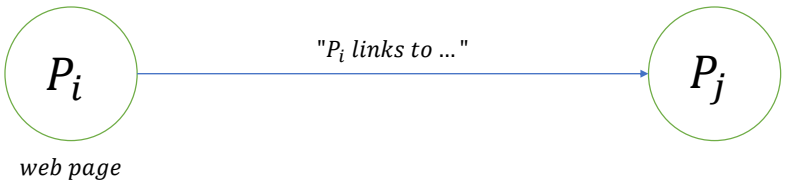
Intelligent Data Analysis
Lecture Notes on
PageRank

Peter Tiňo

1 "Mining the Web"

We will now look at how to quantify the 'significance' of documents/websites on the web. In previous lectures we were able to represent documents as vectors and analyse how similar they are against a set of terms in our query. The problem comes when we want to find documents on the massive World Wide Web; there are likely to be many documents that match our query rather well! We need to decide on some form of ranking of the query results, hopefully giving the most 'interesting' results the priority. How can this be achieved?

The web is a huge interconnection structure among pages that can be represented as a graph: the web pages correspond to 'nodes' and the hyperlinks linking these pages become 'edges' between the nodes. We can exploit the 'democratic' nature of the web in order to assign a 'significance factor' to each page based on how it is embedded into the web, meanwhile neglecting its actual content. The rough idea is that if a given page has many other pages pointing to it, its content is probably worthwhile.



2 The PageRank equation

For a web page P_i, let $x_i \geq 0$ be its 'authority' that somehow quantifies its 'importance' or 'significance'. The authority x_i needs to take into account:

i) The number of incoming links or citations - more links to P_i should indicate higher significance,

ii) The authority of the pages that cite the page in question - the web page will 'look good' if more authoritative pages cite it rather than a lot of 'weaker' pages.

We denote by $Pa(P_i)$ the set of parents of P_i (the set of pages that link (point) to P_i), and by h_j the number of *outgoing* links from page P_j. These ingredients are used to form the PageRank equation, derived by Google founders Larry Page and Sergey Brin while they were still students at Stanford in 1998.

Each page $P_j \in Pa(P_i)$ pointing to page P_i increases the authority x_i by x_j/h_j. This is because only a fraction of total authority x_j flows to P_i - the authority x_j of page P_j is evenly distributed among all outgoing links from P_j. So the authority collected by P_i from the web is

$$\tilde{x}_i = \sum_{P_j \in Pa(P_i)} \frac{x_j}{h_j}.$$

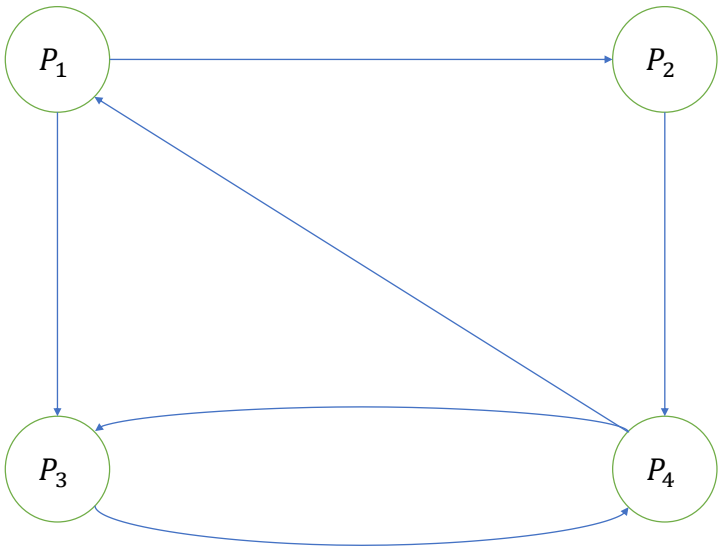
Now imagine that we would like all pages to have at least some default authority. This would give some minimal visibility to weak or new pages on the web. Let us construct the total authority of P_i from two ingredients: (1) an authority of 1 that is given by default to all pages and (2) the authority $\tilde{x}_i$ earned by P_i through citations from the web. The authority of P_i is then obtained as a convex combination of (1) and (2):

$$x_i = (1-d) \cdot 1 + d \cdot \tilde{x}_i,$$

where the mixing weights are determined by the 'discount factor' $0 < d < 1$. We can finally formulate PageRank equations (Brin & Page, 1998) for determining authorities of N pages on the web:

$$x_i = (1-d) + d \sum_{P_j \in Pa(P_i)} \frac{x_j}{h_j}, \quad i = 1, 2, ..., N. \tag{1}$$

We will apply PageRank to a small example. Consider a web with pages P_1, P_2, P_3 and P_4. They link to each other according to the following diagram:



We calculate the authority of each page as follows:

$$\begin{aligned} x_1 &= (1-d) + d\left(\frac{x_4}{2}\right) \\ x_2 &= (1-d) + d\left(\frac{x_1}{2}\right) \\ x_3 &= (1-d) + d\left(\frac{x_1}{2} + \frac{x_4}{2}\right) \\ x_4 &= (1-d) + d\,(x_2 + x_3) \end{aligned}$$

Notice that to calculate the authority of a page it is not enough to just use PageRank; solving for each x_i requires us to solve a system of N linear equations of N unknowns x_i, $i = 1, 2...., N$ (there are N pages on the web) given as the matrix equation

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = (1-d) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + d \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,N} \\ w_{2,1} & w_{2,2} & \dots & w_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N,1} & w_{N,2} & \dots & w_{N,N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \tag{2}$$

written in a compact form as

$$\mathbf{x} = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x}.$$

Elements of the coupling matrix $\mathbf{W}$ are

$$w_{i,j} = \begin{cases} \frac{1}{h_j} & \text{, if there is a link from } P_j \text{ to } P_i \\ 0 & \text{, otherwise.} \end{cases}$$

This system of equations can be solved using standard linear algebra as follows ($\mathbf{I}_N$ is the $N \times N$ identity matrix) :

$$\begin{aligned} & \mathbf{x} = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x} \\ \Leftrightarrow \quad & (\mathbf{I}_N - d \cdot \mathbf{W})\mathbf{x} = (1-d) \cdot \mathbf{1}_N \\ \Leftrightarrow \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ \Leftrightarrow \quad & \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \end{aligned}$$

where $\mathbf{A} = \mathbf{I}_N - d \cdot \mathbf{W}$ and $\mathbf{b} = (1-d) \cdot \mathbf{1}_N$ and we assume $\mathbf{A}$ is invertible[1].

We see that finding the authority of each page involves calculating an inverse matrix. However, this is a major problem since computing $\mathbf{A}^{-1}$ for very large N is infeasible. To work around this issue, we will need a bit of knowledge on dynamical systems.

3 Contraction mappings

For a vector $\mathbf{x}$, define $F(\mathbf{x})$ to be an *affine mapping* that takes a vector input and outputs another vector. Such mappings are of the form

$$F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{b}$ is a 'shift' vector and $\mathbf{A}$ is a matrix defining a linear transformation. In the case of PageRank,

$$F(\mathbf{x}) = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x},$$

so we see that $d \cdot \mathbf{W}$ maps $\mathbf{x}$ to another vector while $(1-d) \cdot \mathbf{1}_N$ further shifts this mapped vector.

We are interested in finding special vectors $\mathbf{x}_*$ where applying F on them does not do anything at all i.e. $F(\mathbf{x}_*) = \mathbf{x}_*$. We say that $\mathbf{x}_*$ is a *fixed point* of the mapping F. One way to find these points is by using a iterative method; suppose that $\mathbf{x}(0)$ is a random initial vector we start with at time $t = 0$. Using this vector, obtain $\mathbf{x}(1) = F(\mathbf{x}(0))$, the vector we get at time $t = 1$. Under some assumptions on F, we can continue with this process until we find a fixed point:

[1] If $\mathbf{A}$ is not invertible, or is ill-conditioned, we would need to resort to pseudoinverse, or SVD. This is beyond the scope of this module.

$$\begin{aligned} \mathbf{x}(1) &= F(\mathbf{x}(0)) \\ \mathbf{x}(2) &= F(\mathbf{x}(1)) \\ &\vdots \\ \mathbf{x}(t+1) &= F(\mathbf{x}(t)) \\ &\vdots \\ \mathbf{x}_* &= F(\mathbf{x}_*) \end{aligned} \tag{3}$$

But what exactly are the properties of F such that its fixed point $\mathbf{x}_*$ can be found by simple repeated applications of F from arbitrary initial condition $\mathbf{x}(0)$ and, in addition, this fixed point is guaranteed to exist and is unique? We will show that it is sufficient to demand that F "shrinks distances" as it maps vectors $\mathbf{x}$ to their images $F(\mathbf{x})$.

Definition 3.1. The mapping F is a *contraction mapping* if for any two points $\mathbf{x}_1$ and $\mathbf{x}_2$,

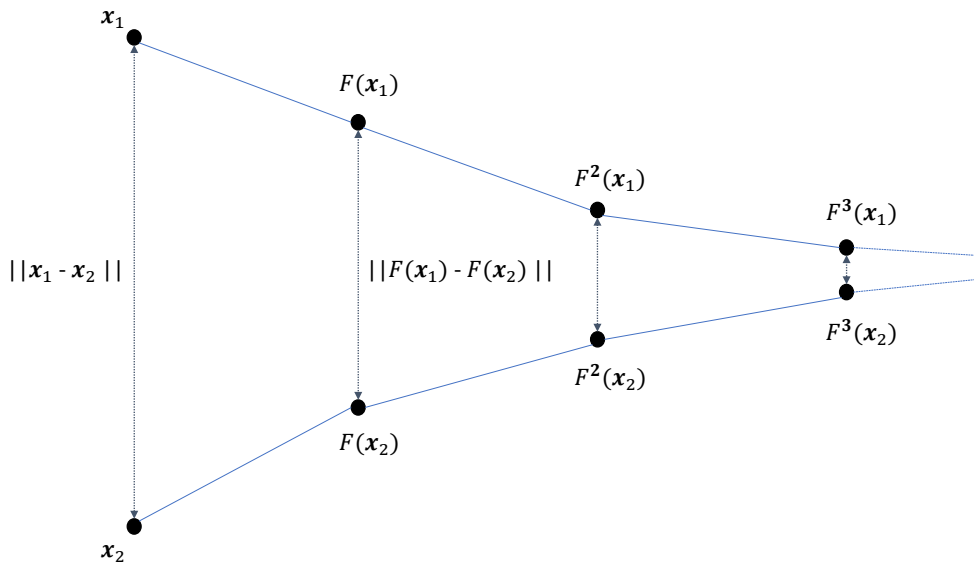
$$||F(\mathbf{x}_1) - F(\mathbf{x}_2)|| \leq \rho||\mathbf{x}_1 - \mathbf{x}_2||, \tag{4}$$

where $\rho \in (0, 1)$ (*Lipschitz constant*).

When we apply F to $\mathbf{x}_1$ and $\mathbf{x}_2$, they get squeezed closer together by a factor of at least ρ in their original distance. Notice that, if $F^n(\mathbf{x})$ is a mapping of $\mathbf{x}$ under F n times, then

$$\begin{aligned} ||F^n(\mathbf{x}_1) - F^n(\mathbf{x}_2)|| &\leq \rho||F^{(n-1)}(\mathbf{x}_1) - F^{(n-1)}(\mathbf{x}_2)|| \\ &\leq \rho^2||F^{(n-2)}(\mathbf{x}_1) - F^{(n-2)}(\mathbf{x}_2)|| \\ &\leq \\ &\vdots \\ &\leq \rho^n||\mathbf{x}_1 - \mathbf{x}_2||. \end{aligned}$$

We see that the distances decay exponentially fast since $0 < \rho < 1$. The following figure illustrates this observation, showing how these two vectors move closer together every time F is applied.



Now we address the following question: *If I keep applying a contraction mapping F, starting from any vector $\boldsymbol{x}$, will the images converge to a point?* Observe that from contractivity of F, we have:

$$||F^{n-1}(\mathbf{x}) - F^n(\mathbf{x})|| \leq \rho||F^{n-2}(\mathbf{x}) - F^{n-1}(\mathbf{x})||.$$

This says that the distances between subsequent images of $\mathbf{x}$ under F get smaller and smaller, converging (in a complete metric space) to a point that does not move at all under the action of F, which is our fixed point.

To show that contractive mappings cannot have two distinct fixed points, suppose $\mathbf{x}$ and $\mathbf{y}$ are fixed points of F, i.e. $F(\mathbf{x}) = \mathbf{x}$ and $F(\mathbf{y}) = \mathbf{y}$. Note that in this case,

$$||F(\mathbf{x}) - F(\mathbf{y})|| \leq \rho||\mathbf{x} - \mathbf{y}|| \Leftrightarrow ||\mathbf{x} - \mathbf{y}|| \leq \rho||\mathbf{x} - \mathbf{y}||.$$

But this inequality is possible only when $\mathbf{x} = \mathbf{y}$! Therefore, we've shown that contraction mappings always exhibit a single *unique* fixed point.

Lastly, we still need to answer one very important question: *Is the PageRank mapping*

$$F(\mathbf{x}) = (1 - d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x}$$

a contraction? The answer is *yes*. To see this, consider any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^N$. We have,

$$\begin{aligned} ||F(\mathbf{x}_1) - F(\mathbf{x}_2)|| &= ||(1 - d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x}_1 - (1 - d) \cdot \mathbf{1}_N - d \cdot \mathbf{W}\mathbf{x}_2|| \\ &= d||\mathbf{W}\mathbf{x}_1 - \mathbf{W}\mathbf{x}_2|| \\ &= d||\mathbf{W}(\mathbf{x}_1 - \mathbf{x}_2)|| \\ &< d||\mathbf{x}_1 - \mathbf{x}_2||. \end{aligned}$$

The last inequality holds because we are multiplying $(\mathbf{x}_1 - \mathbf{x}_2)$ by the page coupling matrix $\mathbf{W}$, where absolute majority of the entries are 0 (the page interconnection structure is very sparse) and non-zero entries are usually fractions < 1. Maximum entry value is 1, if a page points to a single page.

We have thus shown that to solve the PageRank equation

$$\mathbf{x} = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x},$$

we can simply start with an arbitrary assignment of authorities to pages, $\mathbf{x}(0)$, and iterate the PageRank equation few times

$$\mathbf{x}(t+1) = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x}(t).$$

This process will approach the solution

$$\mathbf{x}_* = (1-d) \cdot \mathbf{1}_N + d \cdot \mathbf{W}\mathbf{x}_*,$$

rather quickly, since the distances between the iterates $\mathbf{x}(t)$ decay exponentially fast.

4 Authority of web page communities

We will now investigate what authority PageRank assigns to a *collection of web pages* G, which we will call a *community* of web pages. We will categorise out web pages into three different sets:

i) $Into(G)$ - This set contains the set of pages outside G that have a link pointing to any page in G,

ii) $Out(G)$ - The set of pages from within G that link to pages outside G,

iii) $dp(G)$ - The set of *dangling pages* within G. These are pages without any outgoing link, i.e. pages that that do not link to any other page.

The authority of a community G is accumulated through authorities of individual pages inside it:

$$E_G = \sum_{p \in G} x_p. \tag{5}$$

E_G thus represents the *visibility* or *energy* of G.

Figure 1 provides an illustration of a community of web pages (the blue nodes), with one dangling page present (the grey node) and a number of external pages that link to G (the green nodes). Of course, the more authority our pages in G have, the more visible G is on the web. However, designers of communities of web pages need to address the following question: *How do we maximise* (5) *if we have no control over how web pages on the web link to* G*?*

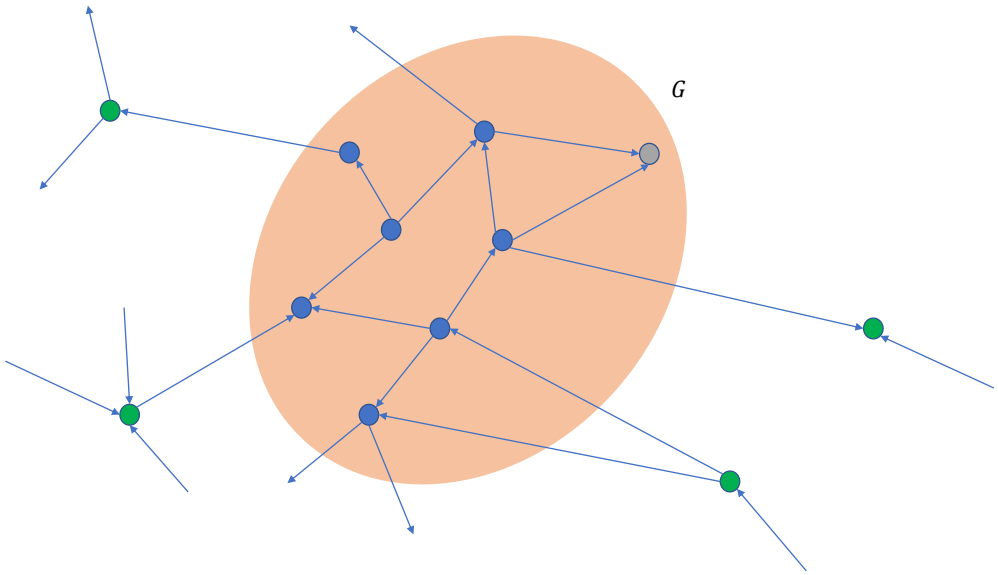


Figure 1: An example of a community of web pages denoted by G. The grey node is an example of a dangling page as it has no 'out' links, and the green nodes outside G point to or receive a link from a page in the community, affecting the overall 'energy' or 'visibility' it has.

Some researchers have investigated this matter (Bianchini, Gori & Scarselli, 2005), and found that E_G can be decomposed based on our sets $Into(G)$, $Out(G)$ and $dp(G)$ as follows:

$$E_G = |G| + E_G^{Into} - E_G^{Out} - E_G^{dp}, \tag{6}$$

where $|G|$ is the size of G (number of pages in the community); E_G^{Into} is the energy flowing to G from pages outside G that point to it; E_G^{Out} is the energy that leaks out from G through links from pages in G pointing outside it; finally, E_G^{dp} is the energy lost due to dangling pages.

Before we write expressions for the decomposition terms in (6), note that the entity we are interested in now is not a single page, but a set of pages G. Recall that if a page p points to another page q, only the fraction x_p/h_p of authority x_p flows to q, since the authority x_p is equally split among all outgoing links from p. In our setting, if a page p points to G from outside G, it can point to one or several pages in G. Hence to establish what fraction of x_p flows to G, we must calculate the fraction ρ_p of outgoing links from p that point to G. The authority flowing from p to G is then $\rho_p x_p$.

Consider now a page p inside the community G. The fraction of its outgoing links that stay inside G is (of course) ρ_p, meaning that the fraction of its outgoing links that point outside G is $1 - \rho_p$. Hence, the fraction of authority x_p leaking out of G is $(1 - \rho_p)x_p$. We are now ready to express the decomposition terms in (6):

$$
\begin{aligned}
E_G^{Into} &= \frac{d}{1-d} \sum_{p \in Into(G)} x_p \rho_p, \\
E_G^{Out} &= \frac{d}{1-d} \sum_{p \in Out(G)} x_p (1-\rho_p), \\
E_G^{dp} &= \frac{d}{1-d} \sum_{p \in dp(G)} x_p.
\end{aligned} \tag{7}
$$

From (7), we can learn useful lessons about how to structure our community of web pages to ensure we maximise its energy:

i) Whenever possible, avoid creating dangling pages within the community. At least put a link to the main page of the community.

ii) Care about referencing pages outside the community; linking to other pages from your main page may not be a good idea, as more authority flows out.