodes, we have that $X_{n,0} + X_{n,1} + X_{n,2} + \ldots = n+1$. Notice that $X_{n,k}$ and $Y_{n,k}$ are connected through the S_n sum of degrees as follows:

$$S_n = \sum_{k\geq 1} kX_{n,k} = \sum_{k\geq 1} Y_{n,k}, \tag{44}$$

since in both summations we counted the nodes of degree k exactly k times. At the nth step the probability that an old vertex of degree k is connected to the newcomer node is defined as $\lambda k/S_{n-1}$, where the proportionality coefficient λ is less than 2.

Furthermore, let $\mathcal{F}_n$ denote the σ-field, generated by the first n steps of the model, moreover let $\Delta_{n,k}$ be the number of new edges into the set of old vertices of degree k at the iteration step n. With $\Delta_{n,k}$ we can formulate the total number of new edges at time n as:

$$\Delta_n = \sum_{k\geq 1} \Delta_{n,k}. \tag{45}$$

Notice, that $\Delta_{n+1,k}$ conditioned to $\mathcal{F}_n$ has binomial distribution with parameters $X_{n,k}$ and $\lambda\frac{k}{S_n}$, since during the iteration step of $n+1$, an edge is drawn to an old vertex of degree k according to a Bernoulli distributed random variable, with probability parameter $\lambda\frac{k}{S_n}$. Furthermore, the edge additions are independent from each other, thus the $\Delta_{n+1,k}$ new edges to the set of old k-degree nodes consist of $X_{n,k}$ trials of Bernoulli distributed events, which is by definition has Binomial distribution with parameters $X_{n,k}$ and $\lambda\frac{k}{S_n}$. Hence we have that

$$\mathbb{E}(\Delta_{n+1} \mid \mathcal{F}_n) = \mathbb{E}\left(\sum_{k\geq 1} \Delta_{n+1,k} \,\middle|\, \mathcal{F}_n\right) = \sum_{k\geq 1} \mathbb{E}\big(\Delta_{n+1,k} \,\big|\, \mathcal{F}_n\big) = \tag{46}$$

$$= \sum_{k\geq 1} \mathbb{E}\left(Bin\left(X_{n,k}, \lambda\frac{k}{S_n}\right)\right) = \sum_{k\geq 1} X_{n,k}\lambda\frac{k}{S_n} = \tag{47}$$

$$= \lambda\frac{1}{S_n}\sum_{k\geq 1} kX_{n,k} = \lambda\frac{1}{S_n}S_n = \lambda, \tag{48}$$

i.e. the expected number of new edges at each step is λ.

Before we prove a strong law of large numbers for the maximum degree, we have to understand the asymptotics of S_n.

Theorem 7. *[82]*

$$S_n = 2\lambda n + o\left(n^{\frac{1}{2}+\varepsilon}\right), \quad \forall \varepsilon > 0.$$

Proof. With $\Delta_1 = 1$ let us define $\zeta_n = \sum_{j=1}^{n} \left(\Delta_j - \lambda\right) = \frac{S_n}{2} - n\lambda$, since the sum of degrees agrees with the half of the number of edges. Notice that ζ_n is a martingale with respect to $\mathcal{F}_n$ since

$$\mathbb{E}(\zeta_{n+1} - \zeta_n \mid \mathcal{F}_n) = \mathbb{E}(\Delta_{n+1} - \lambda \mid \mathcal{F}_n) = \lambda - \lambda = 0, \tag{49}$$

what is more, $(\zeta_n, \mathcal{F}_n)$ is a square integrable martingale, since the variance of ζ_n is equal to $\sum_{j=1}^{n} \sum_{k\geq 1} \mathrm{Var}(\Delta_{j,k})$ and then the exact variance can be easily calculated using the law of total variance. By the convexity of the square function, ζ_n^2 is a submartingale, so the process A_n in its Doob–Meyer decomposition is increasing. A_n is also called the predictable quadratic variation of (ζ_n^2) and usually denoted with angle brackets as $\langle \zeta_n^2 \rangle$. By the Doob–Meyer decomposition we have that

$$A_n = \sum_{j=2}^{n} \mathrm{Var}(\Delta_j | \mathcal{F}_{j-1}) = \sum_{j=2}^{n-1} \sum_{k\geq 1} \mathrm{Var}\left(Bin\left(X_{j,k}, \lambda \frac{k}{S_j}\right)\right) = \tag{50}$$

$$= \sum_{j=2}^{n-1} \sum_{k\geq 1} X_{j,k} \frac{k\lambda}{S_j}\left(1 - \frac{k\lambda}{S_j}\right) \leq n\lambda. \tag{51}$$

From [84], we know that $\zeta_n = o\left(A_n^{\frac{1}{2}+\varepsilon}\right)$ almost everywhere on the event $A_n \to \infty$, i.e. we obtained that $\frac{S_n}{2} - n\lambda = o\left(n^{\frac{1}{2}+\varepsilon}\right)$, which completes the proof.□

Now we can move on to the proof of the maximum degree.

Theorem 8. *Let $M_n = \max\{k : X_{n,k} > 0\}$ denote the maximum degree of the (modified) BA model after n iteration steps. Then we have*

$$[82] \lim_{n\to\infty} \frac{M_n}{\sqrt{n}} = \mu \tag{52}$$

almost surely, where the limit μ differs from zero with positive probability.

Proof. Let $W_{n,j}$ denote the degree of vertex j after the nth step (recall that j is its creation time), with initial values $W_{n,j} = 0$ for $n < j$, $W_{1,0} = W_{1,1} = 1$ and $W_{j,j} = \Delta_j$. Then M_n can be rewritten as

$$M_n = \max\{W_{n,j} : j \geq 0\}.$$

(Note that non-rigorous counterpart of $W_{n,j}$ is $d_j(n)$ from Eq. (39).)

Now let us introduce $c_{n,k}$ normalizing terms as follows:

$$c_{n,k} = \prod_{i=1}^{n-1} \frac{S_i}{S_i + k\lambda}, \quad n \geq 1,\ k \geq 1. \tag{53}$$

For $n \to \infty$, with probability 1, we have that

$$c_{n,k} = \exp\left(-k\lambda \sum_{i=1}^{n-1} \frac{1}{S_i} + \frac{k^2\lambda^2}{2} \sum_{i=1}^{n-1} \frac{1 + o(1)}{S_i^2} \right) \tag{54}$$

since

$$\ln(c_{n,k}) = \sum_{i=1}^{n-1} \ln\left(\frac{S_i}{S_i + k\lambda}\right) = \sum_{i=1}^{n-1} \ln\left(\frac{1}{1 + \frac{k\lambda}{S_i}}\right) = \tag{55}$$

$$= -\sum_{i=1}^{n-1} \ln\left(1 + \frac{k\lambda}{S_i}\right) = -\sum_{i=1}^{n-1} \left(\frac{k\lambda}{S_i} - \frac{1}{2}\left(\frac{k\lambda}{S_i}\right)^2 + \dots \right) = \tag{56}$$

$$= -k\lambda \sum_{i=1}^{n-1} \frac{1}{S_i} + \frac{k^2\lambda^2}{2} \sum_{i=1}^{n-1} \frac{1 + o(1)}{S_i^2}, \tag{57}$$

where in Eq. (56) we used the Taylor series expansion of $\ln(1 + x)$ at 0. From Theorem 7 we know that $\frac{1}{S_i} = \frac{1}{2\lambda i}(1 + o(i^{-\frac{1}{2}+\varepsilon}))$, substituting this into Eq. (57), and using the fact that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i} - \ln(n) = \gamma,$$

we obtain that for $n \to \infty$, $\ln(c_{n,k})$ differs from $-\frac{k}{2}\ln(n)$ only by a term converging

with probability 1. Thus, we have that

$$c_{n,k} \sim \gamma_k n^{-\frac{1}{2}}, \tag{58}$$

with an appropriate positive random variable γ_k.

Furthermore, we clearly know that

$$\mathbb{E}\big(W_{n+1,j} \,\big|\, \mathcal{F}_n\big) = W_{n,j} + \lambda \frac{W_{n,j}}{S_n} = W_{n,j} \frac{S_n + \lambda}{S_n}. \tag{59}$$

Let us define $Z^1_{n,j}$ as

$$Z^1_{n,j} = c_{n,1} W_{n,j}.$$

From Eq. (59) for $n \geq \max\{j, 1\}$ we have that $(Z^1_{n,j}, \mathcal{F}_n)$ is either a positive martingale (can be easily checked by definition) or constant zero, hence it converges almost surely to some ζ_j. To estimate the moments of ζ_j, consider

$$Z^k_{n,j} = c_{n,k} \binom{W_{n,j} + k - 1}{k}. \tag{60}$$

Since in a step the degree of the node j is either increases by one or remains the same, $W_{n+1,j} - W_{n,j}$ is equal to either 1 or 0, hence by using the binomial coefficient's recurrence relation, namely that $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$, we can write that

$$\begin{aligned}
\binom{W_{n+1,j} + k - 1}{k} &= && (61)\\
&= \binom{W_{n,j} + k - 1}{k} + \big(W_{n+1,j} - W_{n,j}\big) \binom{W_{n,j} + k - 1}{k-1} = && (62)\\
&= \binom{W_{n,j} + k - 1}{k} \left(1 + \big(W_{n+1,j} - W_{n,j}\big) \frac{k}{W_{n,j}}\right), && (63)
\end{aligned}$$

since

$$\binom{W_{n,j}+k-1}{k-1} = \frac{(W_{n,j}+k-1)\cdot(W_{n,j}+k-2)\cdot\ldots\cdot(W_{n,j}+1)}{(k-1)\cdot(k-2)\cdot\ldots\cdot 1} = \tag{64}$$

$$= \frac{k}{W_{n,j}} \frac{(W_{n,j}+k-1)\cdot\ldots\cdot(W_{n,j}+1)\cdot W_{n,j}}{k\cdot(k-1)\cdot(k-2)\cdot\ldots\cdot 1} = \tag{65}$$

$$= \frac{k}{W_{n,j}} \binom{W_{n,j}+k-1}{k}. \tag{66}$$

By taking conditional expectation of Eq. (63) with respect to $\mathcal{F}_n$ we obtain that

$$\mathbb{E}\left(\binom{W_{n+1,j}+k-1}{k-1} \,\middle|\, \mathcal{F}_n \right) = \tag{67}$$

$$= \binom{W_{n,j}+k-1}{k} \left(1 + \frac{k}{W_{n,j}} \mathbb{P}\left(W_{n+1,j} - W_{n,j} = 1 \right) + 0 \right) = \tag{68}$$

$$= \binom{W_{n,j}+k-1}{k} \left(1 + \frac{k}{W_{n,j}} \lambda \frac{W_{n,j}}{S_n} \right) = \tag{69}$$

$$= \binom{W_{n,j}+k-1}{k} \left(1 + \frac{\lambda k}{S_n} \right), \tag{70}$$

hence, for $n \geq \max\{j, 1\}$,

$$\left(Z_{n,j}^k, \mathcal{F}_n \right)$$

is also a convergent martingale. Notice that since $\left(c_{n,1}\right)^k \leq c_{n,k}$, we can majorize $\left(Z_{n,j}^1\right)^k$ by $k! Z_{n,j}^k$.

Now, $c_{n,1} M_n = \max\{Z_{n,j}^1 : 0 \leq j \leq n\}$, is a maximum of an increasing number of nonnegative martingales, which is a submartingale. The rest of the proof can be completed by showing that this submartingale is bounded in L_k, for some $k \geq 1$.

Let us start from the estimation

$$\mathbb{E}(c_{n,1}M_n)^k = \mathbb{E}\left(\max\left\{\left(Z_{n,j}^1\right)^k : 0 \le j \le n\right\}\right) \le \tag{71}$$

$$\le \mathbb{E}\left(\max\left\{k!Z_{n,j}^k : 0 \le j \le n\right\}\right) \le k!\sum_{j=0}^{n}\mathbb{E}\left(Z_{n,j}^k\right) = \tag{72}$$

$$= k!\mathbb{E}(Z_{1,0}^k) + k!\sum_{j=1}^{n}\mathbb{E}(Z_{j,j}^k) = k! + k!\sum_{j=1}^{n}\mathbb{E}\left(c_{j,k}\binom{W_{j,j}+k-1}{k}\right) = \tag{73}$$

$$= k! + k!\sum_{j=1}^{n}\mathbb{E}\left(c_{j,k}\binom{\Delta_j+k-1}{k}.\right) \tag{74}$$

By the law of total expectation (tower-rule) we have that

$$\mathbb{E}\left(c_{j,k}\binom{\Delta_j+k-1}{k}\right) = \mathbb{E}\left[\mathbb{E}\left(c_{j,k}\binom{\Delta_j+k-1}{k}\,\middle|\,\mathcal{F}_{j-1}\right)\right] = \tag{75}$$

$$= \mathbb{E}\left[c_{j,k}\,\mathbb{E}\left(\binom{\Delta_j+k-1}{k}\,\middle|\,\mathcal{F}_{j-1}\right)\right]. \tag{76}$$

Now, notice, that independently of j,

$$\mathbb{E}\left(\binom{\Delta_j+k-1}{k}\,\middle|\,\mathcal{F}_{j-1}\right) \le \mathbb{E}\left(\binom{\pi+k-1}{k}\right), \tag{77}$$

where π denotes a $Poisson(\lambda)$ random variable. Eq. (77) holds, since by definition $\Delta_j = \sum_{k=0}^{j-1}\Delta_{j,k}$, furthermore, we can write that

$$\binom{\Delta_j+k-1}{k} = \sum_{l_0+\ldots+l_j=k}\binom{\Delta_{j,0}}{l_0}\cdot\ldots\cdot\binom{\Delta_{j,j-1}}{l_{j-1}}\binom{k-1}{l_j}. \tag{78}$$

Notice that the binomial coefficients on the right-hand side are conditionally independent by the fact that we draw the edges independently. We have already seen that the conditional distribution of each $\Delta_{j,i}$ is binomial. Let ξ be a $Binomial(n,p)$, and η a $Poisson(np)$ random variable. Then using the factorial

moments of these distributions we have that

$$\mathbb{E}\left(\binom{\xi}{l}\right) = \mathbb{E}\left(\frac{\xi(\xi-1)\dots(\xi-l+1)}{l!}\right) = \frac{1}{l!}\binom{n}{l}p^l l! = \tag{79}$$

$$= \frac{n(n-1)\dots(n-l-1)}{l!}p^l \leq \frac{(np)^l}{l!} = \frac{\mathbb{E}(\eta)^l}{l!} = \tag{80}$$

$$= \mathbb{E}\left(\frac{\eta(\eta-1)\dots(\eta-l+1)}{l!}\right) = \mathbb{E}\left(\binom{\eta}{l}\right). \tag{81}$$

Hence, if we replace every random variables $\Delta_{i,j}$ on the right-hand side of Eq. (78), by conditionally independent *Poisson* variables, the conditional expectation cannot decrease, hence the statement of Eq. (77) follows.

Remembering that we only have to prove the L_k-boundedness of the submartingale $c_{n,1}M_n$, by the previous arguments and Eq. (71) and (75), it is enough to show that

$$\sum_{j=1}^{\infty}\mathbb{E}\left(c_{j,k}\right) < +\infty. \tag{82}$$

From Eq. (58) it is clear, that if $k > 2$ then $\sum_{j=1}^{\infty} c_{j,k}$ converges, but the integrability does not follow immediately.

Let $k = 8$ and $N = \max\{n : S_n > 4\lambda n\}$. Then for $j > N$, we have

$$c_{j,8} = \prod_{i=1}^{j-1}\left(1 - \frac{8\lambda}{S_i + 8\lambda}\right) \leq \prod_{i=N+1}^{j-1}\left(1 - \frac{8\lambda}{4\lambda + 8\lambda}\right) = \tag{83}$$

$$= \prod_{i=N+1}^{j-1}\left(1 - \frac{2}{n+2}\right) = \frac{(N+1)(N+2)}{j(j+1)}, \tag{84}$$

but notice that this obviously holds even for $j \leq N$. Thus, for the boundedness of $\sum_{j=1}^{\infty}\mathbb{E}(c_{j,8})$ it is sufficient to prove that $\mathbb{E}\left(N^2\right) < +\infty$.

By the usual large deviation arguments, i.e. with the help of the exponential version of the Chebyshev's inequality, we have

$$\mathbb{P}(N = n) \leq \mathbb{P}(S_n > 4\lambda n) = \mathbb{P}\left(2^{S_n/2} > 2^{2\lambda n}\right) \leq \frac{\mathbb{E}\left(2^{S_n/2}\right)}{4^{\lambda n}}. \tag{85}$$

The reason why we divided S_n by 2 is because S_n equals to twice the number of edges, hence $\frac{1}{2}S_n = \sum_{i=1}^{n} \Delta_i$. By looking at the result of Eq. (85), notice that in order to finish the proof, we have to estimate the moment generating function of S_n. With $\Delta_1 = 1$, we can write

$$\mathbb{E}\left(2^{S_n/2}\right) = \mathbb{E}\left(\mathbb{E}\Big(2^{S_n/2} \;\Big|\; \mathcal{F}_{n-1}\Big)\right) = \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}}\,\mathbb{E}\Big(2^{\Delta_n} \;\Big|\; \mathcal{F}_{n-1}\Big)\right) = \tag{86}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}}\,\mathbb{E}\Big(2^{\sum_{j=1}^{n-1}\Delta_{n,j}} \;\Big|\; \mathcal{F}_{n-1}\Big)\right) = \tag{87}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \prod_{j=1}^{n-1} \mathbb{E}\Big(2^{\Delta_{n,j}} \;\Big|\; \mathcal{F}_{n-1}\Big)\right) = \tag{88}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \prod_{j=1}^{n-1} \mathbb{E}\left(e^{Bin\left(X_{n-1,j},\frac{\lambda j}{S_{n-1}}\right)\ln(2)}\right)\right) = \tag{89}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \prod_{j=1}^{n-1} \left(\frac{\lambda j}{S_{n-1}} e^{\ln(2)} + 1 - \frac{\lambda j}{S_{n-1}}\right)^{X_{n-1,j}}\right) = \tag{90}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \prod_{j=1}^{n-1} \left(\frac{\lambda j}{S_{n-1}}\right)^{X_{n-1,j}}\right) \leq \tag{91}$$

$$\leq \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \exp\left(\sum_{j=1}^{n-1} \frac{\lambda j X_{n-1,j}}{S_{n-1}}\right)\right) = \tag{92}$$

$$= \mathbb{E}\left(2^{\frac{1}{2}S_{n-1}} \exp\left(\frac{\lambda}{S_{n-1}} \sum_{j=1}^{n-1} j X_{n-1,j}\right)\right) = \tag{93}$$

$$= e^{\lambda}\mathbb{E}\left(2^{\frac{1}{2}S_{n-1}}\right), \tag{94}$$

where in Eq. (89) we used again the fact that the conditional distribution of $\Delta_{n,j}$ with respect to $\mathcal{F}_{n-1}$ is binomial, furthermore in Eq. (90) we applied the well-known moment generating function of the binomial distribution at $t = \ln(2)$.

Therefore, we obtained that $\mathbb{E}\left(2^{S_n/2}\right) \leq e^{\lambda}\mathbb{E}\left(2^{\frac{1}{2}S_{n-1}}\right) \leq \ldots \leq e^{\lambda n}$, which

combined with Eq (85), implies that

$$\mathbb{P}(N = n) \leq \left(\frac{e}{4}\right)^{\lambda n}.$$

Hence, we obtained that, indeed $\mathbb{E}\left(N^2\right) < +\infty$, which completes the proof. □

2.4.3 Dynamical growth model

The dynamical growth (DG) model also called Song–Havlin–Makse model (SHM), was introduced by Song, Havlin and Makse in [14], which was motivated by the difference of the distribution of the $R(k_1, k_2)$ degree correlation ratio (see Definition 23) between fractal and non-fractal real-world networks. They found that the famous fractal network of *E. coli*, $R(k_1, k_2)$ shows an anti-correlation of the degrees, i.e. the high degree nodes are mostly connected to low degree nodes (disassortativity), i.e. there is a "repulsion" between the hubs of the network. On the other hand they also investigated the network of Internet at the router level, which is known to be non-fractal, in which there is a high correlation between degrees, thus their conjecture was disassortativity between the degree of the nodes leads to fractal, and assortativity between the degrees leads to non-fractal network. Thus their dynamical growth model uses these principles, when a new node connects to the already existing ones, in order to create a fractal, non-fractal or a mixture between fractal and non-fractal [14]. The algorithm of the model is as follows:

1. **Initial condition**: In generation $t = 0$, we start from a simple structure of a few nodes e.g. two nodes connected via a link.

2. **Growth**: At each iteration step, the next generation $t+1$ is obtained recursively connecting $m \cdot \deg_t(v)$ new vertices (offsprings) to the already existing v nodes, where m is a predefined parameter, and $\deg_t(v)$ is the degree of the node v at time t. Let us denote the offsprings of v by $(v^{(i)})_{i=1,\ldots,m\deg_t(v)}$.

3. **Rewiring of the edges**: In addition, we identically rewire every old (u, v) edge with probability p, more precisely as a stochastic combination of **Mode I** (with probability p) and **Mode II** (with probability $(1 - p)$)

i) **Mode I**: We keep the old edge (u, v),

ii) **Mode II**: We delete (u, v), and add $x \leq m$ new $(u^{(i)}, v^{(i)})_{i=1,\ldots,x}$ links connecting pairs of the offsprings of the nodes u and v.

Figure 6 illustrates the growing mechanism of one iteration step of the model with parameters $m = 3$, $p = 1$ and both $x = 1$ and $x = 2$. Note that in the case of $x = 1$, the rewiring can be interpreted as the replacement of the (u, v), between a randomly chosen offspring pair. In this model, the key parameter is the p rewiring parameter, because S. Havlin et al. showed that when $p = 0$ the model generates non-fractal graph, with assortative mixing, on the other hand, when $p = 1$ the model leads to fractal graph, with disassortative structure according to the R degree correlation ratio. Furthermore, there is a continuous transition between the the two deterministic states, where the mixture property emerges. Note, that in our experiments, the model generates disassortative graphs for all values of p according to the ρ assortativity coefficient, and varying the rewiring parameter does not change significantly the assortativity [15].

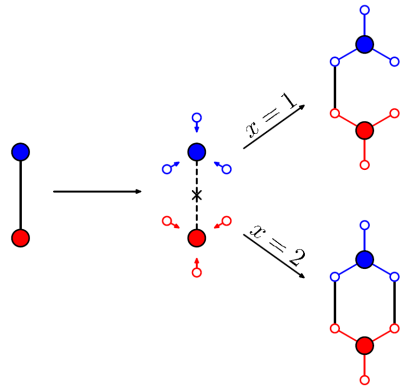


Figure 6: The evolution of the Song–Havlin–Makse model, with parameters $m = 3$, $p = 1$ and $x = 1, 2$. The figure is from [57]

2.4.4 Hub attraction dynamical growth model

We have already mentioned that Song et al. showed that the collaboration network of actors is fractal [16]. On the other hand, due to the social domain of the network, there is a high probability, that high degree actors are connected, i.e. that

they have collaborated in at least one movie, which in the language of network theory means that in this network there is a great chance, that the hubs are connected, which contradicts the aforementioned repulsion-between-hubs principle. Thus, motivated by this observation L. Kuang et al. modified the DG model, such that with the new mechanism the model can generate fractal networks, with strong hub attraction [18]. Hence, the hub attraction dynamical growth (HADG) model is based on the previously described DG model from 2.4.3, with the following modification applied: firstly, the rewiring probability of the model is flexible, i.e. it depends on the degree of the endpoints of the links. The other modification, is what they call within-box link-growth method, which means that after the egde rewiring, the model adds additional edges between the newly added offsprings, in order to increase the clustering coefficient of the network. The within-box link growth method was motivated by the fact that real-world networks are usually highly clustered, especially social networks [4], yet the SHM networks are either trees, or contains only long cycles, i.e. its global clustering coefficient is zero. The evolution of the HADG model is defined as follows [18]:

1. **Initial condition** and **growth**: The start and the growth of the model is exactly the same as in the DG model (see step 1 and step 2).

2. **Flexible edge rewiring**: We rewire the (u, v) edge with probability a if $\frac{\deg_t(u)}{\deg_t^{\max}} > T$ and $\frac{\deg_t(v)}{\deg_t^{\max}} > T$, and rewire it with probability b otherwise. Formally the $p_{(u,v)}$ edge rewiring probability of the (u, v) edge at time $t + 1$ is given by:

$$p_{(u,v)} = \begin{cases} a, & \text{if} \frac{\deg_t(u)}{\deg_t^{\max}} > T \text{ and } \frac{\deg_t(v)}{\deg_t^{\max}} > T \\ b, & \text{otherwise,} \end{cases} \tag{95}$$

   where $\deg_t^{\max}$ is the maximum degree in the network at time t and $a, b, T \in [0, 1]$ are predefined parameters. Thus, if de define $a < b$, then hubs will have higher probability to be connected than non-hubs.

3. **Within-box link-growth**: At step $t + 1$, for each old v nodes, we add $\deg_t(v)$ edges between the newly generated offsprings of v.

Note that the last step does not effect the fractal and scale-free property of the model [18]. The conclusion of the article [18] is that there are fractal networks, with assortative behaviour, i.e. where the most connected nodes can be connected, and with this model, with appropriate parameter settings, we can generate such graphs. Figure 7b shows the results of the box-counting algorithm on this model, with parameters $a = 0.1, 0.65, 1$, the other two parameters were chosen to be fixed values of $b = 0.1$ and $T = 0.4$.

2.4.5 Repulsion based fractal model

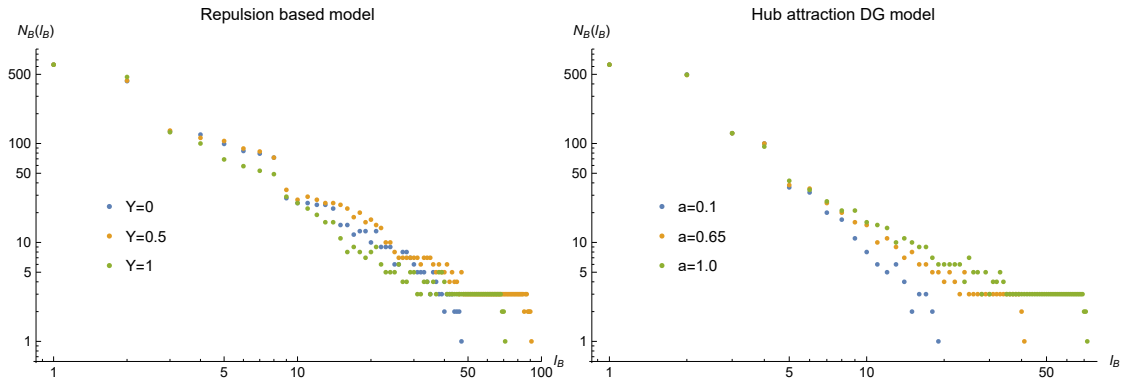
Although Kuang et al. highlighted the existence of fractal networks with hub-connection and showed that the repulsion-between-hubs principle of Song et al. cannot be the exceptional origin of fractality, they did not explained nor investigated that then which other rules give rise to fractal structure of the networks. In my Bachelor thesis[15], we introduced a new model, called repulsion based fractal (RBF) model, which resolves the contradiction between the arguments of the two articles. Here we also attempt to show that the apparent contradiction is due to the famous Third Variable Problem. While, in the SHM model the extent of fractality indeed correlates with the magnitude of disassortativity, we assume that this relationship is not causal, and there is a third variable in the background, which affects both fractality and degree correlation. The growing mechanism of the repulsion based fractal model is as follows:

1. **Initial condition** and **growth**: The start and the growth of the model is exactly the same as in the DG and the HADG model (see step 1 and step 2).

2. **Dynamic edge rewiring**: The edge rewiring probability of the (u, v) edge depends on the average of the degrees of u and v, i.e. the $p^Y_{(u,,v)}$ dynamic edge rewiring probability at time $t + 1$ can be calculated by:

$$p^Y_{(u,,v)} = 1 - \left| Y - \frac{\deg_t(u) + \deg_t(v)}{2 \deg_t^{\max}} \right|, \tag{96}$$

where $0 \leq Y \leq 1$ is a predefined parameter, and $\deg_t^{\max}$ is the maximum degree in the graph at time t. With the Y parameter, we assign high edge rewiring



(a) (b)

Figure 7: The log-log plot of the result of the box-counting on the repulsion based and the hub attraction DG model. Figure 7a shows repulsion based model with different Y parameters, and 7b shows the hub attraction model with different a parameters, while $b = 0.1$ and $T = 0.4$.

probability to those edges, which endpoints' average degree is close to $Y \cdot \deg_t^{\max}$, for example if $Y = 0$, with high probability we rewire those edges, which connect nodes with relative small degrees, on the other hand in case of $Y = 1$, with high probability we rewire the edges, that are linked between nodes with large degrees, i.e. between hubs. The speciality of this model, is that it gives rise to fractal graphs for all $Y \in [0, 1]$. For example Figure 7a shows the distribution of $N_B(l_B)$ on a log-log scale of this model with $Y = 0, 0.5, 1$ parameter settings.

Our conjecture is that the mentioned third variable is the "repulsion", which naturally increases the mean graph distances, and we have seen that the less small-world a graph is the more fractal it is. On the other hand repulsion of course effects the degree correlations, since if the repulsion is between hubs, i.e. in the model $Y = 1$, then hubs are only connected with nodes of small degree, thus the degrees are anti-correlated, while when the repulsion is between the small degree nodes, that means that the hubs are connected, and there are long paths consist of relative small degree nodes, hence there is significantly larger correlation between the degrees.

2.4.6 Mixture model

The following model was motivated by our observation, that fractal real-world networks often show assortative mixing patterns, which is in contrast to the results of [14], namely that fractal networks are often disassortative. As we have already mentioned, the gridlike networks such as metro or road networks (Figure 3 and 5) are fractal, while the degrees of these networks are highly correlated. Our novel model embraces both preferential attachment mechanism and the "geometric" nature structure, that emerges in the infrastructure related networks [79], furthermore in blood vessels and trabecular bones [80], that can be embedded in two- and three-dimensional Euclidean spaces respectively. The model is defined as follows:

1. **Initial condition**: We start with a k-dimensional $G_{n_1,\dots,n_k}$ grid graph with $n_1 \times \ldots \times n_k$ vertices, where k and $n_1 \ldots, n_k$ are predefined parameters.

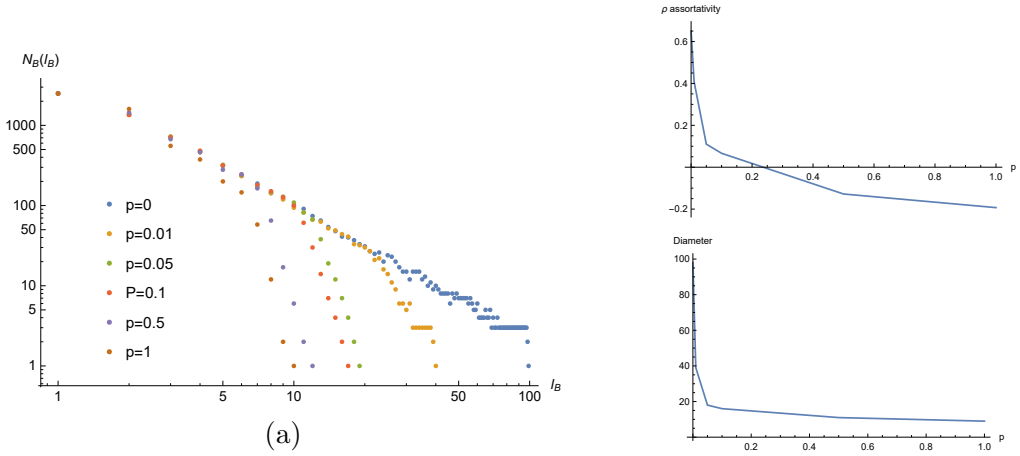
2. **Edge rewiring**: Then similarly to the WS model, with probability p every (v_i, v_j) edge is replaced by (v_i, v_k), but the v_k node is not chosen uniformly, but according to the preferential attachment mechanism. We only rewire those edges, which were originally present in the grid graph, i.e. one edge is only rewired once.

Thus, varying the p parameter from 0 to 1, the model initially generates fractal, non-small-world, assortative graphs, which then transform into non-fractal, small-world, disassortative networks, this phenomena is well-illustrated on Figure 8.

However, if we define the probabilities as in the original preferential attachment model, then this procedure does not lead to scale-free notworks, but if we modify the probabilities, such that we increase the attraction of the nodes that degree is greater than half of the maximal degree, then it results to emergence of larger hubs, and our conjecture is that degree distribution starts to scale as a power-law. To the modification we used the well-known sigmoid function.

Let denote the p_{v_j} probability that the v_k endpoint is replaced to v_j is proportional to

$$p_{v_j} = \frac{1}{1 + \exp\left(-a\left(\frac{\deg(v_j)}{\deg^{\max}} - \frac{1}{2}\right)\right)},$$



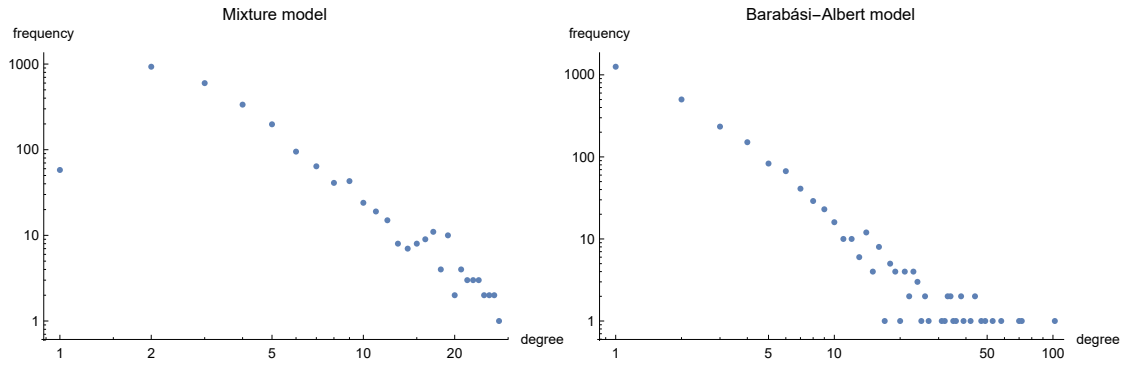
(a) (b)

Figure 8: The transition of fractality, assortativity and diameter of the mixture model. Figure 8a shows the N_B of the model with different p parameters, and Figure 8b shows the change of ρ assortativity and diameter as a function of p.

where a is a positive constant, which defines the sharpness of the "S"-shaped curve, i.e. for $a \to \infty$ this function converges to the Heaviside step function shifted to the right by $\frac{1}{2}$.

Figure 8 shows that the model indeed at $p = 0$ is pure fractal indeed, which transforms into a mixture of fractal and small-world. Note that, while the $l_{B,\max}$ decreases the fractal dimension does not vary, this suggests that the local fractal structure does not change. As we anticipated, Figure 8b shows that as p increases, the assortativity and the diameter of the model decreases drastically, since as we rewire the edges, we create shortcuts, furthermore the assortativity changes, since hubs emerge which are connected with numerous nodes of relative small degree.

The subfigure 9a of Figure 9 shows the log-log plot of the empirical degree distribution of the sigmoid-modified mixture model with rewiring probability of $p = 1$ and as a reference Figure 9b shows the log-log plot of the degree distribution of a similar sized graph generated by the Barabási–Albert model with $m = 2$ parameter setting. The plot suggests that the mixture model becomes scale-free as $p \to 1$, however the analytical proof of this conjecture requires more research of this model.



(a) (b)

Figure 9: The log-log plot of the degree distribution of the modified mixture model and the scale-free Barabási–Albert model.

3 Data-driven analysis of network metrics

In this section we present a data-driven analysis of real-world complex networks. Machine learning techniques provide powerful tools to discover patterns in data, hence it can be applied in network analysis as well in order to effectively find the common and different aspects of real networks from different domains. To this end we collected 584 networks from five different domains (brain, cheminformatics, social, food and miscellaneous), and then we calculated a rich set of graph metrics of these graphs. The chosen graph measurements are detailed in Subsection 3.1. To the best of our knowledge, we have gathered the largest dataset, which besides the basic descriptive properties contains several sophisticated graph metrics of real networks, furthermore this is the only dataset which contains information about the fractality of the networks. Here we set multiple goals with this dataset.

Firstly, we attempt to determine the most influential metrics, by which the domains of the networks can be identified. To this end, we will use two different approach: in Subsection 3.2 we use exploratory data analysis, mainly data visualizations of the different metrics, and study their relations and distributions on the different domains. In Subsection 3.3.2 we will teach different machine learning models to be able to distinguish the different domains by the graph metrics, this task in the language of data science is called classification. Note that there is a recent brief study [85], where similar methods have been used, but their re-

sults are arguable, since they also used the size of the graphs and the number of edges as explanatory variables which could identify the different domains alone. Furthermore, they did not normalized the attributes such as the number of total triangles, which has significantly different possible range of values in small and in large graph. The importance of the normalization is discussed in the subsequent subsection.

Another important problem, that whilst there are numerous graph metrics introduced in the last few decades, but it is still unclear that, whether there exists a complete non-redundant set of graph metrics that can fully describe the networks. The importance of this question reflects in the fact that there are several approaches, trying to find this set of attributes, and data-driven techniques seem to be the solution; for example in [26] Garcia-Robledo et al. gathered data about the network of the Internet, and followed data-driven approach to study the correlation of the graph metrics, and to find a complete non-redundant set. Similarly in [34] Filkov et al. gathered data from 113 real-world networks in order to find a set of metrics which enables comprehensive comparison between any two networks. Hence, in order to study this problem we again propose two different approaches, firstly in Section 3.2.1 we investigate the correlation of the graph metrics and then we will assemble a set of uncorrelated i.e. non-redundant variables. Furthermore, in Subsection 3.3.1 we will use machine learning techniques, namely feature selection algorithms, which primary goal is to select the relevant features for use in machine learning model training.

As we have already mentioned that our dataset is unique not just by its size but also by the involved fractality related measurements such as the estimated Zipf, Pareto and ML power-law fractal scaling parameters. Hence in the following subsections we will also apply the previously described data analysis techniques to explore the relationship of fractal related metrics with every other graph attributes, and compare the observations to the existing results of [14]. For example in Subsection 3.3.3 we will use regression techniques to estimate the fractal scaling parameters, and in Subsection 3.2.1 we will discuss the correlation of fractality with other graph metrics considering both real-world networks and models.

3.1 Dataset

In this section we introduce our gathered dataset, describe its real-world networks and detail the studied graph measurements. In the following subsection we introduce the background of our networks, and the necessary data-preparations, then in Subsection 3.1.2 we detail the chosen graph measurements.

3.1.1 Networks

As we have already mentioned, five different domains are studied, namely brain networks (mainly human neural networks), social networks (mostly Facebook friendship and Twitter retweet networks), food webs i.e. consumer-resource networks (graph representation of what-eats-what in an ecological community), cheminformatics networks (protein-protein interactions), and there is a fifth category of miscellaneous networks, which half is from the collection of DIMACS [36] and the other half is from the misc group from Network Repository [29], which contains a few real-world networks such as power-grid and flight networks, but there are a great deal of "synthetic" networks such as networks of optimization problems[11] from SuiteSparse [37] as well. The gathered dataset is balanced, meaning that the number of graphs from the different domains are proportional, namely there are 106 food, 100 brain, 100 misc, 99 cheminformatics, 61 DIMACS and 84 social networks.

The graphs were collected one by one from several different online databases [29, 30, 31, 32, 33], thus unfortunately the format of the graphs was not uniform, for example in some cases the graphs were given in weighted edge-list form, but sometimes in adjacency matrix, incidence matrix, or in the case of food networks in a third type of matrix representation was used. After importing these graphs in Wolfram Mathematica, we realized that some of the graphs are disconnected or directed, hence we took the largest connected component of each graph, converted to undirected, and removed the self-loops.

[11] For example see `https://sparse.tamu.edu/MathWorks/QRpivot`

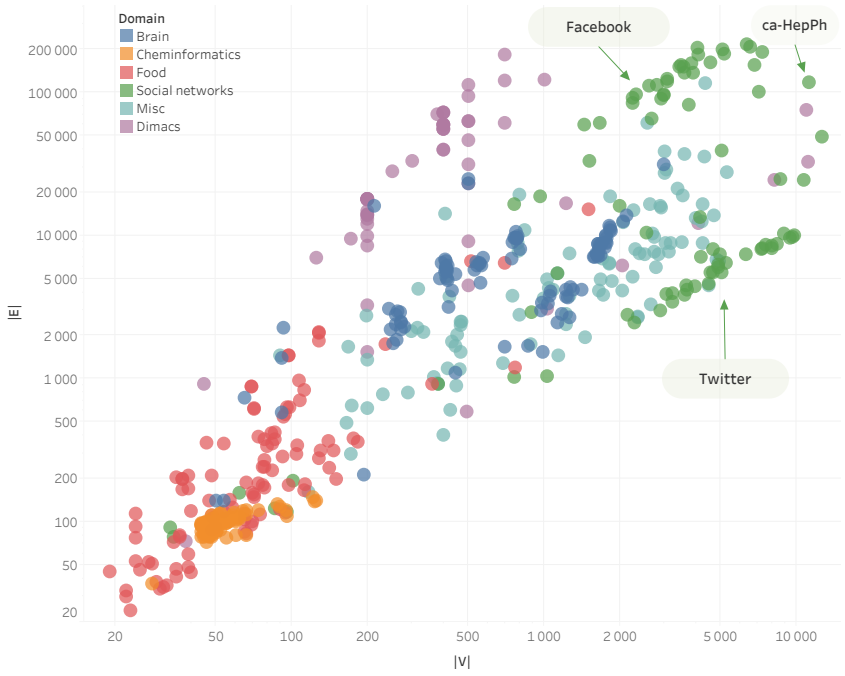


Figure 10: The number of edges and nodes of the gathered graphs on logarithmic scale. The different colors indicate different network domains. The figure was created in Tableau.

3.1.2 Metrics

After we succesfully imported every network, we calculated 32 different graph measurements listed in Table 3, and exported the results into a single dataset, which consists of 550 rows and each row corresponds to a graph, and the columns contain the different measurements, plus the name and domain of the graphs. The most important data preparation task of the obtained dataset is the normalization of the measures, since most of the metrics have different meaning considering different sized networks, specially the distance related attributes, for example a diameter of 10 can have different implications for a network of size 50 or 5000. As we have already mentioned in the introduction of Section 3, unfortunately there are studies where the authors did not take into consideration the normalization of the size-dependent measurements, which is only excusable, when the size of the studied networks are proportional to each other, but for example the food webs and protein-protein interaction networks are usually consists of a few dozens of

nodes, while the social networks topically involve thousands of participants.

Figure 10 shows the number of nodes and edges of the collected networks on a logarithmically scaled scatter plot. The figure also well illustrates that without the miscellaneous group, the rest of the domains could be easily partitioned by the number of edges and nodes, thus we excluded these variables from the set of explanatory attributes. Note that in the figure, while Facebook and Twitter networks have roughly the same size, the retweet graphs are significantly sparser. The scattered green (social) points between the two social media networks are the collaboration networks. For example one of our largest graph is the *ca-HepPh* is the well studied [86, 87] High Energy Physics - Phenomenology (HEP-PH) collaboration network, which represents scientific collaboration between authors who have papers submitted to HEP-PH category.

In the normalization process, we followed [24] where a great deal of metrics and their normalization constant are detailed, and for the variables that were not applied in [24] we proposed natural normalizing constants. We detail the used metrics and their normalization in Table 3, and we remind that the definitions of the metrics can be found in Section 1. Note that we did not include the average of the degrees, since in the case of undirected graphs, the normalized (by $|V|-1$) mean vertex degree coincides with the graph density:

$$\frac{\frac{1}{|V|}\sum_{v_i \in V(G)} \deg(v_i)}{|V-1|} = \frac{2|E|}{|V|\left(|V|-1\right)}.$$

The `MeanDegreeConnectivity` function of the Wolfram language computes the so-called $\langle k_{nn} \rangle$ neighbour connectivity, which gives a list of the k-mean degree connectivity for the graph for succesive $k = 0, 1, \ldots, \deg^{\max}$, where the k-mean degree connectivity is the average of the mean neighbor degrees of vertices of degree k. Thus, if this function is increasing in k then the network is assortative, and alternatively if the function is decreasing that means that the high degree nodes tend to connect to nodes of lower degrees, i.e. the graph is disassortative. The slope of these datapoints are denoted with the variable *slopeOfMeanDegreeConnectivity*.

Note that we have to interpret the meaning of the variables (*ParetoParameter, ZipfParameter, MaximumLikelihoodParameter*) of the estimated parameters of the fitted Pareto, Zipf and power-law distributions differently as their name

would suggest. We fitted these distributions to the previously calculated $N_B(l_B)$ datapoints of all of the networks, which means that we fitted power-law functions to non-power-law distributed data as well, and as a result of this in case of non-power-law data points the value of the estimated scaling parameters are lower (i.e. closer to one) than the estimated parameters of the power-law distributed data points. That is because if we consider the formula (26) of the maximum likelihood estimate of the scaling parameter, then if the x_i data points decay exponentially i.e. if instead of $x_i^{-\alpha}$ we have e^{-x_i} decay, then substituting this into the equation (26) or (30), then due to the cancellation of logarithm of exponentials, the summands are proportional to x_i, whilst in power-law case the summands are proportional to $\ln x_i$, hence after taking the reciprocal of the sum, for non-power-law distributed data points we obtain lower value of estimated scaling parameters, since $\ln x \ll x$. This argument also applies to the parameters which were obtained by the Wolfram Mathematica's built-in parameter estimator, since these parameters were also achieved by maximum likelihood method. Thus, for a given network these attributes indicates how fast is the decay of the $N_B(l_B)$ is, the more close they are to 1 the faster N_B decays.

Table 3: The graph metrics and their descriptions and normalization constants.

Variable	**Description**	**Normalized by**
Name	Unique name of the graph	—
Category	Domain of the network	—
$\|V\|$	Number of nodes	—
$\|E\|$	Number of edges	—
minDeg	Smallest degree in the graph	Maximum possible degree: $\|V\| - 1$
maxDeg	Largest degree in the graph	Maximum possible degree: $\|V\| - 1$
Scalefreeness	The absolute correlation of $\log P(k)$ and $\log k$	—
scalingOfDegreeDistribution	Scaling parameter of the logarithm of the degree distribution, obtained by linear regression procedure	—
scaelingOfMeanNeighbourDegrees	Scaling parameter of the logarithm of the mean neighbour degree distribution, obtained by linear regression method	—
slopeOfMeanDegreeConnectivity	Slope of the $\langle k_{nn} \rangle$ function	—
correlationOfMeanDegreeConnectivity	Correlation of $\langle k_{nn} \rangle$ and k	—
Diameter	Diameter of the graph	Longest possible path: $\|V\| - 1$
Radius	Radius of the graph	Longest possible path: $\|V\| - 1$
MeanGraphDistance	Mean of the graph distances	Longest possible path: $\|V\| - 1$
VertexConnectivity	Percentage of the vertex connectivity	—
EdgeConnectivity	Percentage of the edge connectivity	—
GraphLinkEfficiency	The link efficiency of the graph	—
GraphDensity	The graph density	—
maxBetweennessCentrality	Maximum of the betweenness centralities	Number of node pairs $\|V\|\,(\|V\| - 1)/2$
minBetweennessCentrality	Minimum of the betweenness centralities	Number of node pairs $\|V\|\,(\|V\| - 1)/2$
avgBetweennessCentrality	Mean of the betweenness centralities	Number of node pairs $\|V\|\,(\|V\| - 1)/2$
minEigenvectorCentrality	Minimum of the eigenvector centralities	—
maxEigenvectorCentrality	Maximum of the eigenvector centralities	—
maxEdgeBetweennessCentrality	Maximum of the edge betweenness centralities	Number of node pairs $\|V\|\,(\|V\| - 1)/2$
minEdgeBetweennessCentrality	Minimum of the edge betweenness centralities	Number of node pairs $\|V\|\,(\|V\| - 1)/2$
Assortativity	ρ assortativity coefficient	—
GlobalClusteringCoefficient	Global clustering coefficient	—
avgLocalClusteringCoefficient	Mean of the local clustering coefficients	—
VarLocalClusteringCoefficient	Unbiased sample variance of the local clustering coefficients	Squared mean local clustering coefficient
VarMeanNeighborDegree	Unbiased sample variance of the average neighbour degrees	Squared mean neighbor degree
VarVertexDegree	Unbiased sample variance of the degrees of the nodes	Squared average degree
ParetoParameter	Estimated scaling parameter of the fitted Pareto distribution	—
ZipfParameter	Estimated scaling parameter of the fitted Zipf distribution	—
MaximumLikelihoodParameter	MLE of the scaling parameter, obtained by the formula (30)	—

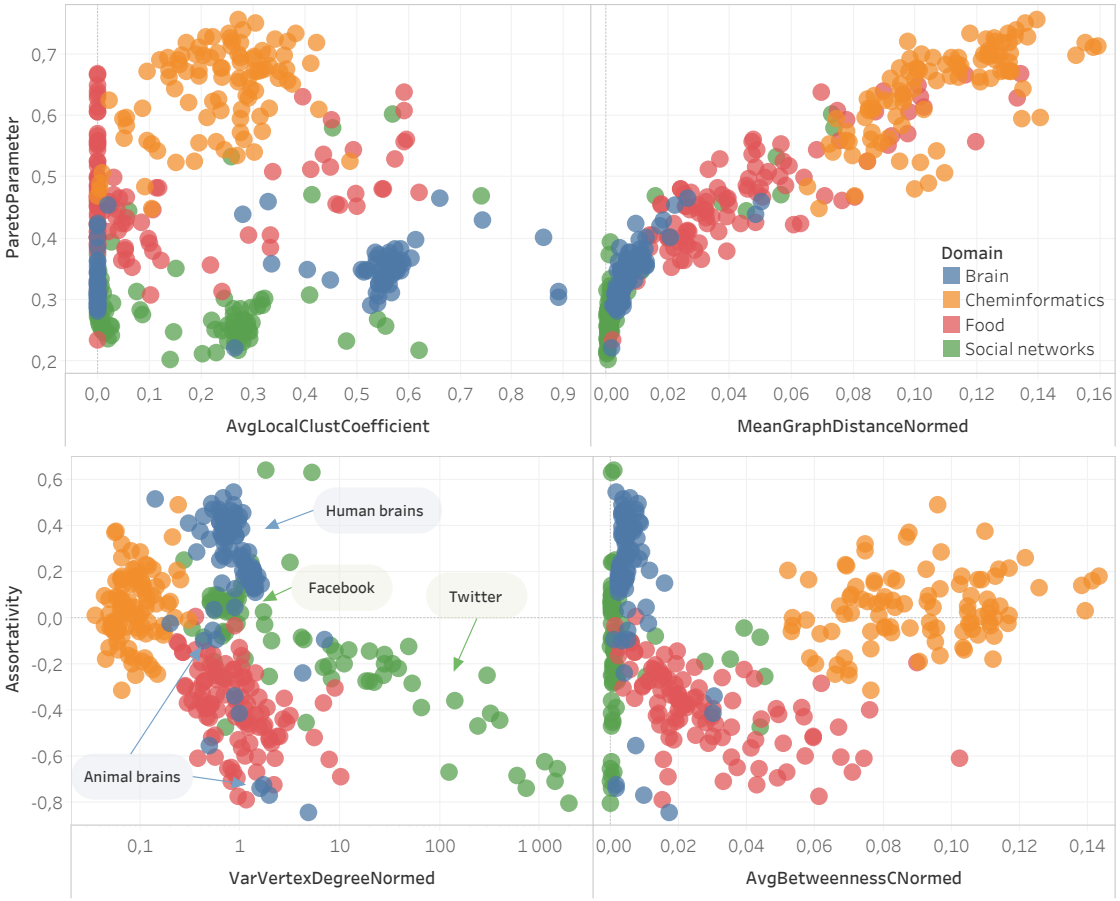


Figure 11: Scatter plots of the different attribute pairs of the networks. The different colors denote different domains. The figure was created in Tableau.

3.2 Exploratory data analysis

In this section we provide a brief exploratory analysis of the obtained dataset in order to give insights to the main characteristics of the domains of the networks. Figure 11 shows scatter plots of some of the determinative graph metrics. In the first row, the plots show the average local clustering coefficient and the normalized mean graph distance against the estimated scaling parameter of the Pareto distribution[12] of the previously calculated $N_B(l_B)$ datapoints. The bottom row represents the normed variance of the degrees on logarithmic scale and the the

[12]More precisely the scaling parameter minus one, since by the original definition of the Pareto distribution, for value x the probability density is proportional to $x^{-\alpha-1}$.

normalized average betweenness centrality versus the ρ assortativity coefficient. Considering the assortativity coefficient, the plots suggest that the food webs are typically disassortative, while the brain networks are assortative and the cheminformatics networks are symmetrically scattered meaning that they can be both assortative, disassortative and uncorrelated as well.

Note that even if two different graphs come from the same domain, they can differ significantly, for example while all of the blue points represent brain networks, the human neural networks are the assortative ones and the few disassortative points correspond to animal brain networks. More thorough study of this observation requires more neural graphs of the different animal species, however there are a few efforts that are related to our work, for example in the highly cited paper of Bullmore and Sporns [88] studied the graph properties of the human neural networks obtained by fMRI and electrophysiological techniqes, in addition L. Deuker et al. studied assortativity, clustering, density and shortest paths related metrics of the human brain functional networks derived from magnetoencephalography [89]. Similarly on the other three plots, the blue points which are in a cluster are the human neural networks, and all the "outlier" blue points, are originated from animals.

Similarly, in spite of the fact that both Facebook and Twitter networks are social, they differ in many characteristics, for example the variance of the degrees in retweet networks are in many orders of magnitudes larger, and the average local clustering coefficient is near zero in Twitter graph, while considering Facebook networks this coefficient is around 2.5.

The second plot suggests that there is a strong correlation between the scaling parameter and the normalized mean graph distance. We remind that we mentioned that these scaling parameters show that how quick the decay of the $N_B(l_B)$ is, and the higher the value of the parameter is, the more fractal the graph is. Thus, in accordance with the three-parameter-identification method introduced in Eq. (35), the second subfigure in Figure 11 suggests that, the social networks are rather on the non-fractal and small-world endpoint of the spectrum except a few collaboration network, the famous Karate club [13] and the social network of bottlenose dolphins [90]. Similarly, the brain networks are rather on non-fractal endpoint of

[13] See `https://en.wikipedia.org/wiki/Zachary%27s_karate_club`

the spectrum except the few animal neural networks. On the other hand, the food networks behave in a diverse range, they can be found on the whole spectrum, but mostly on the intermediate section, to what we referred to as mixture between fractal and small-world, i.e. in these networks the fractal scaling holds for small l_B's and for large values of l_B the small-world property dominates. Finaly, the networks of cheminformatics represent the rather fractal and non-small-world part of the palette, which is consistent with the observation of Song et al. in [14].

Note that the miscellaneous group of networks are filtered from the plots, since they were too scattered due to the various origin domains. However, considering the whole set of the networks, some correlations cancel out each other for example the assortativity and the average betweenness centrality, but if we focus on a single domain, we can observe, that there are domain-specific correlations, namely in food webs these attributes are correlated. Similarly, this is the case in social networks with the assortativity and the variance of degrees. We will discuss the domain dependent correlations in more detail later in the paragraph 3.2.1.1.

3.2.1 Correlation analysis

In this subsection we reveal the correlations between the different graph metrics, furthermore examine if these relationships are universal or are there any domain-dependent relations. The majority of the measurements can be categorized by the related graph properties as follows:

1. **Centrality related variables**: *AvgBetweennessCentrality, MaxEdgeBetweennessCentrality, MaxEigenvectorCentrality etc.*

2. **Shortest paths related variables**: *Diameter, Radius, LinkEfficiency, MeanGraphDistance etc.*

3. **Density related variables**: *GraphDensity, maxDeg, minDeg, vertex and edge connectivity etc.*

4. **Clustering coefficients**: *Assortativity, Global and local clustering coefficient, MeanDegreeConnectivity etc.*

Figure 13 shows that there are high correlations among the variables inside these groups as one would anticipate with good reason, furthermore, as Table 13 suggests, there are also non-trivial correlations among these groups. For example the shortest path related variables such as the diameter and mean graph distances are heavily correlated with the betweenness centralities. This is consistent with the study [24], which argued that it can be analytically shown that the normalized average betweenness centrality is linearly proportional to the normalized average path length. What is more, these two groups of the centralities and graph distances are also highly correlated with the $N_B(l_B)$ scaling parameters such as *Zipf*, *Pareto* and *MLE*, which is also well illustrated in the top-right sub-figure of Fig. 11.

Note that in [14] Song et al. argue that fractal networks are more robust than non-fractals, i.e. they are more resistant to targeted attacks. This is consistent with the fact that *EdgeConnectivityPercentage* variable is highly correlated with the fractal scaling parameters, however it is interesting that the *VertexConnectivityPercentage* is rather correlated with the density related attributes, but it is still moderately correlated with the fractal related *MLE* parameter. Further surprising phenomena is that the maximum degree and the maximum betweenness centrality are more connected to the clustering related measures specially to the assortativity, then to their own groups.

The normalized minimum betweenness centrality feature turned out to be the most uncorrelated variable, that is because this measurement is usually zero in inhomogeneous networks. Although the computation of the *Scalefreeness* parameter is rather based on heuristic arguments, its most correlated companion is the scaling parameter of the degree distribution, which indicates that it indeed denotes the scale-freeness of the networks.

Figure 12 shows the community plot of a weighted adjacency graph as an alternative visualization of the considerably correlated measurements with the absolute coefficients greater than 0.43. The community structure of the graph was obtained via the modularity-based clustering method, introduced by Newman et al. in [91]. The graph illustrates well that the centrality and the distance related metrics are so correlated, that they have been merged into a single clique denoted with the red nodes. The other two communities marked by purple and yellow, are the density-

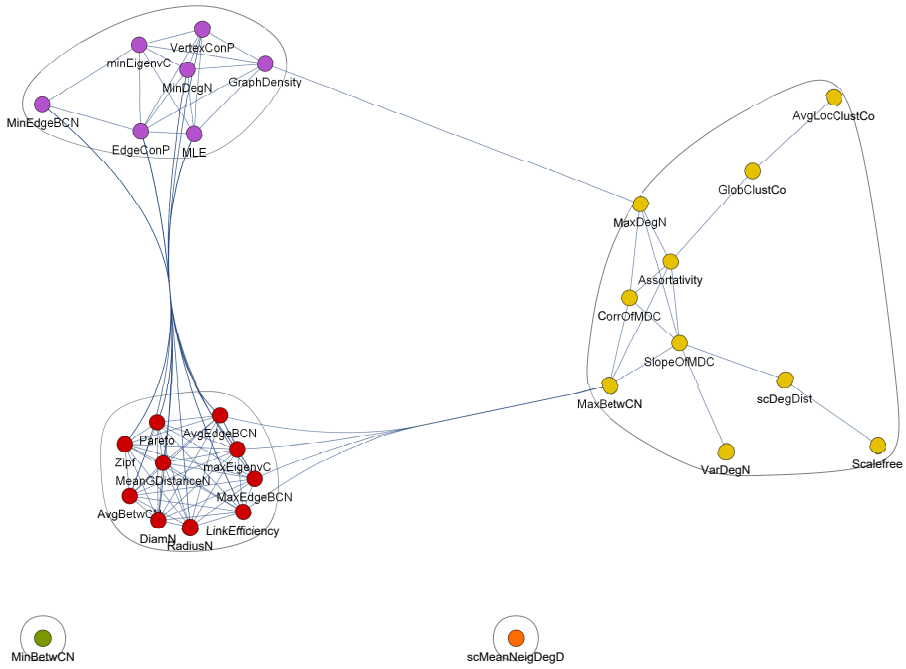


Figure 12: Weighted community graph of the measurements, whose absolute correlation is greater than 0.43. The figure was created in Wolfram Mathematica with the help of `CommunityGraphPlot` function.

robustness related (vertex- and edge connectivity, graph density) and node-level features (assortativity, clustering coefficients and degree distribution) respectively. Note that the graph also shows that while the density and distance metrics are in different groups, they are connected through multiple variables, which is in contrast of the observation of [24], where the authors find that density and distance metrics form two orthogonal groups of highly correlated metrics, but there are other existing results that support our results [26].

The graph representation of the strong correlation also helps to select a non-redundant set of the metrics, since all we have to do is to chose one highly connected metric from each community. The completeness of the chosen non-redundant set can be evaluated by machine learning models, by setting the explanatory variables to be the selected non-redundant set. For more details, and results see Subsection 3.3.2.

Note that when we crated the correlation tables, we excluded the graphs of the DIMACS collection, because these graphs were very distinct from all of the

other domains, and behaved oppositely as the real-world networks in almost every attribute, hence these graphs significantly distorted the correlation of the graph metrics.

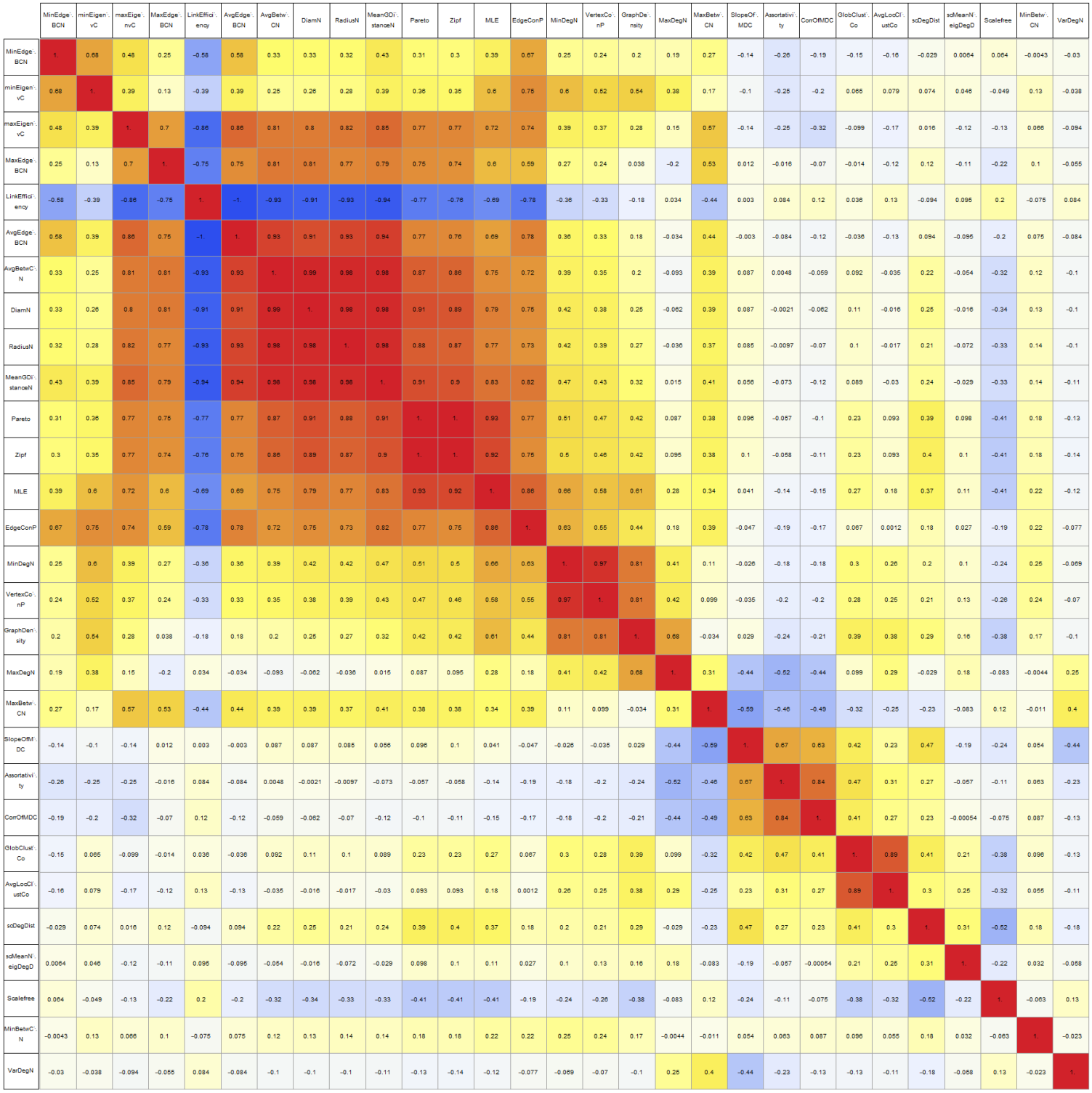


Figure 13: The temperature map of the correlation table of the graph metrics. The shades of the color red indicates positive, and the shades of the color blue indicates negative correlation. The more vivid the tone is, the stronger the relationship is. The variables are ordered, such that the highly correlated variables are close to each other. The figure was created in Wolfram Mathematica.

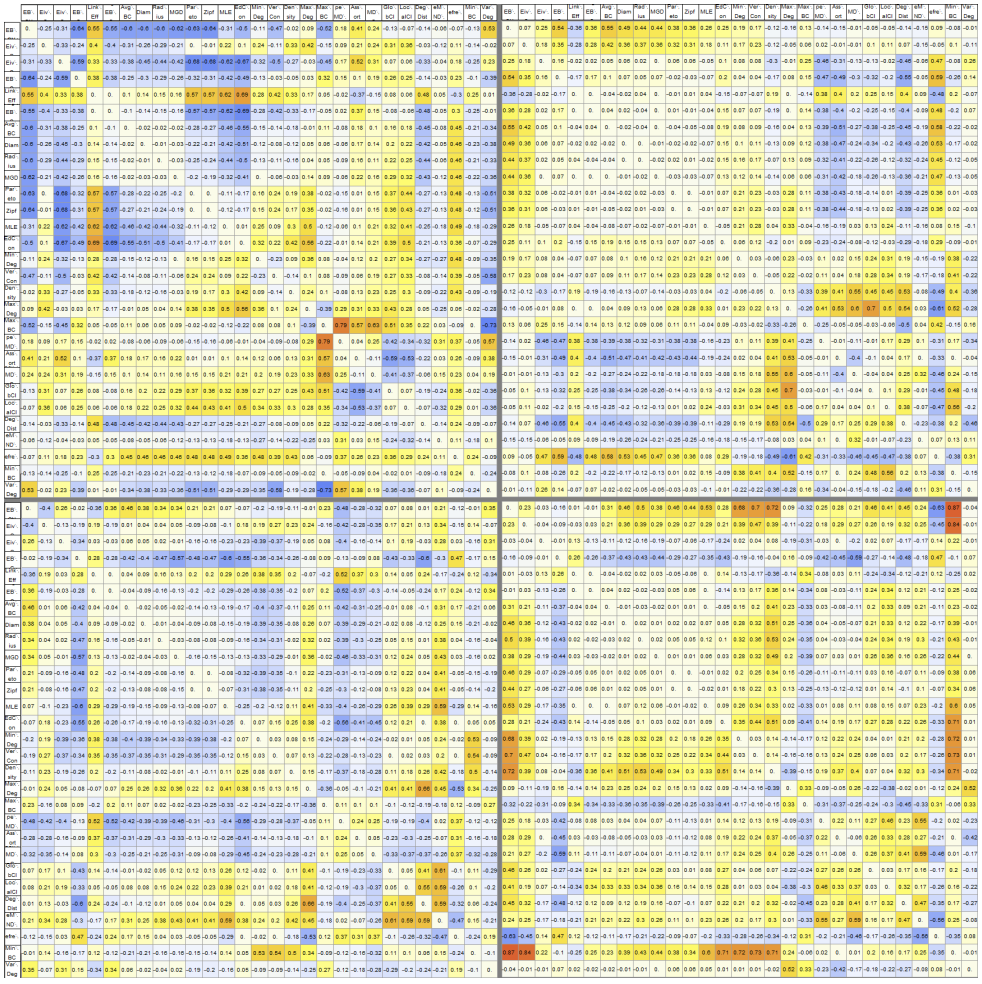


Figure 14: The correlation differences from the correlation Table 13. In the first row, the first table corresponds to the cheminformatics, the second to the food networks. In the second row the first correlation map belongs to the brain, the second belongs to the social networks.

3.2.1.1 Domain specific correlations

However real-world networks indeed share some common characteristics, there are unique properties as well. As the author of [24] pointed out, we have to be cautious when we investigate networks from different fields together. We have already seen that normalization is a crucial step that cannot be omitted before comparing different networks, but networks do not only differ in their sizes but there are more latent disparities concealed in the structure of the graphs, such

as the relationships of the metrics. Due to the domain-specific properties there are even some contradictory observations regarding the relationship of density and distance related metrics in [24] and [26].

In this paragraph we attempt to reveal these latent unique traits of the different network domains. Figure 14 illustrates the domain-specific correlations, more precisely, the figure consist of the cheminformatics, food, brain and social networks correlation matrix difference from the overall correlation matrix represented in Figure 13, i.e. we calculated the correlations of the variables using only the rows of the database that corresponds to a single domain, then from each domain-specific correlation matrix we subtracted the "average" correlation matrix (Figure 13), which was calculated using all rows of the database.

The cheminformatics domain is the most separate domain, which is also discernible on the scatter plots of Figure 11. On the top-left part of the correlation difference figure the vivid blue colors indicate that in the protein-protein interaction networks the correlation between the centrality measures (except the minimum eigenvector centrality) and the fractality related parameters are not as strong as in the other domains, moreover the fractality measures are rather correlated with the density-robustness related metrics. Since, these networks seem to have fractal structure, this supports the connection of robustness and fractality mentioned in [14].

The top-right temperature map corresponds to the food networks, and here the blue shades represents the negative correlation of the assortativity and the betweenness centrality related measures. Hence this is a food web specific phenomena and it was also illustrated in the fourth plot of Figure 11. Note that the food webs are from the only domain where the assortativity and the fractality have moderate anti-correlation, hence the conjecture of Havlin et al. that disassortativity leads to fractal structure indeed holds under certain circumstances, but it does not apply generally to every network. Furthermore, there is a slightly higher connection between the density and the clustering related measurements.

In the brain networks (left-bottom), the degree distribution related variables are relatively highly correlated with the density related variables. Overall, the social networks behave normally, i.e. the values of correlations within this domain agrees with the correlations obtained from the whole dataset. On the other hand,

the the correlation between fractal related measurements and other metrics differ significantly from what we observed in other domains. This phenomenon is due to the fact that most of the social networks are non-fractal and the measure of their fractality is rather concentrated around a value as it can be seen in Figure 11.

Note that some of these correlation differences are corrupted by a very few number of outliers, but before drawing conclusions, we inspected these correlation differences on scatter plots, and we only highlighted those cases, which were not distorted by outliers. In addition, the reason behind these unique correlation differences requires further investigations and deeper understating of the underlying domains.

3.3 Statistical learning

Supervised learning is a machine learning task, where we approximate (or learn) a function that maps an input to an output based on the input-output pairs of the provided dataset [92]. The output is usually called as label, and the label is chosen according to the aspired goal. Here we apply machine learning algorithms to achieve multiple goals, one of them is to study whether the collected graph metrics are able to identify the different domains. In this case we will set the label to be the names of the domains. However, in Subsection 3.3.3 we will estimate graph metrics such as the scaling parameter of the $N_B(l_B)$ function, thus here we will set the label to be value of these parameters.

The difference between supervised and unsupervised learning is that in case of the latter one we do not have labels, hence in an unsupervised learning task we try to deduce a function that describes the hidden structure of the data, for example cluster analysis and density estimation. In supervised learning we partition our data to training set and test set. The models learn the connections between the attributes on the training set, and then we evaluate the accuracy of the models on the test set, i.e. perform predictions and compare the results to the *true* values of the labels. Naturally, we say that the better the model performs, the more close its prediction to the real value, the concept of closeness depends on the type (categorical of quantitative) of the label. Note that when the label is a categorical variable, such as the domains, we refer to it as *classification* and when the label

is quantitative variable we call it *regression* task.

Countless machine learning algorithms have been introduced over the last decades, and most of them can be applied to both classification and regression tasks. Here we will only apply the most frequently used models, namely Decision Tree (DT), Naive Bayes (NB), Generalized Linear Model (GLM), Random Forest (RF), Gradient Boosted Trees (GBT) and Deep Learning (DL). Note that even though some of the models rely on deep and beautiful mathematics, for reasons of space, here we cannot go into detail, but for a great overview and thorough understanding of this topic we recommend the book of Elements of Statistical Learning [93]. In this work for data preparation, modelling and evaluating we used RapidMiner, a visual workflow designer data science tool.

Roughly speaking, the goal of every supervised learning model can be written as follows: Let $\mathbf{X}$ denote the $N \times d$ input matrix, which consists of N observations of d attributes. The observations of the jth attribute is denoted by $\mathbf{X}_{\cdot,j}$, and the ith observation vector is denoted by $\mathbf{X}_{i,\cdot}$. The label or target variable is typically denoted by $Y = (y_1, \ldots, y_N)$. Then in a function-fitting paradigm we say that the model attempts to learn an f function by the examples, for which $f(\mathbf{X}) \approx Y$, where the "$\approx$" can be interpreted in many ways, but the goal is that in some way minimmize the difference between $f(\mathbf{X})$ and Y. The most usual objective function is the mean squares of the deviations, since the square function can be easily handled due to its convexity and differentiability. The approximations that minimizes this objective are called Least Squares estimators. Note that the $f(\mathbf{X})$ notation is sliglthly ill, and more precisely we should write $f(\mathbf{X}_{i,\cdot}) \approx y_i$, $\forall i$, since the the input of f and $\hat{f}$ is a vector, and the output is a scalar.

The learning algorithms' estimated f function is usually denoted by $\hat{f}$, and these algorithims have the ability to modify its input-output relationship $\hat{f}$ in response to differences of $y_i - \hat{f}(\mathbf{X}_{i,\cdot})$ between the original and the estimated outputs. This process is known as learning by examples, and upon the completion of the learning process, the hope is that the estimated and the real outputs will be close enough to be useful for all sets of inputs [93], and that the models' estimated function helps us to understand the underlying data and its origin.

Table 4: Most important and relevant metrics according to different techniques

Method	**Selected variables**
Forward selection	maxEdgeBetweennessCentrality, Diameter, VertexConnectivityPercentage, MaxDegree, MeanGraphDistance, avgBetweennessCentrality Scalefreeness
Backward elimination	All variables except the VarVertexDegree scalingOfDegreeDistribution, ParetoParameter and minBetweennessCentrality
Evolutionary optimization	maxEigenvectorCentrality, Radius, Assortativity maxEdgeBetweennessCentrality, minDegree, maxDegree, GlobalClusteringCoefficient, maxBetweennnessCentrality, ZipfParameter avgBetweennessCentrality, EdgeConnectivityP, GraphDensity, Scalefreeness,CorrOfMDC
Correlation (Global importance)	ParetoParameter, maxEigenvectorCentrality, MeanGraphDistance, MLEParameter, Diameter, avgBetweennessCentrality, GraphLinkEfficiency
Correlation graph	GraphDensity, Assortativity MeanGraphDistance, GlobalClusteringCoefficient

3.3.1 Feature selection

In this subsection we will perform different feature selection methods. We apply both machine learning algorithms, and correlation-based techniques, namely we will examine the correlation of the variables with the response variable, furthermore we will employ the results of Subsection 3.2.1, especially the graph representation of the correlation temperature map.

The ultimate goal of every feature selection (FS) algorithm, is to identify a relevant subset of the explanatory variables for use in model construction and for better understanding of the data. These techniques are used for multiple reasons, for example a model that uses a few but relevant attributes is easier to interpret, if we successfully narrowed down the subset of attributes, we can avoid the curse of dimensionality [94], furthermore we can reduce the risk of overfitting. Here we use three different FS algorithm; Forward Selection, Backward Elimination and Evolutionary Optimization. These FS algorithms are detailed in [93], but briefly

they use simple machine learning models, here we used generalized linear model[14], and the forward selection starts with an empty set and then iteratively adds one variable to the model (and the set) at a time. In each step every variable (which is not in the already selected set) is tested for inclusion in the model, and then that variable is selected which improved the most significantly the performance of the model. Backward elimination is using analogous idea but it works backwards, i.e. starts with every variable and eliminates the non-relevant ones.

The evolutionary FS algorithm is a Genetic algorithm which belongs to a larger class of Evolutionary algorithms, which provide solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. In the genetic FS algorithm selection mutation means switching features on and off and crossover means interchanging used features, furthermore selection is done by the specified selection scheme, here we used tournament selection [95].

The correlation based variable importance calculations are independent from the models. After keeping one variable from the highly correlated attributes, we simply assign weights to the variables according to their correlation with the target variable. Alternatively we can chose the most connected nodes from the communities of the graph shown in Figure 12.

Table 4 shows the most relevant and important features selected by the three feature selection algorithm, and the two correlation based technique. We set the target variable to be the name of the domains. The table suggests that both forward and backward selection algorithms reach their maximization relatively early. We will evaluate the selected subsets with machine learning models in the next Subsection 3.3.2.

3.3.2 Classification

In this subsection we examine whether is it possible to efficiently predict the domain of the networks using the graph measurements. As we have already mentioned before, for the predictions we will train several machine learning algorithms on different training sets obtained by the different feature selection methods. Note

[14]See `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html`

Table 5: Confusion matrix of a decision tree, using only the $|V|$ and $|E|$ attributes.

Accuracy: 79.59%		**True domain**					
		Social	**Misc**	**Food**	**Chems**	**Brain**	**Precision**
Predicted domain	**Social**	13	2				86.67%
	Misc	3	14	1		2	70.00%
	Food			20	4	3	74.07%
	Chems	1			16		94.12%
	Brain		4			15	78.95%
	Recall	76.47%	70.00%	95.24%	80.00%	75.00%	

that we excluded the number of nodes and edges from the set of explanatory variables, since these variables can determine effectively the origin domain of the networks (see Table 5).

The performance of the models are calculated associated with the confusion matrix. The confusion matrix in itself is not a performance metric, however almost every performance metric is based on the values in this matrix. In our case we have five different class, hence the confusion matrix is a 5×5 table and the rows correspond to the predicted values and the columns correspond to the real values of the domains, and in the cell of the ith row and jth column we write the number of cases when the actual class was the jth domain and the predicted class is the ith domain. For example Table 5 shows the confusion matrix of the decision tree classifier, which only used the number of edges and nodes as explanatory variables. Whit this example we wanted to highlight the predictive power of these two variables, and to illustrate why these variables should not be involved in the set of explanatory variables.

The class precisions, class recalls and the total accuracy are also shown in Table 5. These are frequently used performance metrics, the accuracy is the ratio of the number of hits (total of the diagonal) and the size of the test set. Note that the accuracy is a good measure of the performance only if the target variable classes in the data are nearly balanced, i.e. there are roughly equal number of graphs from each domains. The class precisions are interpreted rowwise, i.e. it is the number of accurate predictions divided by the total number of predictions of a given class. The definition of class recall is analogous to the precision, but the recall is interpreted columnwise i.e. the recall of a given C class is the number of

Table 6: Performance of the machine learning models on the different sets created by the FS methods detailed in Table 4

	Feature selection method					
	Backward	Forward	Evolutionary	Correlation	Graph	average
NB	69.4%	55.1%	69.4%	50.0%	**72.4%**	63.26%
GLM	77.6%	81.6%	**90.8%**	75.5%	68.4%	78.78%
DT	81.6%	84.7%	84.7%	71.4%	**85.7%**	81.62%
DL	88.8%	77.6%	**88.8%**	64.3%	67.3%	77.36%
RF	**90.8%**	83.7%	89.8%	82.7%	87.8%	86.96%
GBT	90.8%	87.8%	**91.8%**	89.8%	89.8%	**90.00%**
average	83.17%	78.42%	**85.89%**	72.28%	78.56%	

hits of the class C divided by the total number of examples that true label is C.

Formally let $M = (m_{ij})$ denote the $k \times k$ confusion matrix of a k-class problem. Then the recall and the precision of the ith class is defined as:

$$Recall_i = \frac{m_{ii}}{\sum_{j=1}^{k} m_{ij}},$$

and

$$Precision_i = \frac{m_{ii}}{\sum_{j=1}^{k} m_{ji}}$$

respectively, furthermore the accuracy of the classifier is defined as

$$Accuracy = \frac{\sum_{i=1}^{k} m_{ii}}{\sum_{j,l=1}^{k} m_{jl}}.$$

For further performance metrics and their evaluations we refer to [96]. As both Figure 10 and Table 5 suggest, the most difficultly identifiable class is the Miscellaneous, and the best performances were obtained on Cheminformatics and Food network domains.

Table 6 shows the accuracy of several machine learning models that were trained on different sets acquired from the 5 feature selection methods detailed in the previous subsection and Table 4. The table suggest that the poorest performances were obtained apparently on the correlation based method, on the other hand the best accuracies are belong to the method of evolutionary optimization.

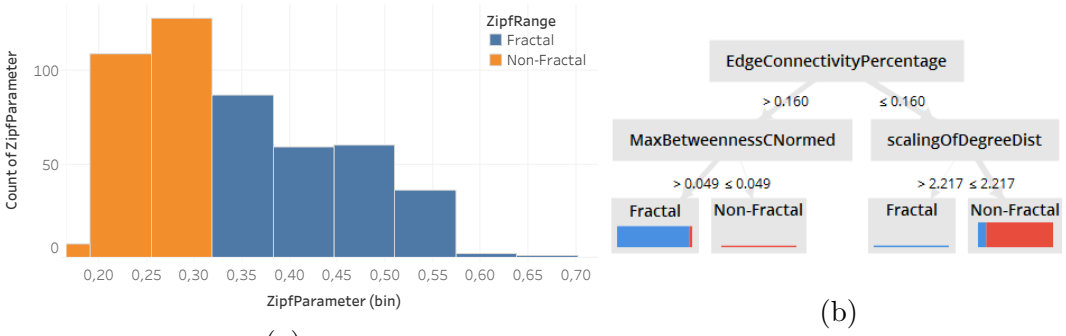
Overall the most accurate model was the Gradient Boosted Tree with accuracy of 91.8%, but the achievement of the Random Forest and the Generalized Linear Model are also remarkable. The reason why the Deep Learning did not produce high efficiency is because this model performs better on datasets of larger number of rows. Thus we can conclude that the domains of the networks are unique enough to be effectively identified by their normalized metrics.

Note that the training set was set to be the 70% of the original dataset, and it was built by stratified sampling i.e. we built random subsets ensuring that the class distribution in the subsets is the same as in the whole dataset. Furthermore, some of the hyperparameters of the models were previously optimized via grid-optimization, for example the number of trees and the depth of the trees in the tree-based models.

3.3.2.1 Fractality classification

Furthermore, we solved a fractality related classification problem as well. We divided the networks into two equal sets according to the *ZipfParameter*, and for now lets call those networks fractal whose *ZipfParameter* is in the range of larger scaling parameters. Figure 15a shows the histogram of the *ZipfParameter*, and the different colors indicate these newly defined sets. We used the attributes that were suggested by the evolutionary optimization FS algorithm, except the metrics that are highly correlated with the *ZipfParameter*, namely the centrality and the distance related measures. Then we set the label to be this newly defined binary variable and predicted the binarized fractality. The performance of the models are listed in Table 7 and their ROC curves are shown in Figure 16.

A ROC curve is an easily interpretable visual representation of the diagnostic ability of a binary classifier, and most frequently it is used to compare the performance of different models. In binary classification problems we usually refer to the clasess as positive and negative. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (TPR) at various threshold setting. The TPR is the binary version of the recall and it can be viewed as the statistical power. The FPR is the Type I error i.e. it is the number of false alarms i.e. false positives divided by the number of true negatives. The greater the



(a) (b)

Figure 15: Figure 15a shows the histogram of the *ZipfParameter*. The divisor point is at 0.319. Figure 15b shows the decision tree of the binarized fractal classification problem.

area under the ROC curve (AUC) is the more accurate the model is. Hence even without the highly correlated metrics, we are able to succesfully predict the binarized fractality of the networks. The decision tree, shown in Figure 15b, uses only *EdgeConnectivity*, *MaxBetweennessCentrality* and the *scalingOfDefreeDistribution* metrics to predict the class, and still achieves performance of 92.9% accuracy.

	Accuracy	AUC
NB	90.8%	0.987
GLM	92.9%	0.973
DT	92.9%	0.930
DL	93.9%	0.981
RF	93.9%	0.983
GBT	93.9%	0.992

Table 7: Accuracy and AUC of the models in the the binary fractality classification task

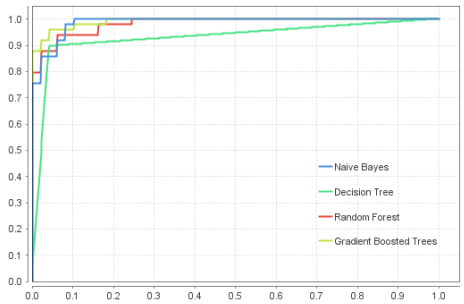


Figure 16: ROC curves of the models regarding the binary fractality classification

3.3.3 Regression

In this section we use regression analysis to estimate the value of the continuous graph measures, such as the scaling parameters of the $N_B(l_B)$ function, assortativity, centrality and distance related metrics. We apply similar procedures, used

in the previous subsection, i.e. after selecting a target variable, we train the models both including and excluding the variables that are highly correlated with the target variable. In other words, considering the correlation graph in Figure 12, we delete the node of the label variable and all of its neighbours, and the remaining nodes of variables make up the set of explanatory variables, for example if the target variable is the *ParetoParameter*, then we exclude the whole community marked by red.

Since, here we work with continuous variables, other metrics should be considered to measure performance. In this work we use the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the lenient version of the Mean Absolute Percentage Error (MAPE), defined as follows: Let $\mathbf{Y} = (y_1, \ldots, y_n)$ denote the vector of n observed values of the target variable, and $\hat{\mathbf{Y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ be the predictor's estimation of $\mathbf{Y}$. Then the Mean Squared Error (MSE) of the predictor is computed as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 , \tag{97}$$

and the RMSE is defined as the square root of the MSE, i.e.

$$RMSE = \sqrt{MSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)^{\frac{1}{2}} , \tag{98}$$

i.e. RMSE is the ℓ^2 norm of the error vector $\mathbf{Y} - \hat{\mathbf{Y}}$, which is the distance of the $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ vectors in the ℓ^2_n metric space. The MAE is analogously defined as the ℓ^1_n-distance of the vectors $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, i.e.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| . \tag{99}$$

The lenient version of the MAPE is related to the absolute error, namely the absolute error is divided by the maximum of the actual value and the predicition. Note that in the strict version of MAPE, the absolute error is divided by the

minimum of those terms. Formally it is defined as

$$LMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\max\{y_i, \hat{y}_i\}} \right|. \tag{100}$$

Note that in the original version of MAPE the denominator is the actual value y_i, and sometimes it is referred to as Mean Relative Error.

Table 8 and Table 9 shows the efficiency of the ordinary machine learning models according to the previously introduced performance metrics. Regarding Table 8, we can conclude that the fractality of the networks i.e. the estimated *ParetoParameter* is well identifiable even without using the highly correlated attributes.

On the other hand, in case of the *Assortativity* the estimation appears to be a difficult task, since the best performance in LMAPE is 42.2%. This suggests that assortativity behaves like an independent variable, i.e. even if most of the graph metrics of two different networks are proportional to each other, it is not impossible that one of them is assortative and the other is disassortative. This is well illustrated in the bottom-left plot of Figure 11, where we have seen that protein-protein interaction networks can be both assortative and disassortative, furthermore we have have also shown fractal models in Subsection 2.4 that differ in assortativity.

In Table 9, the labels were set to be the variables *Diameter* and *AvgBetweennessCentrality*, and the obtained performance metrics suggest that the level of difficulty of these tasks are equivalent. Overall the best two models were again the Gradient Boosted Tree and the Random Forest, however the simple Decision Tree also achieved remarkable efficiency, specially in Table 10, where the cells contain the LMAPE performance of the predictors, that could use the highly correlated variables as well. The results detailed in Table 10 are also consistent with the difficulty of assortativity identification, since even if its correlated companions were included in the training set, the best performance of LMAPE equals to 27.8%.

Table 8: Performance of the regressions of the *ParetoParameter* and the *Assortativity* metrics

Label	**ParetoParameter**			**Assortativity**		
	RMSE	**MAE**	**LMAPE**	**RMSE**	**MAE**	**LMAPE**
DT	0.061	0.043	10.4%	0.218	0.155	47.6%
DL	0.058	0.040	10.0%	0.182	0.130	52.1%
GLM	0.053	0.042	10.8%	0.227	0.162	58.8%
RF	0.049	0.034	8.8%	0.186	0.124	48.7%
GBT	**0.047**	**0.033**	**8.5%**	**0.177**	**0.113**	**42.2%**

Table 9: Performance of the regressions of the normed *Diameter* and the *AvgBetweennessCentrality* metrics

Label	**Diameter**			**AvgBetweennessCentrality**		
	RMSE	**MAE**	**LMAPE**	**RMSE**	**MAE**	**LMAPE**
DT	0.025	0.015	32.6%	0.009	0.005	32.3%
DL	0.034	0.024	43.5%	0.010	0.007	47.8%
GLM	0.033	0.023	50.4%	0.011	0.008	57.9%
RF	**0.026**	**0.015**	**32.4%**	**0.007**	**0.004**	**30.0%**
GBT	0.028	0.016	34.3%	0.009	0.006	40.6%

Table 10: LMAPE of the regressions of the *ParetoParameter, Diameter, Assortativity* and the *AvgBetweennessCentrality* mesasures by involving the highly correlating variables in the set of explanatory variables.

	Pareto	**Assort.**	**Diam.**	**AvgB.C.**
DT	**0.5%**	35.9%	11.3%	**10.9%**
DL	2.9%	32.5%	48.0%	60.2%
GLM	3.9%	40.3%	37.8%	40.2%
RF	3.0%	31.7%	16.0%	17.0%
GBT	**0.4%**	**27.8%**	**7.8%**	24.0%

4 Summary and conclusion

The purpose of this work was to explore and understand the fractality moreover the common and unique structure of complex networks from different domains. We conducted a comprehensive literature review, but also highlighted the contingent deficiency of mathematical rigour of related papers, proposed new methods, and pointed out the possible pitfalls of the wildly used techniques.

After giving a general introduction of network theory and data science, we detailed the most important concepts of the topic, we acquired an extensive understanding of fractal behavior of complex networks. We presented graph related box-covering algorithm, adopted from fractal theory, by which the fractality of the networks is defined. Besides the frequently used compact box burning box-covering algorithm, we presented, implemented and compared other novel alternative algorithms by both their running time and efficiency.

The problem of accurate identification of fractal dimension of networks, relies on the correct detection and fitting of power law distribution in the empirical data, obtained by the box-covering algorithm. We pointed out, that in several articles the validation of power-law distributions are carried out non-rigorously, hence we uncovered a statistical framework that utilize maximum-likelihood fitting for discerning and quantifying power law behavior in data, and later in this work we applied multiple variants of this proposed technique.

Motivated by our observations, and by the contrariety of the pure small-world and pure fractal properties, we proposed an extended, mathematically more rigorous definition of fractal networks, allowing a network to be locally fractal but globally small-world, and we showed several real-world and model generated examples that embrace this phenomena.

Throughout this paper, we followed two main distinct approaches, firstly we attempted to understand the fractality through well-studied mathematical network models from the literature and our newly proposed models. We investigated how the graph metrics vary as the models transit from non-fractal to fractal and vice versa. We also detailed a strong law of large numbers for the maximum degree of a modified version of the Barabási–Albert model.

For the second approach, we gathered a unique dataset, containing graph met-

rics of numerous real-world networks, and we followed data-driven analysis techniques for a rather general purpose. We did not only focus on the fractal property of the networks, but investigated a rich set of metrics that describe every aspects of the networks. We applied both traditional statistics and machine learning methods to find out how the values of the graph metrics are distributed on real-world networks, how the metrics are correlated with each other, how the correlations vary on different network domains. Furthermore, efficiently solved classification and regression tasks, such as identify the network domains, or the fractality by the metrics, and estimate an appointed graph measurement value using different selections of the remaining metrics.

There are several theoretical, empirical and technical open questions in this topic, such as analytically investigate the relationship of fractality and other graph metrics, and understand why fractal, non-fractal and mixture between fractal and non-fractal network are present in nature, and in different network domains. There are several possible future directions for the further research such as analytical investigation of the proposed novel network models, introducing new statistical algorithms for measuring fractality, and employing other data-driven techniques such as ANOVA and PCA analysis for better understanding of the networks. Finally, with data science techniques, we aim to tune the parameters of the models, such that they describe the real-world networks as accurately as possible.

References

[1] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific american*, vol. 288, no. 5, pp. 60–69, 2003.

[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[3] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the national academy of sciences*, vol. 97, no. 21, pp. 11 149–11 152, 2000.

[4] M. E. Newman, "Properties of highly clustered networks," *Physical Review E*, vol. 68, no. 2, p. 026121, 2003.

[5] K. Klemm and V. M. Eguiluz, "Highly clustered scale-free networks," *Physical Review E*, vol. 65, no. 3, p. 036123, 2002.

[6] C. I. Del Genio, T. Gross, and K. E. Bassler, "All scale-free networks are sparse," *Physical review letters*, vol. 107, no. 17, p. 178701, 2011.

[7] P. Erdos and A. Rényi, "On the evolution of random graphs. 1960," *Publ. Math. Inst. Hungar. Acad. Sci*, vol. 5, p. 17, 1973.

[8] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, "Cycles structure and local ordering in complex networks," *arXiv preprint cond-mat/0212026*, 2002.

[9] A.-L. Barabási, *Network science*. Cambridge university press, 2016.

[10] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[11] M. P. Stumpf and M. A. Porter, "Critical truths about power laws," *Science*, vol. 335, no. 6069, pp. 665–666, 2012.

[12] A. D. Broido and A. Clauset, "Scale-free networks are rare," *arXiv preprint arXiv:1801.03400*, 2018.

[13] M. Golosovsky, "Power-law citation distributions are not scale-free," *Physical Review E*, vol. 96, no. 3, p. 032306, 2017.

[14] C. Song, S. Havlin, and H. A. Makse, "Origins of fractality in the growth of complex networks," *Nature Physics*, vol. 2, no. 4, p. 275, 2006.

[15] M. Nagy, *Fractal Networks and Assortativity*. Bachelor's Thesis, Department of Stochastics, Budapest University of Technology and Economics, 2016.

[16] C. Song, S. Havlin, and H. A. Makse, "Self-similarity of complex networks," *Nature*, vol. 433, no. 7024, p. 392, 2005.

[17] S.-H. Yook, F. Radicchi, and H. Meyer-Ortmanns, "Self-similar scale-free networks and disassortativity," *Physical Review E*, vol. 72, no. 4, p. 045105, 2005.

[18] L. Kuang, B. Zheng, D. Li, Y. Li, and Y. Sun, "A fractal and scale-free model of complex networks with hub attraction behaviors," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–10, 2015.

[19] B. Zheng, H. Wu, L. Kuang, J. Qin, W. Du, J. Wang, and D. Li, "A simple model clarifies the complicated relationships of complex networks," *Scientific reports*, vol. 4, p. 6197, 2014.

[20] D. Li, X. Wang, and P. Huang, "A fractal growth model: Exploring the connection pattern of hubs in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 471, pp. 200–211, 2017.

[21] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi *et al.*, "A survey of statistical network models," *Foundations and Trends® in Machine Learning*, vol. 2, no. 2, pp. 129–233, 2010.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, "Characterization of complex networks: A survey of measurements," *Advances in physics*, vol. 56, no. 1, pp. 167–242, 2007.

[24] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles," *Physical Review E*, vol. 85, no. 1, p. 016117, 2012.

[25] A. Jamakovic and S. Uhlig, "On the relationships between topological measures in real-world networks," *Networks and Heterogeneous Media*, vol. 3, no. 2, p. 345, 2008.

[26] A. Garcia-Robledo, A. Diaz-Perez, and G. Morales-Luna, "Correlation analysis of complex network metrics on the topology of the internet," in *Emerging Technologies for a Smarter World (CEWIT), 2013 10th International Conference and Expo on*. IEEE, 2013, pp. 1–6.

[27] R. Kohavi, "Glossary of terms," *Machine Learning*, vol. 30, pp. 271–274, 1998.

[28] M. Nagy and R. Molontay, "Predicting dropout in higher education based on secondary school performance," in *Intelligent Engineering Systems (INES), 2018 IEEE 22nd International Conference on*. IEEE, 2018.

[29] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. [Online]. Available: http://networkrepository.com

[30] E. T. Aaron Clauset and M. Sainz, "The colorado index of complex networks," 2016. [Online]. Available: https://icon.colorado.edu/

[31] N. Kasthuri and J. Lichtman, "Neurodata's graph database," 2008. [Online]. Available: http://openconnecto.me/graph-services/download/

[32] J. Kunegis, "KONECT – The Koblenz Network Collection," in *Proc. Int. Conf. on World Wide Web Companion*, 2013, pp. 1343–1350. [Online]. Available: http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-koblenz-network-collection.pdf

[33] J. G. Diego Vázquez and R. Naik, "Interaction web database," 2003. [Online]. Available: https://www.nceas.ucsb.edu/interactionweb/resources.html

[34] V. Filkov, Z. M. Saul, S. Roy, R. M. D'Souza, and P. T. Devanbu, "Modeling and verifying a broad array of network properties," *EPL (Europhysics Letters)*, vol. 86, no. 2, p. 28003, 2009.

[35] F. Grando and L. C. Lamb, "Estimating complex networks centrality via neural networks and machine learning," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[36] Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), "Dimacs challenge." [Online]. Available: http://dimacs.rutgers.edu/Challenges/

[37] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, p. 1, 2011. [Online]. Available: https://sparse.tamu.edu/

[38] S. Bornholdt and H. G. Schuster, *Handbook of graphs and networks: from the genome to the internet*. John Wiley & Sons, 2006.

[39] R. Cohen and S. Havlin, "Scale-free networks are ultrasmall," *Physical review letters*, vol. 90, no. 5, p. 058701, 2003.

[40] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.

[41] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[42] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[43] D. Williams, *Probability with martingales*. Cambridge university press, 1991.

[44] R. Molontay, "Fractal characterization of complex networks," Master's thesis, Department of Stochastics, Budapest University of Technology and Economics, 2015.

[45] L. Pietronero and E. Tosatti, *Fractals in physics*. Elsevier, 2012.

[46] D. K. Ludlow, "Fractals in chemistry by walter g. rothschild (wayne state university). wiley publishers: New york. 1998. 206 pp. 69.95. isbn 0-471-17968-x." 2000.

[47] F. S. Labini and A. Gabrielli, "Fractals in cosmology," in *CONFERENCE SERIES-INSTITUTE OF PHYSICS*, vol. 173. Philadelphia; Institute of Physics; 1999, 2003, pp. 305–308.

[48] M. F. Osborne, "Brownian motion in the stock market," *Operations research*, vol. 7, no. 2, pp. 145–173, 1959.

[49] A. Kelley, "Layering techniques in fractal art," *Computers & Graphics*, vol. 24, no. 4, pp. 611–616, 2000.

[50] B. B. Mandelbrot, *Fractals*. Wiley Online Library, 1977.

[51] M. W. Palmer, "Fractal geometry: a tool for describing spatial patterns of plant communities," *Vegetatio*, vol. 75, no. 1-2, pp. 91–102, 1988.

[52] B. Mandelbrot, "How long is the coast of britain? statistical self-similarity and fractional dimension," *science*, vol. 156, no. 3775, pp. 636–638, 1967.

[53] B. B. Mandelbrot, *The fractal geometry of nature*. WH freeman New York, 1983, vol. 173.

[54] R. F. Voss, "Fractals in nature: from characterization to simulation," in *The science of fractal images*. Springer, 1988, pp. 21–70.

[55] R. Orbach, "Dynamics of fractal networks," *Science*, vol. 231, no. 4740, pp. 814–819, 1986.

[56] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse, "How to calculate the fractal dimension of a complex network: the box covering algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 03, p. P03006, 2007.

[57] H. D. Rozenfeld, L. K. Gallos, C. Song, and H. A. Makse, "Fractal and transfractal scale-free networks," in *Encyclopedia of Complexity and Systems Science*. Springer, 2009, pp. 3924–3943.

[58] H. D. Rozenfeld, C. Song, and H. A. Makse, "Small-world to fractal transition in complex networks: a renormalization group approach," *Physical review letters*, vol. 104, no. 2, p. 025701, 2010.

[59] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.

[60] S. R. de la Torre, J. Kalda, R. Kitt, and J. Engelbrecht, "Fractal and multifractal analysis of complex networks: Estonian network of payments," *The European Physical Journal B*, vol. 90, no. 12, p. 234, 2017.

[61] Z.-J. Zeng, C. Xie, X.-G. Yan, J. Hu, and Z. Mao, "Are stock market networks non-fractal? evidence from new york stock exchange," *Finance Research Letters*, vol. 17, pp. 97–102, 2016.

[62] W.-X. Zhou, Z.-Q. Jiang, and D. Sornette, "Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling," *Physica A: Statistical Mechanics and its Applications*, vol. 375, no. 2, pp. 741–752, 2007.

[63] C. Yuan, Z. Zhao, R. Li, M. Li, and H. Zhang, "The emergence of scaling law, fractal patterns and small-world in wireless networks," *IEEE Access*, vol. 5, pp. 3121–3130, 2017.

[64] S. Deng, W. Li, J. Gu, Y. Zhu, L. Zhao, and J. Han, "Measuring fractal dimension of metro systems," in *Journal of Physics: Conference Series*, vol. 604, no. 1. IOP Publishing, 2015, p. 012005.

[65] Y. Deng, W. Zheng, and Q. Pan, "Performance evaluation of fractal dimension method based on box-covering algorithm in complex network," in *Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on.* IEEE, 2016, pp. 682–686.

[66] P. Mattila, *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability.* Cambridge university press, 1999, vol. 44.

[67] M. Barnsley, "Fractals everywhere (new york: Academic)," 1988.

[68] D. Wei, B. Wei, Y. Hu, H. Zhang, and Y. Deng, "A new information dimension of complex networks," *Physics Letters A*, vol. 378, no. 16-17, pp. 1091–1094, 2014.

[69] A. Rényi, "Dimension, entropy and information," in *Trans. 2nd Prague Conf. Information Theory*, 1960, pp. 545–556.

[70] H. Zhang, Y. Hu, X. Lan, S. Mahadevan, and Y. Deng, "Fuzzy fractal dimension of complex networks," *Applied Soft Computing*, vol. 25, pp. 514–518, 2014.

[71] D. R. Cox, *Inference and asymptotics.* Routledge, 2017.

[72] P. Hall, "On some simple estimates of an exponent of regular variation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 37–42, 1982.

[73] D. M. Mason, "Laws of large numbers for sums of extreme values," *The Annals of Probability*, pp. 754–764, 1982.

[74] B. M. Hill *et al.*, "A simple general approach to inference about the tail of a distribution," *The annals of statistics*, vol. 3, no. 5, pp. 1163–1174, 1975.

[75] A. Clauset, M. Young, and K. S. Gleditsch, "On the frequency of severe terrorist events," *Journal of Conflict Resolution*, vol. 51, no. 1, pp. 58–87, 2007.

[76] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[77] J. Komjáthy and K. Simon, "Generating hierarchial scale-free graphs from fractals," *Chaos, Solitons & Fractals*, vol. 44, no. 8, pp. 651–666, 2011.

[78] R. Molontay, *Networks and fractals*. Bachelor's Thesis, Department of Stochastics, Budapest University of Technology and Economics, 2013.

[79] G. Csányi and B. Szendrői, "Fractal–small-world dichotomy in real-world networks," *Physical Review E*, vol. 70, no. 1, p. 016122, 2004.

[80] P. Wlczek, A. Odgaard, and M. Sernetz, "Fractal 3d analysis of blood vessels and bones," in *Fractal Geometry and Computer Graphics*. Springer, 1992, pp. 240–248.

[81] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády *et al.*, "The degree sequence of a scale-free random graph process," *Random Structures & Algorithms*, vol. 18, no. 3, pp. 279–290, 2001.

[82] Z. Katona and T. F. Móri, "A new class of scale free random graphs," *Statistics & probability letters*, vol. 76, no. 15, pp. 1587–1593, 2006.

[83] A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1-2, pp. 173–187, 1999.

[84] J. Neveu, *Discrete-parameter martingales*. Elsevier, 1975, vol. 10.

[85] J. P. Canning, E. E. Ingram, S. Nowak-Wolff, A. M. Ortiz, N. K. Ahmed, R. A. Rossi, K. R. Schmitt, and S. Soundarajan, "Network classification and categorization," *arXiv preprint arXiv:1709.04481*, 2017.

[86] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 985–1042, 2010.

[87] M. E. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Physical review E*, vol. 64, no. 1, p. 016131, 2001.

[88] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, p. 186, 2009.

[89] L. Deuker, E. T. Bullmore, M. Smith, S. Christensen, P. J. Nathan, B. Rockstroh, and D. S. Bassett, "Reproducibility of graph metrics of human brain functional networks," *Neuroimage*, vol. 47, no. 4, pp. 1460–1468, 2009.

[90] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. Suppl 2, pp. S186–S188, 2003.

[91] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.69.026113

[92] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[93] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[94] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017, pp. 314–315.

[95] N. Chaikla and Y. Qi, "Genetic algorithms in feature selection," in *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 538–540.

[96] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.