The PageRank algorithm

See Jeremy Kun's PageRank blogs:

`jeremykun.com/2011/06/12/googles-pagerank-introduction`
`...-a-first-attempt`
`...-the-final-product`

A search engine $\begin{cases} \text{finds the pages relevant to the given keywords} \\ \text{lists the most "important" hits} \end{cases}$

Which pages are important? We need a way to rank all web pages on the internet.

First idea: Important are those that many links point to.

(But: it should matter what kind of pages point to it.)

Better: Important are those that many important pages point to.

$v_1, v_2, \ldots$ the pages
$x_1, x_2, \ldots,$ their ranks

$$\left(x_i = \sum_{j \to i} x_j, \quad \text{where } j \to i \text{ is a link from } v_j \text{ to } v_i \right)$$

(But: here a page with many links gains more influence than its due)

Even better:

$$x_i = \sum_j a_{ij} x_j, \quad \text{where} \quad a_{ij} = \frac{\#\, links\ j \to i}{\#\, links\ j \to}$$

that is, every page gives the proportionate part of its rank to its links.

This way the rank vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}$ is "defined with itself".

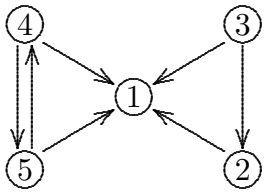
We want to find a (nonzero) solution for this homogeneous system of linear equations.

If $A = [a_{ij}]_{i,j}$ is the link matrix, $\mathbf{x}$ the rank vector, then

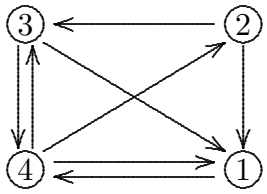
$$\mathbf{x} = A\mathbf{x}$$

- Is there a nontrivial solution?
- Is the solution unique up to scalar multiples?

Though we are talking about huge matrices (there are close to 2 billion web pages on the internet), to understand how this ranking works, let's look at two very small examples: a graph of 5 and 4 web pages, where arrows represent the links.



G_1



G_2

In G_1 there is no link from v_1 (v_1 is a dangling node), so the matrix A above is not defined. However, it is intuitively clear that v_1 should have the highest rank.

Exercise: Find the link matrix A for G_2, and find a rank vector $\mathbf{x}$ if possible.

Solution:

$$A = \begin{bmatrix} 0 & 1/2 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1/3 \\ 0 & 1/2 & 0 & 1/3 \\ 1 & 0 & 1/2 & 0 \end{bmatrix} \qquad \begin{array}{rcl} A\mathbf{x} & = & \mathbf{x} \\ & \Updownarrow & \\ (A-I)\mathbf{x} & = & \mathbf{0} \end{array}$$

$$\left[\begin{array}{cccc|c} -1 & 1/2 & 1/2 & 1/3 & 0 \\ 0 & -1 & 0 & 1/3 & 0 \\ 0 & 1/2 & -1 & 1/3 & 0 \\ 1 & 0 & 1/2 & -1 & 0 \end{array}\right] \mapsto\mapsto\mapsto \left[\begin{array}{cccc|c} 1 & 0 & 0 & -3/4 & 0 \\ 0 & 1 & 0 & -1/3 & 0 \\ 0 & 0 & 1 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \Rightarrow \mathbf{x} = \begin{bmatrix} 3/4 \\ 1/3 \\ 1/2 \\ 1 \end{bmatrix}$$

So v_4 has the highest rank though it has only two incoming edge, while v_1 has three. But: the only link from v_1 goes to v_4, transferring all its importance to v_4, and v_4 has an extra "vote" from v_3, as well.

Do we always find a solution?

Theorem
If the graph has no dangling node (a node with no outgoing edges) then $A\mathbf{x} = \mathbf{x}$ has nontrivial solutions.

Proof: $\mathbf{x} \neq \mathbf{0}$, $A\mathbf{x} = \mathbf{x}$ means that $\mathbf{x}$ is an eigenvector of A with eigenvalue 1. Since every column of A adds up to 1, every row of A^T adds up to 1, that is,

$$A^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \Rightarrow$$

1 is an eigenvalue of A^T, so 1 is an eigenvalue of A, as well.
$\big($ $\det(A^T - I) = \det(A - I)^T = \det(A - I)$. $\big)$

Definition
Let $A \in \mathbb{R}^{m\times n}$ be a real matrix.
$\mathbf{A} \geq \mathbf{0}$ if $a_{ij} \geq 0$ for all i, j (**nonnegative matrix**).
$\mathbf{A} > \mathbf{0}$ if $a_{ij} > 0$ for all i, j (**positive matrix**).
A is a **stochastic matrix** if $A \geq 0$, and every column of A adds up to 1.
A vector $\mathbf{x}$ is nonnegative/positive/stochastic if it has this property as an $n \times 1$ matrix.

Notation $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.
$J = J_n \in \mathbb{R}^{n\times n}$ is the matrix whose each entry is 1.

Note that a nonnegative matrix $A \in \mathbb{R}^{n\times m}$ is stochastic $\Leftrightarrow \mathbf{1}^T A = \mathbf{1}^T$.

Remark: The link matrix in the previous theorem is stochastic, and the proof can be applied to any stochastic matrix, so if $A \in \mathbb{R}^{n\times n}$ is stochastic then the equation $\mathbf{x} = A\mathbf{x}$ has a nontrivial solution.

Another approach leading to the same ranking

Random surfing:
— Start at a random page.
— Click randomly on one of the links in this page.
— Continue clicking at random links, wherever you arrive.

If the probability of getting at page v_i at the t'th step is $x_i^{(t)}$, then:

$$x_i^{(0)} = \frac{1}{n} \qquad n \text{ is the number of web pages}$$
$$x_i^{(t+1)} = \sum a_{ij} x_j^{(t)}, \text{ i.e. } \mathbf{x}^{(t+1)} = A\mathbf{x}^{(t)} \quad \forall t,$$

provided that the surfer doesn't get to a linkless page (say, there are no dangling nodes). Here A is the link matrix defined earlier.

Let $\mathbf{x} = \lim_{t\to\infty} \mathbf{x}^{(t)}$, if it exists. Then

$$\begin{array}{ccc} \mathbf{x}^{(t+1)} & = & A\mathbf{x}^{(t)} \\ & \downarrow & \\ \mathbf{x} & = & A\mathbf{x}, \end{array}$$

so it will be the ranking we want. Thus the rank is the limit of the probability that a random surfer arrives at a given page after many clicks.

Note that A and $\mathbf{x}^{(0)}$ is stochastic, so any $\mathbf{x}^{(t)}$ is stochastic: the product of two nonnegative matrices is clearly nonegative, and $\mathbf{1}^T\mathbf{x}^{(t+1)} = \mathbf{1}^T A\mathbf{x}^{(t)} = \mathbf{1}^T\mathbf{x}^{(t)} = \ldots = \mathbf{1}^T\mathbf{x}^{(0)} = 1$. Hence its limit vector $\mathbf{x}$ is also stochastic.

— What can we do if there are dangling nodes? Pick a random URL.

— What can we do if the surfer gets to a part of the graph that has no outgoing edges to the rest of the graph?

Modify the surfing:
the surfer types in a random URL with probability p (even if he could choose a link),
he clicks at a link with probability $(1-p)$,
if there is no link then he chooses a random page.

So let A be the link matrix where we put $\frac{1}{n}\mathbf{1}$ in the columns of the dangling nodes, so A is stochastic. $\hat{A} := (1-p)A + p\frac{1}{n}J$. Then the new matrix, $\hat{A}$ is still stochastic and it is positive.

Theorem
Let $A \in \mathbb{R}^{n\times n}$ be a positive stochastic matrix. Then
(1) 1 is an eigenvalue with a one-dimensional eigenspace generated by a unique stochastic eigenvector $\mathbf{x}$;
(2) for any nonzero stochastic vector $\mathbf{v}$, $\lim_{t\to\infty} (\hat{A})^t\mathbf{v} = \mathbf{x}$.

(We shall return to this theorem later.)

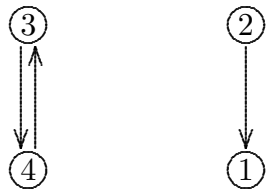
Remark: When we want to determine the eigenvector $\mathbf{x}$, the Gaussian elimination cannot be done for very large matrices. Instead, we can approximate the eigenvector with $(\hat{A})^t\frac{1}{n}\mathbf{1}$: if A is sparse (has relatively few nonzero entries), then $\mathbf{y} \mapsto A\mathbf{y}$ can be calculated easily, and even with the positive matrix $\hat{A}$:

$$\hat{A}\mathbf{y} = (1-p)A\mathbf{y} + p\frac{1}{n}J\mathbf{y} = (1-p)A\mathbf{y} + p\frac{1}{n}\mathbf{1},$$

if $\mathbf{y}$ is stochastic.

Exercise: Calculate the rank vector for G_1 if we modify the link matrix so that the first column is $\frac{1}{5}\mathbf{1}$ (i.e. we add an arrow from v_1 to every vertex, including itself).

Exercise: Take the graph G_3:



G_3

Calculate the rank vector from the corresponding modified link matrix $\hat{A}$, when we choose p to be $\frac{1}{4}$ or $\frac{1}{2}$, respectively.