Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

A Probabilistic Model for Webpage Activity and Its Effect on PageRank

Report for the
Delft Institute of Applied Mathematics
as part of

the degree of

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

LEONARD HUIJZER

Delft, The Netherlands
June 2016

BSc report APPLIED MATHEMATICS

"A Probabilistic Model for Webpage Activity and Its Effect on PageRank"

LEONARD HUIJZER

Delft University of Technology

Thesis Adviser

Dr. N.V. Budko

Other members of the graduation committee

Drs. E.M. van Elderen

Dr. D.C. Gijswijt

June, 2016

Delft

Contents

1 Introduction

In 1998 Google's founders Sergey Brin and Larry Page published a paper in which they presented a mathematical method they developed to rank webpages according to importance.[1] The ranking, called PageRank, they proposed uses only the link structure of the internet to measure the importance of webpages. When a user enters a search query into Google, within milliseconds Google returns the search results. These results are not only based on the relevance to the search query, but also on the importance of a webpage. The importance of a webpage is measured by the number of recommendations it has from other important webpages, a recommendation being a hyperlink. This – seemingly circular – definition can be stated explicitly in mathematical terms. The PageRank model will be introduced, it will be shown that the PageRank vector – containing the PageRank value for each known webpage – is an eigenvector of the Google matrix.[2]

A lot of research has been done on several aspects of the PageRank model. The sensitivity of the PageRank vector to change in the underlying network structure has been a popular topic in recent years, see for example [4], [5] and [12]. These papers mainly study the results of some deterministic change on PageRank. E.g., the effect of rank one updates of the hyperlink matrix on the new PageRank vector. A lot of this research focused on providing upper bounds for change in the PageRank vector, given some specific sort of change in the underlying network. On the other hand, the evolution of the structure of the internet as a whole has seen a lot of attention as well. See for example, [6], [10] and [11]. These papers focus mainly on the effects of random change in networks. E.g., what can be said about the connectivity of the internet in case of a random breakdown.

In this report things will be considered from a different perspective. Change will not be thought of as a given, as something that just happens, but it will be investigated why there would be change, what sort of change happens in a network and how this would affect PageRank. Mainly numerical methods will be used, as the application of analytical methods is limited.

First, in chapter 3, the PageRank algorithm will be introduced. In chapter 4 a model will be introduced in which network change will be modelled in terms of webpage owner/administrator activity. A basic model will be proposed in section 4.1, in which the owner/administrator activity will be assumed to be equal for every webpage. The results of this basic model will be shown in section 4.2. In sections 4.3-4.6, this model will be adjusted so as to take into account that some webpages are more active than others. In the objective extended model of section 4.3 page owners/administrators are assumed to be rational in the sense that they will make decisions based on the PageRank of other pages. The subjective extended model of section 4.5 will simulate webpage activity by owners/administrators making decisions based on the number of links that point to other webpages. The behaviour of the models can be investigated using numerical methods. First the general behaviour of the models will be shown using numerical methods, in sections 4.4 and 4.6. In section 4.7, based on both extended models, some tests will be devised to measure the influence of page owner/administrator activity on the PageRank of the webpage. The results indicate the optimal strategy to be used by owners/administrators for maximizing the PageRank value of a webpage.

[1] See [1].

[2] Note that the current PageRank model used by Google is still based on the original proposal, however it is almost certain that Google uses a more advanced model, hence the PageRank model as discussed in this report may not be a completely up-to-date depiction of the algorithm.

2 Preliminaries

Definition 1. Let $\mathbf{x} \in \mathbb{R}^n$. The **1-norm** on $\mathbb{R}^n$ is defined as

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|$$

Definition 2. Let $\|\cdot\|_1$ be the 1-norm on $\mathbb{R}^n$ and let $\mathbf{M}$ be an $n \times n$ matrix. Then the **induced matrix norm** corresponding to $\|\cdot\|_1$ is defined as

$$\|\mathbf{M}\|_1 := \max_{1 \leq j \leq n} \sum_{i=1}^{n} |M_{ij}|,$$

the largest absolute column sum.

Definition 3. An $n \times n$ matrix $\mathbf{M}$ is called **row substochastic** if:

1. $M_{ij} \geq 0$ for all $1 \leq i \leq n$ and $1 \leq j \leq n$
2. $\sum_{i=1}^{n} M_{ij} \leq 1$ for all $1 \leq i \leq n$.

Definition 4. An $n \times n$ matrix $\mathbf{M}$ is called **row stochastic** if:

1. $M_{ij} \geq 0$ for all $1 \leq i \leq n$ and $1 \leq j \leq n$
2. $\sum_{i=1}^{n} M_{ij} = 1$ for all $1 \leq i \leq n$.

Definition 5. An $n \times n$ matrix $\mathbf{M}$ is called **irreducible** if for each element M_{ij} of $\mathbf{M}$ there exists an $n \in \mathbb{N}_{>0}$ such that $(\mathbf{M}^n)_{ij} > 0$. $\mathbf{M}$ is called **reducible** if it is not irreducible.

Definition 6. An $n \times n$ matrix $\mathbf{M}$ is called **aperiodic** if for all $1 \leq i \leq n$:

$$\gcd\{k \in \mathbb{N}_{>0} : (\mathbf{M}^k)_{ii} > 0\} = 1$$

Moreover, if for every $1 \leq i \leq n$ it holds that $\mathbf{M}_{ii} > 0$, then $\mathbf{M}$ is aperiodic. $\mathbf{M}$ is called **periodic** if it is not aperiodic.

Definition 7. An $n \times n$ matrix $\mathbf{M}$ is called **primitive** if and only if $\mathbf{M}$ is aperiodic and irreducible. $\mathbf{M}$ is called **imprimitive** if it is not primitive.

Lemma 1. Let $\mathbf{M}$ be a stochastic $n \times n$ matrix. Then 1 is an eigenvalue of $\mathbf{M}$, and moreover, for each eigenvalue μ it holds that $|\mu| \leq 1$, i.e. 1 is the largest eigenvalue of $\mathbf{M}$.

Lemma 2. Let $\mathbf{M}$ be an $n \times n$ matrix and let λ be an eigenvalue of $\mathbf{M}$. Then λ is also an eigenvalue of $\mathbf{M}^T$.

Theorem 1. (Perron-Frobenius) If the $n \times n$ matrix $\mathbf{M} \geq \mathbf{0}$ is irreducible then the following properties hold.

1. $\mathbf{M}$ has a simple maximum eigenvalue $\lambda > 0$ (λ has an algebraic multiplicity of one).
2. There exists an eigenvector $\mathbf{x} > \mathbf{0}$ such that $\mathbf{M}\mathbf{x} = \lambda\mathbf{x}$
3. There is a unique vector $\mathbf{p} > \mathbf{0}$ such that $\mathbf{M}\mathbf{p} = \lambda\mathbf{p}$ and $\|\mathbf{p}\|_1 = 1$.

3 The PageRank algorithm

In *The Anatomy of a Large-scale Hypertextual Web Search Engine* Sergey Brin and Larry Page made their PageRank algorithm known to the rest of the world. This algorithm uses the structure of the web to calculate the importance of every known webpage on the internet. This value is then used to rank search results for a particular search query according to importance of the webpages.

3.1 Structure of the internet

The basic idea behind PageRank is that the internet can be viewed as a directed graph. Each webpage can be seen as node in the graph. A hyperlink on webpage P_1 to webpage P_2 can be seen as a directed edge from node P_1 to P_2.[3] This hyperlink is called an **outlink** of webpage P_1 and an **inlink** of webpage P_2.

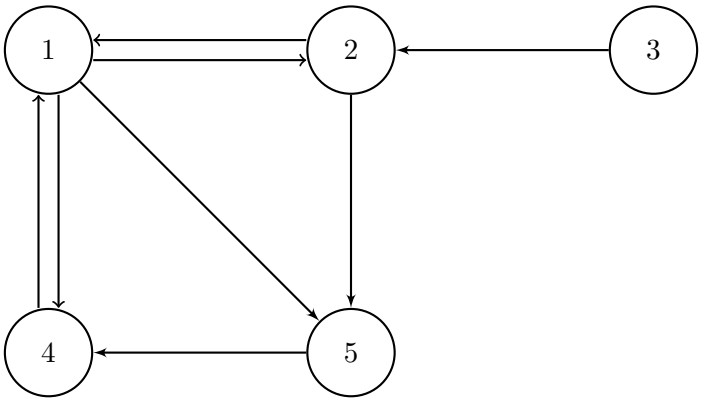


Figure 1: A network of webpages.

The Google founders viewed an inlink as a recommendation for a webpage. The more recommendations a webpage has, the more important and popular it must be. The problem with this approach is that one can easily manipulate the number of recommendations for a webpage. The status of the webpage providing the inlink should also be considered. The more outlinks a webpage has, the less should it be considered an important recommendation. When a webpage has an inlink from an important webpage, it increases the importance of the webpage being linked to. If however, the important webpage has many outlinks it will not have as much as an impact. This is the basic idea behind the PageRank model.

3.2 Importance of webpages

In the following sections it will be explained that some changes to this basic idea are necessary to make it mathematically robust. These change lead to a more robust interpretation of importance. The idea is that of a **random surfer**. Imagine a person who browses the internet by randomly clicking links from webpage to webpage. After this process is continued for an indefinite amount of time, the relative importance of the webpages is obtained. The more webpages link to a certain webpage, the more often the random surfer will visit that webpage and the more important it is. This process can be simulated by a Markov chain. It turns out that

[3]Note that only unique hyperlinks are considered. Multiple hyperlinks from one webpage to another are counted as one.

the ranking of a network is the stationary distribution of the Markov chain, corresponding to eigenvalue 1.

3.3 Ranking

A few definitions will be introduced in order to formalize the underlying basic idea of PageRank.

Definition 8. Let δ_{ij} be defined as

$$\delta_{ij} = \begin{cases} 1, & \text{if there is a hyperlink from webpage } P_i \text{ to webpage } P_j \\ 0, & \text{otherwise} \end{cases}$$

Definition 9. The **outdegree** of node P_i is defined as:

$$|P_i| = \sum_{j=1}^{n} \delta_{ij}$$

The outdegree of a node is simply the amount of unique outlinks on the webpage (multiple links to the same webpage are not counted). Using these definitions the idea of a ranking based on the link structure as explained in the previous section can be defined.

Definition 10. Let $B(P_j)$ denote the set of nodes outlinking to node P_j (also called the set of pages backlinking or inlinking to P_j) and let n be the total number of webpages. The $1 \times n$ vector $\boldsymbol{\pi}^T$ is called a **ranking** if

1. $\pi_j \geq 0$, for all $1 \leq j \leq n$.
2. $\sum_{j=1}^{n} \pi_j = 1$
3. $\pi_j = \sum_{P_i \in B(P_j)} \frac{\pi_i}{|P_i|}$

This means that a ranking must satisfy a system of linear equations. E.g., the ranking of the network in figure 1 must satisfy the linear system:

$$\begin{cases} \pi_1 = \frac{1}{2}\pi_2 + \pi_4 \\ \pi_2 = \frac{1}{3}\pi_1 + \pi_3 \\ \pi_3 = 0 \\ \pi_4 = \frac{1}{3}\pi_1 + \pi_5 \\ \pi_5 = \frac{1}{3}\pi_1 + \frac{1}{2}\pi_2 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1 \\ \pi_1, \pi_2, \pi_3, \pi_4, \pi_5 \geq 0 \end{cases}$$

The solution for this linear system is given by $\boldsymbol{\pi}^T = [6/16 \quad 2/16 \quad 0 \quad 5/16 \quad 3/16]$. In this example P_1 would be the most important webpage, followed by P_4, etc. The sixth equation ($\sum_{j=1}^{5} \pi_j = 1$) in the linear system guarantees the uniqueness of the solution. The system of linear equations exactly models the basic idea of important webpages having lots of important recommendations.

3.4 Hyperlink matrix H

A system of linear equations can be written in a more clear and compact matrix form. In order to achieve this the following matrix is introduced.

Definition 11. Let the $n \times n$ matrix $\mathbf{H}$ be defined as follows:

$$H_{ij} = \begin{cases} 1/|P_i|, & \text{if } \delta_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$\mathbf{H}$ is called the **row normalized hyperlink matrix** of the corresponding network.

The non-zero elements of row i correspond to the outlinks of P_i, similarly, the non-zero elements of column j correspond to the inlinks of P_j. More specifically, the elements of row i denote the probabilities to surf from P_i to any other page if a hyperlink on P_i were selected randomly.

Definition 12. If $|P_i| = 0$ then P_i is called a **dangling node**.

If a page P_i is a dangling node, i.e. if it does not have any outlinks, the i^{th} row of $\mathbf{H}$ contains only zero elements. Figure 2 is an example of a network with a dangling node.

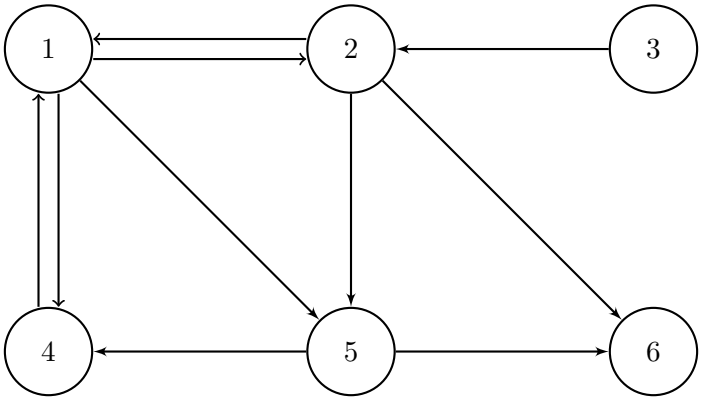


Figure 2: A network of webpages with a dangling node.

The normalized hyperlink matrix $\mathbf{H}$ corresponding to figure 2 is given as follows. Node P_6 is a dangling node, hence the 6^{th} row of $\mathbf{H}$ contains only zero elements.

$$\mathbf{H} = \begin{bmatrix} 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

If $\boldsymbol{\pi}$ is a ranking then $\mathbf{H}^T\boldsymbol{\pi} = \boldsymbol{\pi}$, since for element π_i:

$$\pi_i = \sum_{k \in B(P_i)} \frac{\pi_k}{|P_k|} = \sum_{k=1}^{n} H_{ik}^T \pi_k = (\mathbf{H}^T\boldsymbol{\pi})_i,$$

i.e. a ranking $\boldsymbol{\pi}$ is a right eigenvector of the matrix $\mathbf{H}^T$ associated with eigenvalue 1.

The hyperlink matrix $\mathbf{H}$ has the following valuable properties.

1. **H** is a sparse matrix, i.e. a large proportion of its elements are zero. This means that if a storage scheme is used in which only the non-zero elements of **H** and their locations are stored, then **H** does not require much storage space.

2. Since **H** is sparse, it does not require the $O(n^2)$ computation for matrix multiplication. In fact, the average webpage on the internet has been estimated to have about 10 outlinks, which means that **H** contains about $10n$ non-zero elements. As a consequence, matrix multiplication involving **H** requires only an $O(n)$ computation.

3. **H** is almost row stochastic. Only the rows corresponding to dangling nodes contain only zero elements. All the other rows are stochastic, hence **H** is row substochastic.

4. **H** can be periodic and/or reducible.

3.5 Stochastic matrix S

A ranking $\boldsymbol{\pi}$ satisfies the equation $\mathbf{H}^T\boldsymbol{\pi} = \boldsymbol{\pi}$. The natural question that one could ask is under what conditions for $\mathbf{H}^T$ does a stationary distribution exist? And, if it exists, under what conditions is it unique? It turns out that additional conditions on **H** are needed to make sure a unique stationary distribution exists. The Google founders modified the matrix **H** in such a way that the uniqueness and existence are guaranteed, while making sure the intended interpretation still would hold.

The first step was to ensure that **H** is row stochastic. It was noted in the previous section that **H** already is row substochastic. Whenever the random surfer ends up in a dangling node, it will get stuck. The stochasticity modification will ensure that the random surfer does not get stuck.

Definition 13. Let the $n \times 1$ vector **d** be defined as

$$d_i = \begin{cases} 1, & \text{if } |P_i| = 0 \\ 0, & \text{otherwise} \end{cases}$$

then **d** is called the **dangling node vector** corresponding to the network.

Definition 14. A $1 \times n$ vector $\mathbf{v}^T$ is called a **personalization vector** if

- $v_i > 0$ for all $1 \leq i \leq n$
- $\sum_{i=1}^{n} v_i = 1$

This vector can be used for the personalization of rankings. Each element v_i of **v** describes the chance to jump to P_i. The vector elements can be customized so as to correspond to the interests and preferences of the user. For example, if a user is interested in Bitcoins and P_5 is a webpage about Bitcoin, v_5 can be made larger than the other elements. Then the chance that the user will go to P_5 is relatively large compared to the other elements, which corresponds to the users surfing habits.

Definition 15. The row stochastic $n \times n$ matrix **S** is defined as

$$\mathbf{S} = \mathbf{H} + \mathbf{d}\mathbf{v}^T$$

This modification of $\mathbf{H}$ results in a row stochastic matrix $\mathbf{S}$. It simply replaces the zero rows of $\mathbf{H}$ corresponding to dangling nodes with the vector $\mathbf{v}^T$. If the random user ends up in a dangling node, it will jump to page P_i with probability v_i. The matrix $\mathbf{S}$ corresponding to the graph in figure 2 with $\mathbf{v} = (1/n)\mathbf{e}$ is given as follows.

$$\mathbf{S} = \begin{bmatrix} 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}$$

If the random surfer ends up in P_6 the probability that it will visit any webpage P_i next is equal for all six pages.

The matrix $\mathbf{S}$ has the following properties.

1. Since $\mathbf{S}$ is stochastic, the matrix has largest eigenvalue 1 and hence $\mathbf{S}^T$ as well. Certainly a ranking $\boldsymbol{\pi}$ such that $\mathbf{S}^T\boldsymbol{\pi} = \boldsymbol{\pi}$ then exists, however it need not be unique.
2. $\mathbf{S}$ is in general periodic as well as reducible.
3. $\mathbf{S}$ is still sparse, but much less so then $\mathbf{H}$. This is because every row corresponding with a dangling node is replaced by a row of non-zero elements and, since there are quite a lot of dangling nodes on the internet, this makes the matrix denser. As a consequence, matrix multiplications involving $\mathbf{S}$ will take more time and resources.

3.6 Google matrix G

Still another modification to the matrix is necessary to ensure the existence of a unique stationary distribution. This primitivity modification will make the matrix primitive, i.e. aperiodic and irreducible. The Perron-Frobenius theorem then guarantees the uniqueness of the stationary distribution. In order to achieve this, the Google founders extended the model by allowing the random surfer to teleport to another webpage. The interpretation for this is that, a surfer does not only surf the internet by following hyperlinks, but also sometimes by jumping to a webpage by using the browser's address bar.

Definition 16. Let $0 \leq \alpha < 1$ denote the probability that the random surfer will follow a link. α is called the **teleportation parameter**.

Brin and Page used this parameter to define the following matrix.

Definition 17. The **Google matrix G** is defined as

$$\mathbf{G} = \alpha\mathbf{S} + (1 - \alpha)\mathbf{e}\mathbf{v}^T$$

where $\mathbf{e}$ is the $n \times 1$ vector containing all 1's and $\mathbf{v}$ the personalization vector.

The Google matrix models the behaviour of the random surfer explained above. The matrix $\mathbf{S}$ contains the probabilities of jumping from page to page by following hyperlinks, while the matrix $\mathbf{e}\mathbf{v}^T$ contains the probabilities of teleporting from one webpage to another by not following links. Those probabilities are multiplied by α and $1 - \alpha$ respectively, which results in the desired interpretation. It turns out that α directly influences the convergence rate of the power method, which we will be discussed in one of the next sections.

The matrix $\mathbf{G}$ has the following properties.

1. $\mathbf{G}$ is stochastic. $\mathbf{G}$ is a convex combination of $\mathbf{S}$, which is stochastic, and $\mathbf{e}\mathbf{v}^T$, which is also stochastic since $\|\mathbf{v}\|_1 = 1$ and $\mathbf{v} > \mathbf{0}$.

2. $\mathbf{G}$ is a positive matrix ($\mathbf{G} > 0$). This is the result of $\mathbf{v} > 0$ and $\alpha < 1$. This means that storage and matrix operations such as multiplication involving $\mathbf{G}$ require a lot of resources. Luckily however, $\mathbf{G}$ is a convex combination of the sparse matrix $\mathbf{S}$ and the product of the vectors $\mathbf{e}$ and $\mathbf{v}$ which means $\mathbf{G}$ does not have to be stored as a full matrix. This will also simplify matrix operations, which will be discussed in the section about the power method.

3. $\mathbf{G}$ is aperiodic. This is enforced by $\mathbf{G} > 0$.

4. $\mathbf{G}$ is irreducible. $\mathbf{G} > 0$, hence there is a non-zero probability that the random surfer will visit webpage P_j from any starting position P_i.

5. $\mathbf{G}$ is aperiodic and irreducible, hence primitive.

6. If the spectrum of $\mathbf{S}$ is $\{1, \mu_2, \mu_3, \ldots, \mu_n\}$ in descending order, then the spectrum of $\mathbf{G}$ is $\{1, \lambda_2 = \alpha\mu_2, \lambda_3 = \alpha\mu_3, \ldots, \lambda_n = \alpha\mu_n\}$.[4] The link structure of the internet makes it likely that $|\mu_2| \approx 1$ or even $|\mu_2| = 1$, while the biggest eigenvalue of both matrices is 1, hence $|\lambda_2| \leq \alpha < 1$. The second eigenvalue turns out to be important for the convergence rate of the power method for calculating the stationary distribition.

3.7 PageRank

Now that the existence and uniqueness of a stationary distribution is guaranteed, the PageRank vector can be defined in terms of this stationary distribution.

Definition 18. The unique **PageRank vector $\boldsymbol{\pi}$** is defined by

- $\mathbf{G}^T\boldsymbol{\pi} = \boldsymbol{\pi}$
- $\sum_{i=1}^{n} \pi_i = 1$
- $\pi_i > 0$ for all $1 \leq i \leq n$

The uniqueness of the PageRank vector is guaranteed by the Perron-Frobenius theorem. This theorem implies that 1 is a simple eigenvalue of $\mathbf{G}^T$ (1 has an algebraic multiplicity of one), and that there exists a positive eigenvector $\boldsymbol{\pi}$ corresponding to the eigenvalue 1, thereby implying the uniqueness of the vector $\boldsymbol{\pi}$ with $\|\boldsymbol{\pi}\|_1 = 1$. Now that the PageRank vector has been defined, a method to calculate the actual vector is needed. This will be the subject of the next section.

3.8 Power method

The original method proposed by Brin and Page for computing the PageRank vector, is the power method. This iterative method is used to find the dominant eigenvalue and eigenvector pair of a matrix, in this case the eigenvector corresponding to the dominant eigenvalue 1 of $\mathbf{G}$. The power method is considered a slow method, since the convergence rate depends on the difference between the first and second eigenvalue. The eigengap between the first and second eigenvalue is in general rather small. The eigengap for the Google matrix however, is not that small and can even be regulated in a simple way.

[4] As proven in [1] p. 46.

For the actual power method one takes a starting vector $\boldsymbol{\pi}^{(0)}$ with $\|\boldsymbol{\pi}^{(0)}\|_1 = 1$ to begin the iterative process

$$\boldsymbol{\pi}^{(k+1)} = \mathbf{G}^T\boldsymbol{\pi}^{(k)}$$

Usually the uniform vector $\boldsymbol{\pi}^{(0)} = (1/n)\mathbf{e}$ is chosen. As shown in theorem 2 the convergence does not depend on the chosen $\boldsymbol{\pi}^{(0)}$, although of course a starting vector closer to $\boldsymbol{\pi}$ will need fewer iterations to converge.
Since $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$ the process can be expressed as

$$\begin{aligned}
\boldsymbol{\pi}^{(k+1)} &= \alpha\mathbf{S}^T\boldsymbol{\pi}^{(k)} + (1-\alpha)(\mathbf{e}\mathbf{v}^T)^T\boldsymbol{\pi}^{(k)} \\
&= \alpha(\mathbf{H} + \mathbf{d}\mathbf{v}^T)^T\boldsymbol{\pi}^{(k)} + (1-\alpha)\mathbf{v}\mathbf{e}^T\boldsymbol{\pi}^{(k)} \\
&= \alpha\mathbf{H}^T\boldsymbol{\pi}^{(k)} + \alpha\mathbf{v}\mathbf{d}^T\boldsymbol{\pi}^{(k)} + (1-\alpha)\mathbf{v} \\
&= \alpha\mathbf{H}^T\boldsymbol{\pi}^{(k)} + (1-\alpha+\alpha\mathbf{d}^T\boldsymbol{\pi}^{(k)})\mathbf{v}
\end{aligned}$$

So the iteration process can be expressed in terms of a matrix multiplication involving the sparse matrix $\mathbf{H}$, which drastically reduces the required computation time and resources. As seen in section 3.4, multiplication involving $\mathbf{H}$ requires only $O(n)$ effort. Also, in order to calculate the PageRank, only storage for $\mathbf{d}$, $\mathbf{v}$ and $\mathbf{H}$ are necessary, the dense $\mathbf{G}$ does not have to be stored. Theorem 2 will show that the convergence rate can be regulated by α. These are the main reasons why Brin and Page chose for the power method.

Theorem 2. The power method applied to $\mathbf{G}$ will converge to the PageRank vector $\boldsymbol{\pi}$ if $\|\boldsymbol{\pi}^{(0)}\|_1 = 1$ and $\boldsymbol{\pi}^{(0)} \geq \mathbf{0}$.

Proof. Let $v_1, v_2, \ldots, v_n$ be an eigenbasis of $\mathbf{G}^T$ for $\mathbb{R}^n$ corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, assuming $1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$. Furthermore, let $\boldsymbol{\pi}^{(0)}$ with $\|\boldsymbol{\pi}^{(0)}\|_1 = 1$ and $\boldsymbol{\pi}^{(0)} \geq \mathbf{0}$ be the starting vector. Since $v_1, v_2, \ldots, v_n$ is a basis for $\mathbb{R}^n$ there exist $a_1, a_2, \ldots, a_n \in \mathbb{R}$ such that $\boldsymbol{\pi}^{(0)} = \sum_{i=1}^n a_i v_i$.[5] Multiplying this equation by $\mathbf{G}^T$ yields

$$\boldsymbol{\pi}^{(1)} = \mathbf{G}^T\boldsymbol{\pi}^{(0)} = \mathbf{G}^T\sum_{i=1}^n a_i v_i = \sum_{i=1}^n a_i\mathbf{G}^T v_i = \sum_{i=1}^n a_i\lambda_i v_i$$

Repeated multiplication by $\mathbf{G}^T$ yields

$$\boldsymbol{\pi}^{(k)} = (\mathbf{G}^T)^k\boldsymbol{\pi}^{(0)} = (\mathbf{G}^T)^k\sum_{i=1}^n a_i v_i = \sum_{i=1}^n a_i(\mathbf{G}^T)^k v_i = \sum_{i=1}^n a_i\lambda_i^k v_i$$

Since $\lambda_1 = 1$ and $\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$ it follows that

$$\lim_{k\to\infty}\boldsymbol{\pi}^{(k)} = \lim_{k\to\infty}\sum_{i=1}^n a_i\lambda_i^k v_i = a_1\lambda_1 v_1 = a_1 v_1 = a_1\boldsymbol{\pi}$$

So this process converges to a scalar multiple of the eigenvector corresponding to eigenvector 1. It remains to show that $a_1 = 1$. By assumption $\|\boldsymbol{\pi}^{(0)}\|_1 = 1$ and $\boldsymbol{\pi}^{(0)} \geq \mathbf{0}$. Now assume $\|\boldsymbol{\pi}^{(k)}\|_1 = 1$ and $\boldsymbol{\pi}^{(k)} \geq \mathbf{0}$. With the induced matrix norm it follows that

$$\|\boldsymbol{\pi}^{(k+1)}\|_1 = \|\mathbf{G}^T\boldsymbol{\pi}^{(k)}\|_1 \leq \|\mathbf{G}^T\|_1 \cdot \|\boldsymbol{\pi}^{(k)}\|_1$$

[5] Assuming $\mathbf{G}^T$ is diagonalizable. If $\mathbf{G}^T$ is not diagonalizable, the matrix can be written in Jordan form $\mathbf{G}^T = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. For each eigenvalue λ the corresponding Jordan block $\mathbf{J}_\lambda$ can be written as $\boldsymbol{\Lambda} + \mathbf{N}$ with $\boldsymbol{\Lambda}$ a diagonal matrix with λ on the diagonal and a nilpotent matrix $\mathbf{N}$. Hence if $\lambda < 1$, then $\mathbf{J}_\lambda^k \to \mathbf{0}$ for $k \to \infty$. For $\lambda = 1$ this does not hold, therefore the convergence of the power method to the corresponding eigenvector is still ensured.

By assumption $\|\boldsymbol{\pi}^{(k)}\|_1 = 1$ and $\boldsymbol{\pi}^{(k)} \geq \mathbf{0}$. Also $\|\mathbf{G}^T\|_1 = \max_j \sum_{i=1}^n |G_{ij}^T|$, the largest absolute column sum. $\mathbf{G}$ is row stochastic, hence $\mathbf{G}^T$ is column stochastic, therefore $\|\mathbf{G}^T\|_1 = 1$. Note that $\mathbf{G}^T > \mathbf{0}$ as well. So $\|\boldsymbol{\pi}^{(k)}\|_1 = 1$ and $\boldsymbol{\pi}^{(k)} \geq \mathbf{0}$ for all $k \in \mathbb{N}$. Moreover, for $k \geq 1$ it holds that $\boldsymbol{\pi}^{(k)} > \mathbf{0}$, since $\mathbf{G}^T > \mathbf{0}$. Now it follows that

$$1 = \|\boldsymbol{\pi}\|_1 = \|\lim_{k\to\infty} \boldsymbol{\pi}^{(k)}\|_1 = \|a_1 v_1\|_1 = |a_1|\|\boldsymbol{\pi}\|_1 = |a_1|$$

Hence $a_1 = 1$, since $\boldsymbol{\pi}$ is a positive vector. □

The significance of this proof is that it shows the convergence rate of the power method. Since at step k

$$\boldsymbol{\pi}^{(k)} = a_1\lambda_1 v_1 + a_2\lambda_2^k v_2 + \cdots + a_n\lambda_n^k v_n = \boldsymbol{\pi} + a_2\lambda_2^k v_2 + \cdots + a_n\lambda_n^k v_n$$

hence

$$\|\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}\| = \|a_2\lambda_2^k v_2 + \cdots + a_n\lambda_n^k v_n\|$$

If the eigenvalues are assumed to be in descending order, i.e. for each j $|\lambda_j| \geq |\lambda_{j+1}|$, the power method converges with $O(\lambda_2)$. As seen in section 3.6, $|\lambda_2| \leq \alpha$, hence the convergence rate is determined by α.

4 A probabilistic model for link changes

4.1 Basic model

The links between webpages will change over time. This is especially true for links between dynamic webpages like news pages and social media. The changes will affect the matrix **H** and thus the Google matrix **G**. The changes in the Google matrix will lead to changes in the actual PageRank vector. The only deterministic way a change in links can occur is if the owner or administrator of a webpage adds or removes an outlink to another webpage, since webpages have no direct control over inlinks. In order to model link changes it will be sufficient to model changes in outlinks.
First the adjacency matrix **A** representing a network is considered. The matrix contains the following elements

$$A_{ij} = \begin{cases} 1, & \text{if there is a link from } P_j \text{ to } P_i \\ 0, & \text{otherwise} \end{cases},$$

hence a column A_{*j} represents the outlinks of P_j. Let A_{*j} denote the outlink vector of page P_j. For every element A_{ij} in the outlink vector there are four events that can occur, namely:

A. The link from P_j to P_i remains unchanged, given that their was a link. $(P(1 \to 1) = \alpha)$

B. The link from P_j to P_i is removed, given that their was a link. $(P(1 \to 0) = 1 - \alpha)$

C. A link from P_j to P_i is added, given that there was no link. $(P(0 \to 1) = \beta)$

D. There remains no link from P_j to P_i, given that there was no link. $(P(0 \to 0) = 1 - \beta)$

These four events are the only events that can occur. The probability of the first two events sums up to 1, since, given that there is a link, the only possibilities that can happen are that it remains or that it is removed. Let the probability of event A be denoted by $P(1 \to 1) = \alpha$. Then the probability of event B is denoted by $P(1 \to 0) = 1 - \alpha$. Similarly the probabilities of event C and D can be denoted by $P(0 \to 1) = \beta$ and $P(0 \to 0) = 1 - \beta$ respectively. The probability that there is an outlink to P_i at some time t depends on the previous time. By defining the probability of there being a link at time t as

$$P_{ij}(t) := P(A_{ij} = 1 \text{ at time } t),$$

the probability of there not being a link at time t is given as

$$P(A_{ij} = 0 \text{ at time } t) = 1 - P_{ij}(t)$$

since those two occurrences cover the whole probability space ($P(A_{ij} = 0$ at time $t) + P(A_{ij} = 1$ at time $t) = 1$). The probability of there being a link at the next time step is given by

$$P_{ij}(t + \Delta t) = \alpha P_{ij}(t) + \beta[1 - P_{ij}(t)] = [\alpha - \beta]P_{ij}(t) + \beta, \tag{1}$$

and similarly the probability of there not being a link at the next time step is given as

$$1 - P_{ij}(t + \Delta t) = (1 - \alpha)P_{ij}(t) + (1 - \beta)[1 - P_{ij}(t)] = [\beta - \alpha]P_{ij}(t) + [1 - \beta] \tag{2}$$

Or more generally, $P_{ij}(t + (n + 1)\Delta t) = [\alpha - \beta]P_{ij}(t + n\Delta t) + \beta$, the probability of step $n + 1$ depends only on the probability at step n and thus is memoryless. The model introduced here is actually a two-state Markov chain, there being a link and there not being a link between each two pages.

Taking $P_{ij}(t) = 0$ if there is no link at starting point t and $P_{ij}(t) = 1$ if there is a link at t, the desired interpretations are achieved, namely $P_{ij}(t + \Delta t) = \beta$ in the first case and $P_{ij}(t + \Delta t) = \alpha$ in the second case. These probabilities correspond with event C and A respectively.

The assumption being made here is that α and β are fixed, i.e. they do not depend on any webpage or time. In order to obtain realistic results from this model it is assumed that α is relatively big and β is relatively small, that is, the probability that an outlink will remain is relatively high and the probability that a new outlink will be added is relatively small. The condition imposed on those probabilities is that $|\alpha - \beta| < 1$, or rather $\alpha - \beta < 1$ since $\alpha >> \beta$. The probability of there being a link after some number of steps can be stated in terms of the starting probability $P_{ij}(t)$.

Proposition 1. The n^{th} $(n \geq 1)$ step probability of there being a link from P_j to P_i is given by

$$P_{ij}(t + n\Delta t) = (\alpha - \beta)^n P_{ij}(t) + \beta \sum_{k=0}^{n-1} (\alpha - \beta)^k$$

Proof. By induction. For $n = 1$ formula (1) is clearly obtained.
Now assume the formula holds for some $s \in \mathbb{N}$, i.e.

$$P_{ij}(t + s\Delta t) = (\alpha - \beta)^s P_{ij}(t) + \beta \sum_{k=0}^{s-1} (\alpha - \beta)^k$$

By formula (1) the following formula for step $s + 1$ is obtained

$$\begin{aligned} P_{ij}(t + (s+1)\Delta t) &= (\alpha - \beta) \left[(\alpha - \beta)^s P_{ij}(t) + \beta \sum_{k=0}^{s-1} (\alpha - \beta)^k \right] + \beta \\ &= (\alpha - \beta)^{s+1} P_{ij}(t) + \beta \left[1 + \sum_{k=0}^{s-1} (\alpha - \beta)^{k+1} \right] \\ &= (\alpha - \beta)^{s+1} P_{ij}(t) + \beta \left[(\alpha - \beta)^0 + \sum_{k=1}^{s} (\alpha - \beta)^k \right] \\ &= (\alpha - \beta)^{s+1} P_{ij}(t) + \beta \sum_{k=0}^{s} (\alpha - \beta)^k \end{aligned}$$

This concludes the proof. □

Since $\alpha - \beta < 1$ it follows that

$$\lim_{n\to\infty} P_{ij}(t + n\Delta t) = \lim_{n\to\infty} \left[(\alpha - \beta)^n P_{ij}(t) + \beta \sum_{k=0}^{n-1} (\alpha - \beta)^k \right] = 0 + \beta \frac{1}{1 - \alpha + \beta} = \frac{\beta}{1 - \alpha + \beta}$$

As time progresses the probability that there is a link convergences to the probability above. This means that approximately a fraction $\beta/(1-\alpha+\beta)$ of the pages will have an inlink from P_j.

The limit of the convergence process is now known. Now the rate of convergence of

$$\lim_{n\to\infty} P_{ij}(t + n\Delta t) = \lim_{n\to\infty} \left[(\alpha - \beta)^n P_{ij}(t) + \beta \sum_{k=0}^{n-1} (\alpha - \beta)^k \right]$$

to the limit value will be investigated. Obviously the term $(\alpha-\beta)^n P_{ij}(t)$ converges at the rate of $(\alpha-\beta)^n$ to 0, since $\alpha-\beta<1$. As will be shown, the summation term converges at the same rate to $\beta/(1-\alpha+\beta)$. By taking the infinite sum

$$\beta+\beta(\alpha-\beta)+\beta(\alpha-\beta)^2+\ldots$$

and subtracting the first N terms the N^{th}-step error for the geometric series is obtained

$$\begin{aligned} & [\beta+\beta(\alpha-\beta)+\beta(\alpha-\beta)^2+\ldots]-[\beta+\beta(\alpha-\beta)+\cdots+\beta(\alpha-\beta)^{N-1}] \\ = \;& \beta(\alpha-\beta)^N+\beta(\alpha-\beta)^{N+1}+\beta(\alpha-\beta)^{N+1}+\ldots, \end{aligned}$$

which is itself a geometric series which sums to $\beta(\alpha-\beta)^N/(1-\alpha+\beta)$. So the total n^{th}-step error term is given by

$$\varepsilon_{ij}(t+n\Delta t)=\frac{\beta(\alpha-\beta)^n}{1-\alpha+\beta}-(\alpha-\beta)^n P_{ij}(t)=(\alpha-\beta)^n\left(\frac{\beta}{1-\alpha+\beta}-P_{ij}(t)\right) \tag{3}$$

which converges to 0 at the rate of $(\alpha-\beta)^n$.

4.2 Results of the basic model

To illustrate the theoretical results a network containing a subset of 9914 webpages in the `*.cs.stanford.edu` domain will be used. First of all, it will be tested whether or not the convergence of the probability of there being a link matches the theoretical rate of convergence. For this test the values $\alpha=0.95$ and $\beta=0.05$ will be used. Two pages, P_3 and P_5, within column 4 (the outlinks of P_4) have been picked so as to illustrate the convergence of the probability. The probabilities will converge to $0.05/(1-0.95+0.05)=1/2$.

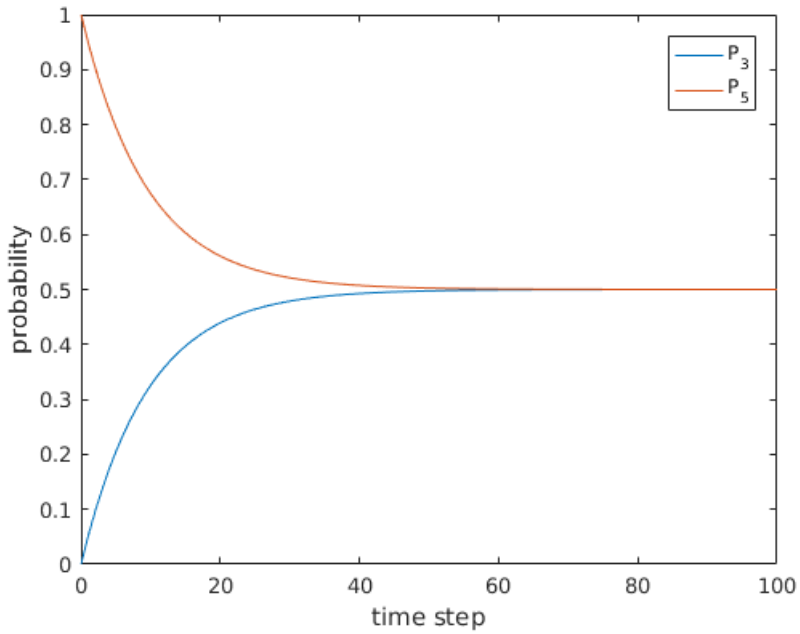


Figure 3: Convergence of the probability of there being an outlink from P_4 to P_3 and P_5

Notice that in figure 3 the convergence is perfectly symmetrical in the 0.5 probability line. As can be seen in equation 3, this is because if $\beta/(1-\alpha+\beta)=1/2$ (i.e. $\beta=1-\alpha$) then the

error of a page with start value 1 is equal to the absolute value of the error of a page with start value 0 at any given step. Figure 3 also shows that after about 50 steps the error is already quite small. By calculating the error after 50 steps by using formula 3 the following values are obtained.

$$\begin{aligned}\varepsilon_{3,4}(t+50\Delta t) &= \frac{0.05(0.95-0.05)^{50}}{1-0.95+0.05} - (0.95-0.05)^{50}\cdot 0 \approx 0.0026\\ \varepsilon_{5,4}(t+50\Delta t) &= \frac{0.05(0.95-0.05)^{50}}{1-0.95+0.05} - (0.95-0.05)^{50}\cdot 1 \approx 0.0026 - 0.0052 = -0.0026\end{aligned}$$

Those values are indeed quite small. In table 1 some more approximate values of the respective probabilities are given.

n	$P_{3,4}(t+n\Delta t)$	$P_{5,4}(t+n\Delta t)$
0	0	1
1	0.05	0.95
2	0.095	0.905
3	0.1355	0.8645
10	0.3257	0.6743
20	0.4392	0.5608
30	0.4788	0.5212
50	0.4974	0.5026
75	0.4998	0.5002
100	0.5	0.5

Table 1: Convergence of probability of there being a link after n steps rounded to four decimals.

By setting an upper bound for the error equation the number of steps required to fall within that bound can be calculated as follows.

$$(\alpha-\beta)^n\left(\frac{\beta}{1-\alpha+\beta} - P_{ij}(t)\right) \leq \delta$$

In the case of $\alpha = 0.95$ and $\beta = 0.05$ $P_{ij}(t)$ can be ignored because of symmetry. Then the number of steps can be calculated by

$$n \geq \frac{\log(2\delta)}{\log(9/10)}$$

Using this formula following table is obtained.

ε	n
10^{-1}	16
10^{-2}	38
10^{-3}	59
10^{-4}	81

Table 2: The upper bound for the error and the number of steps required to achieve it.

Now the actual behaviour in outlinks in the columns are investigated and compared to the behaviour of the probability of there being a link.

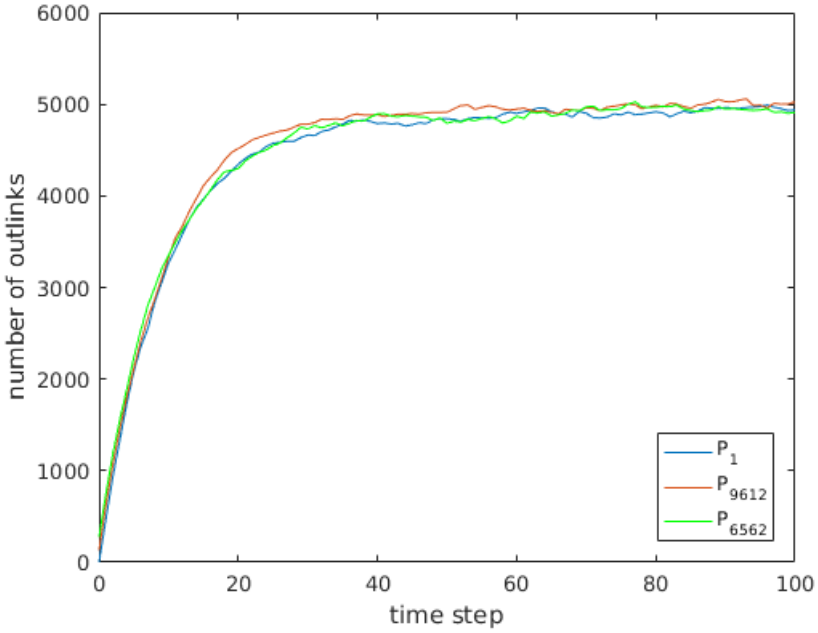


Figure 4: Change in the number of outlinks of P_1, P_{9612} and P_{6562} over time.

Figure 4 shows the change in the number of outlinks over time. The webpages P_1, P_{9612} and P_{6562} have a starting number of outlinks of 0, 133 and 277 respectively. The probability of there being a link at any given position (except from a webpage to itself) will converge to $\beta/(1-\alpha+\beta)$. Taking α and β as before the probabilities converge to $1/2$. This means that after the probabilities are close enough to their limit value the number of outlinks in any given column will fluctuate around approximately $(1/2)\cdot(n-1)$ where n is the total number of webpages. In figure 4 one can see that after about 40 time steps the number of outlinks is already close to the limit value. The actual number of outlinks for the pages in this particular simulation after 40 iterations are 4789, 4887 and 4895 respectively. The difference between the actual and expected number of outlinks can mainly be attributed to the fact that because of the starting number of outlinks the bigger portion of the 9914 webpages does not have an inlink from any of the three webpages. Since, as can seen in figure 3 and table 1, the chance of there being a link while there was none at the beginning will increase to 0.5 and there being more pages with no inlinks the average chance of there being a link in any of the three columns will be below 0.5. As time progresses this effect will disappear and the number of outlinks will fluctuate around the expected value.

The next thing that will be discussed is the effect of the model on the adjacency matrix as a whole and on the PageRank vector. For the individual columns the number of outlinks will eventually fluctuate around $(1/2)(n-1)$ (given α and β as before). This means that for the adjacency matrix the number of outlinks will eventually fluctuate around $(1/2)n(n-1) = 49,138,741$ in this case.

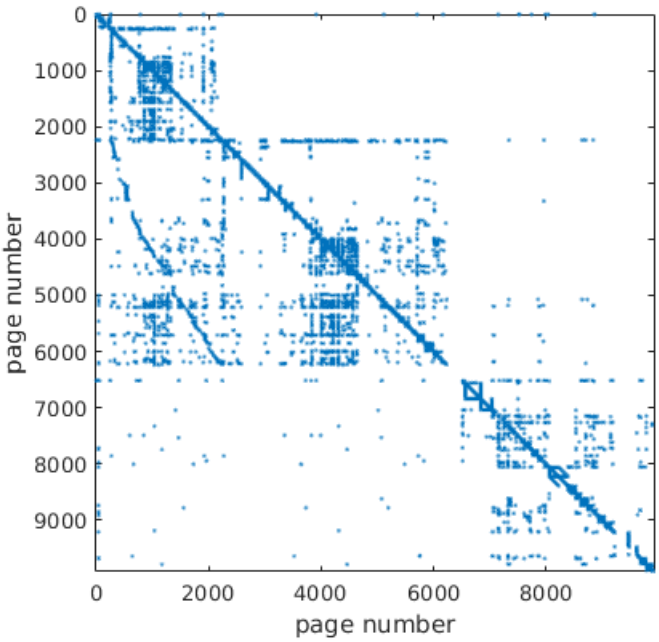


Figure 5: A graphical representation of the links in the 9914×9914 matrix.

Figure 5 shows the structure of the original `*.cs.stanford.edu` matrix. The total number of outlinks is $35,555$, which is only about 0.04% of the total number of possible links ($9914 \times 9913 = 98,277,482$). Figure 6 shows what part of the matrix looks like after one iteration.

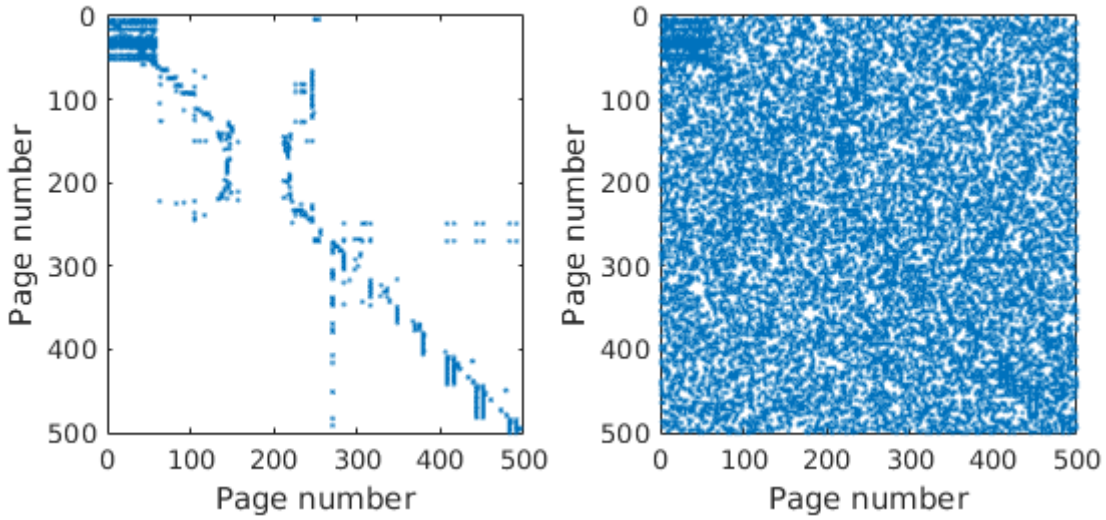


Figure 6: On the left the original links between the first 500 pages and on the right the links between the same pages after one iteration.

To be able to see enough detail, only the links between the first 500 webpages are plotted. After just one iteration it already looks like random noise. The matrix after this particular simulation now contains $4,945,484$ links, which is about 5% of the total number of possible links. Plotting

the number of outlinks over time the following results is obtained.

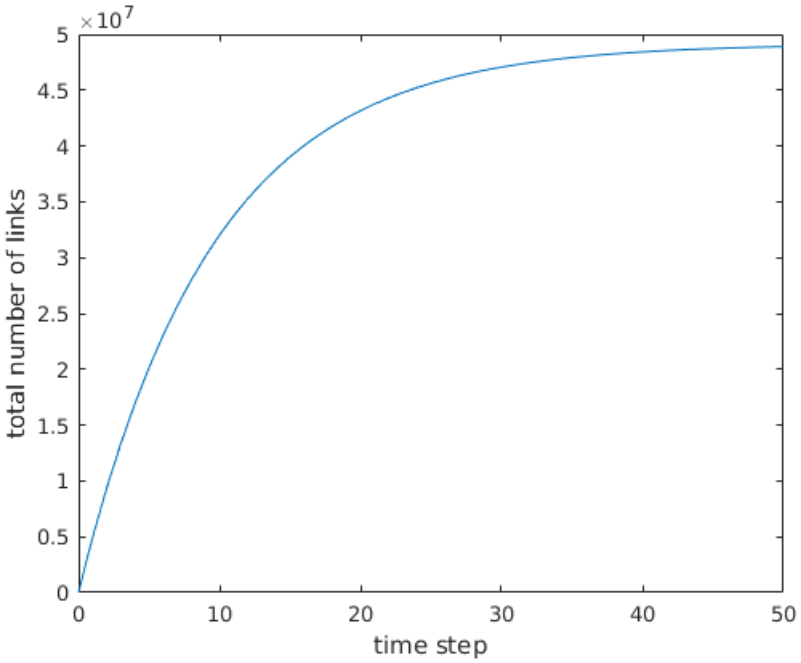


Figure 7: The change in the total number of outlinks over time.

Figure 7 shows no real surprises. The number of outlinks progress towards the predicted value of about $49,138,741$. It seems to take quite a number of steps to get there, but this is explained by the fact that since the number of links in the original matrix is so low, the average probability of there being a link is much lower than 0.5 in the beginning, especially since there are quite a number of dangling nodes, hence it will take longer to actually reach the predicted value.

How do these matrix changes affect the PageRank vector? First the PageRank of the original matrix will be computed and plotted in descending order by using a permutation matrix $\mathbf{P}$ (i.e. $\boldsymbol{\pi}_{\text{sorted}} = \mathbf{P}\boldsymbol{\pi}$). After one time step the PageRank vector will be computed again based on the new adjacency matrix and vector $\boldsymbol{\pi}'$ is obtained. This vector is sorted by using the permutation matrix $\mathbf{P}$ which was used previously, i.e. $\boldsymbol{\pi}'_{\text{sorted}} = \mathbf{P}\boldsymbol{\pi}'$. This means that the pages will be plotted in the same order over for each time step.

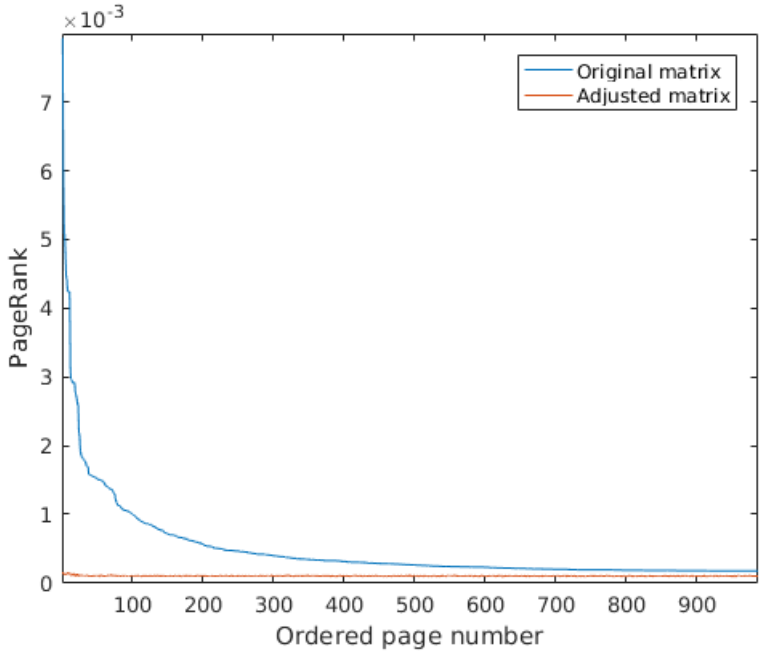


Figure 8: The PageRank for the 1000 pages with highest PageRank based on the original matrix.

As can be seen in figure 8, after only one iteration the PageRank is already almost flattened out. The difference between the maximum and minimum PageRank based on the original matrix is about $7.9 \cdot 10^{-3}$, while the spread based on the adjusted matrix is a mere $6.8 \cdot 10^{-5}$. After 10 iterations it is down to about 10^{-5}. This is exactly what one would expect to happen. The links are randomly distributed, while the number of links in each column will tend towards 9913/2. After 10 iterations the average number of links in a column is about 3229 and since $\mathbf{A}$ is a square matrix, the average per row will be the same. So the number of outlinks and number of inlinks will be about the same for each webpage and hence every webpage will have about the same importance as any other webpage according to the PageRank model. After 10 iterations the maximum number of inlinks is 3431 while the minimum is 3041. For the outlinks the maximum and minimum equal 3432 and 3034 respectively. As time increases the spread will decrease, barring statistical noise, and the PageRank vector will flatten out even more.

In order to obtain a more interesting result in PageRank changes, a more modest β will be used. The original adjacency matrix contains 35555 links. The average number of outlinks per page is then $35555/9914 \approx 3.5$. By taking $\beta = 0.00036$, the expected number of links in the matrix after one time step will be $0.95 \cdot 35555 + 0.00036 \cdot (9914 \cdot 9113 - 35555) \approx 69144$ which is about twice the original number of links. Hence the changes in PageRank are expected to be less drastic since the maximum number of inlinks for the webpages is 344 and the minimum is 0. The expected number for the page with maximum number of inlinks is $0.95 \cdot 344 + 0.00036 \cdot 9569 \approx 330$. PageRank does not solely depend on the number of inlinks, however major changes in outlinks for the webpages are not expected for similar reasons as major changes in the number inlinks are not expected. On average about 3.5 new outlinks per column will be added randomly, which will flatten the PageRank vector somewhat. Figure 9 shows the actual results.

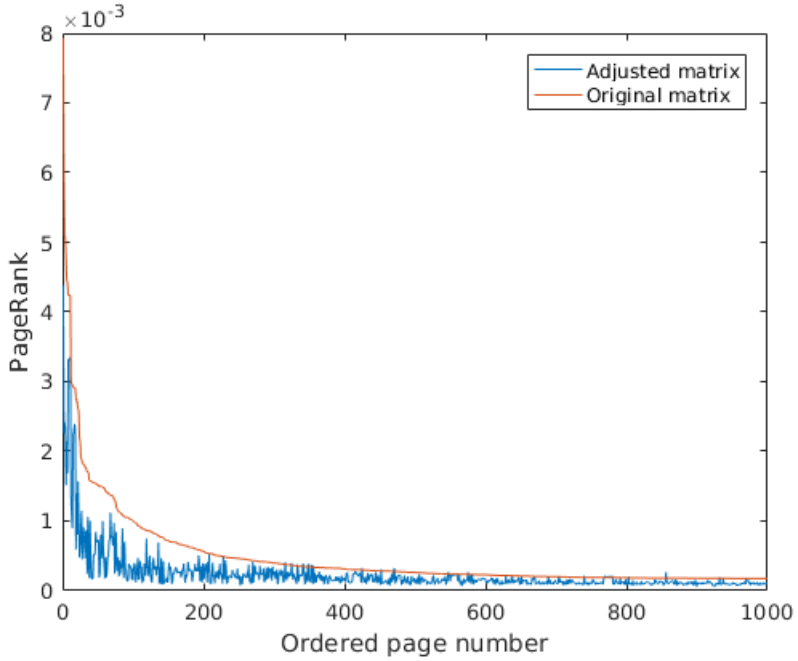


Figure 9: The average PageRank over ten simulations for the 1000 pages with highest PageRank before and after one iteration using a small β .

In figure 8 the difference between the maximum and minimum PageRank was about $7.9 \cdot 10^{-3}$. After averaging 10 simulations of 1 iteration each using $\beta = 0.00036$ the spread is down to $4.3 \cdot 10^{-3}$, which is much less drastic than the mere $6.8 \cdot 10^{-5}$ in figure 8. The results seem to agree with the predictions.

Bigger values for α and smaller values for β like the one used above are especially useful to model shorter time scales. A few more results will be shown using $\alpha = 0.99$ and $\beta = 10^{-3}$ to be able to compare the behaviour between different α's and β's.

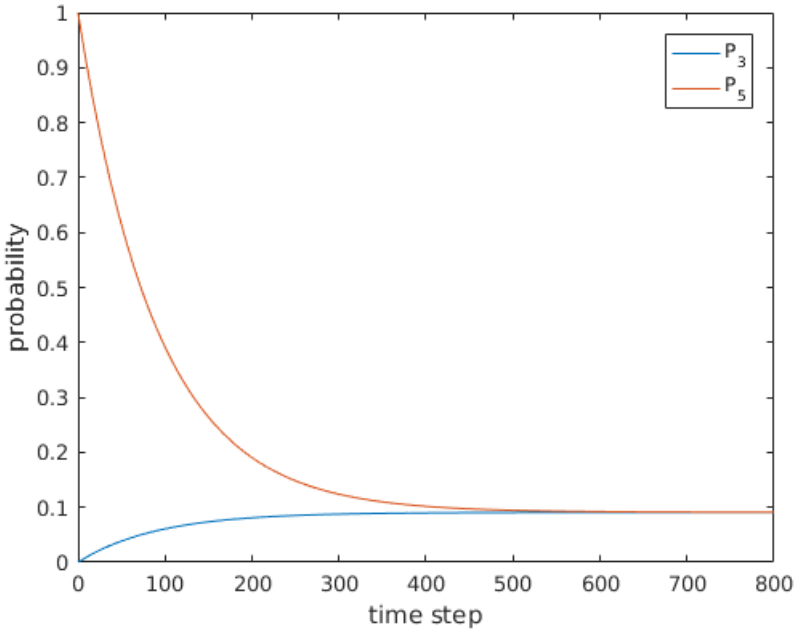


Figure 10: Convergence of the probability of there being a link.

In figure 10 one can see that with the chosen α and β it takes much longer for the probability to converge, compared to convergence with the previously used values. The limit value is $0.001/(1-0.99+0.001) = 1/11$. The convergence is not symmetric either, since the limit value is not equal to $1/2$. In the previous section it was concluded that the rate of convergence was determined by $(\alpha - \beta)^n$, hence the slower convergence can be explained by $\alpha - \beta$ now being closer to 1. Since the probability convergence is much slower, the number of links is expected to converge much slower as well. Using the same webpages as in figure 4, figure 11 shows the number of outlinks of time.

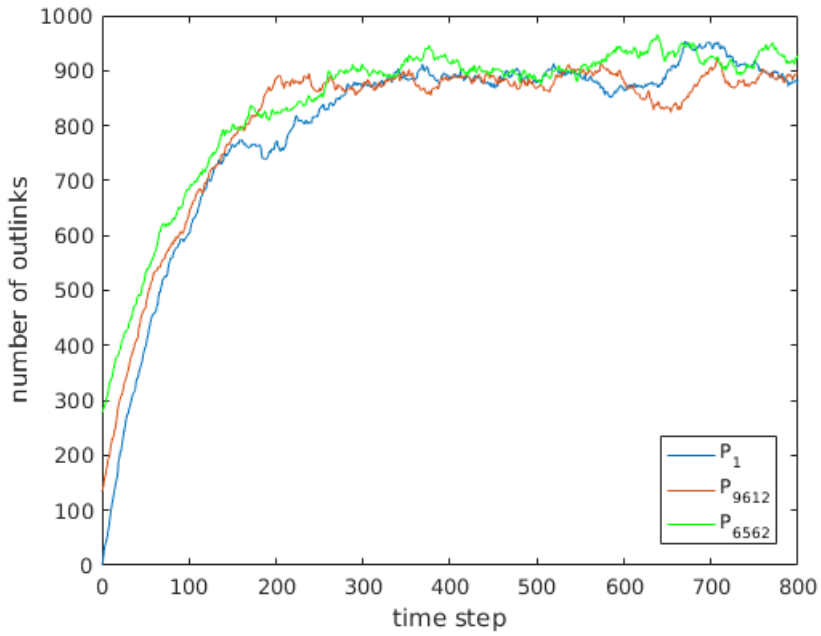


Figure 11: The number of outlinks for page P_1, P_{9612} and P_{6562} over time.

Figure 11 shows a much slower convergence process. The values will eventually fluctuate around $(1/11) \cdot 9913 \approx 901$. Figure 12 shows what the links structure looks like after one iteration.

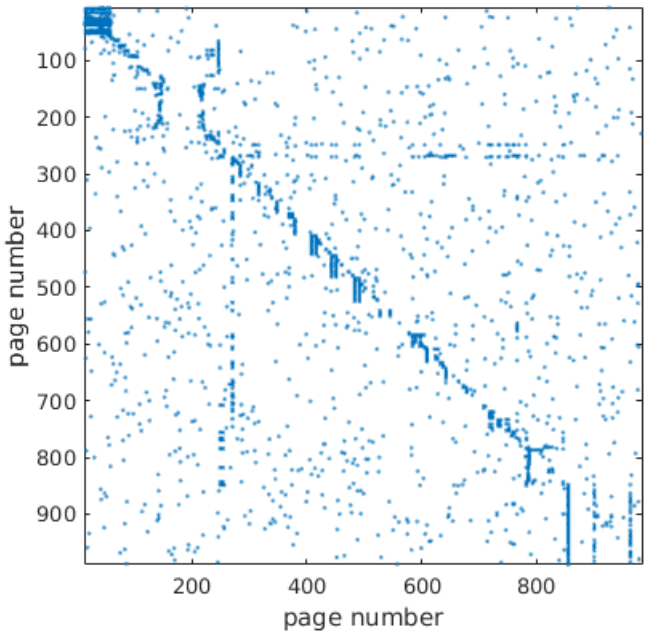


Figure 12: Links between the first 1000 webpages after one iteration.

Figure 12 shows that after one iteration most of the original structure (as can be seen in figure 5) is still intact, as can be expected since the probability for a link to stay is quite high

(0.99). The total number of outlinks is now 133,032, which is about the expected value of $35555 \cdot 0.99 + (9914 \cdot 9913 - 35555) \cdot 0.001 \approx 133,473$.
The expected total number of outlinks is $(9914 \cdot 9913)/11) \approx 8,934,316$.

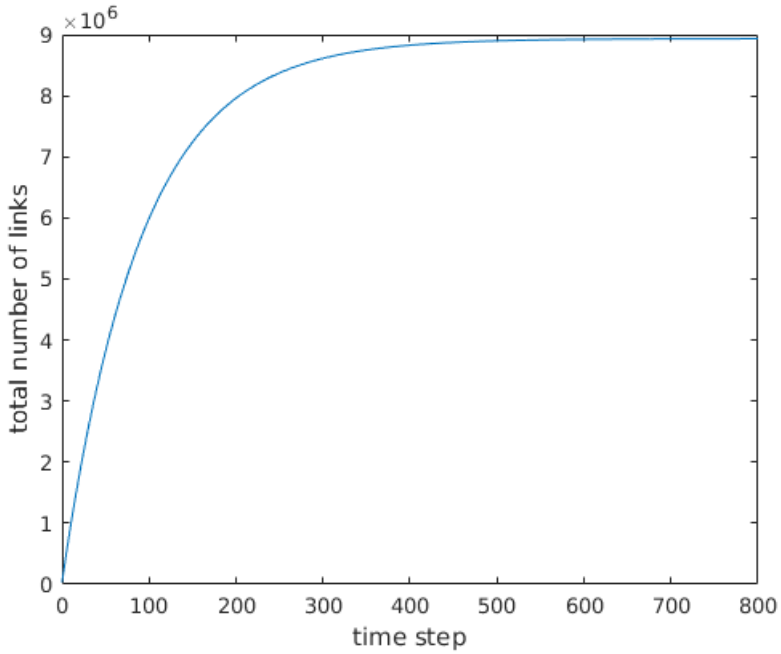


Figure 13: Total number of outlinks over time.

As seen in figure 10 the convergence of the probability of there being a link takes a lot of steps, hence the total number of links will have a low convergence rate as well.

By way of conclusion it can be said that the basic model does not yield very interesting results, since this probabilistic model relies heavily on purely random processes. In the next section a more interesting and realistic model will be introduced.

4.3 Objective extended model

The basic model does not yield very interesting results in terms of realistic behaviour. To make the model less random and more realistic the α and β are made time and webpage dependent, so that α and β are actually $\alpha_{ij}(t)$ and $\beta_{ij}(t)$ respectively. The rationale behind the webpage dependency is that not every webpage changes its outlinks as often. A news page for example will change its outlinks very frequently. A small museum webpage, however, will not change its outlinks all that often, if at all. Hence the probabilities for adding and removing outlinks will depend on P_j, the webpage that provides the outlinks. The probabilities also depend on P_i, the page the potential outlink is pointing at, since the outlinks to more popular pages are less likely to be removed and more likely to be added. For instance, an outlink to a popular webpage as Wikipedia (which probably has a relatively high PageRank) is likely to stay as it is a stable webpage and its content is unlikely subject to major changes. This is also why webpages are more likely to add an outlink pointing at a popular webpage. A popular webpage is usually stable. The popularity of a webpage at a certain point in time is measured by its PageRank $\pi_i(t)$. This important metric will be used in the definition of $\alpha_{ij}(t)$ and $\beta_{ij}(t)$ to simulate the

intended behaviour. Through changes in links, the PageRank of a webpage will change over time, hence the probabilities are time dependent. The model is objective in the sense that webpages owners or administrators base their decisions on an objective metric, the PageRank. An active webpage, a webpage with lots of outlinks at the start, is deemed more likely to change its outlinks. $|P_j|$ can be used in the definition to account for this factor.[6]

Definition 19. The probabilities for keeping an outlink and adding an outlink from P_j to P_i at time t are defined as respectively

$$\begin{aligned} \alpha_{ij}(t) &:= 1 - \left(1 - \gamma \frac{\pi_i(t)}{\max\limits_i \pi_i(t)}\right) \delta \frac{|P_j|}{\max\limits_j |P_j|} \\ \beta_{ij}(t) &:= \frac{\pi_i(t)}{\max\limits_i \pi_i(t)} \varepsilon \frac{|P_j|}{\max\limits_j |P_j|} \end{aligned}$$

The definition of α uses damping factors $0 < \gamma \leq 1$ and $0 < \delta \leq 1$ to ensure that the influence of both the PageRank and outdegree factors are limited. This is due to the observation that outlinks are not that likely to be removed. The damping factors create an artificial lower bound for α. There is also a damping factor $0 < \varepsilon \leq 1$ in the definition of β. This is to ensure that the influence of the outdegree factor is dampened. For some data sets this is necessary, since otherwise the entire link structure of a network could get lost after one iteration because of the enormous amount of new links. Notice that as $\pi_i(t)$ approaches $\max_i[\pi_i(t)]$, α increases, and as $|P_j|$ approaches $\max_j |P_j|$, α decreases. β increases as both $\pi_i(t)$ and $|P_j|$ approach the respective maximums. In both cases the desired behaviour for the probabilities is achieved. Also note that α and β are well-defined, i.e. $0 \leq \alpha, \beta \leq 1$, since $0 < \pi_i(t)/\max_i[\pi_i(t)] \leq 1$ and $0 \leq |P_j|/\max_j |P_j| \leq 1$.

Now equation 1, the probability of there being a link from P_j to P_i after one iteration, becomes

$$P_{ij}(t + \Delta t) = \left[1 - \left(\delta + (\varepsilon - \gamma\delta) \frac{\pi_i(t)}{\max\limits_i \pi_i(t)}\right) \frac{|P_j|}{\max\limits_j |P_j|}\right] P_{ij}(t) + \frac{\pi_i(t)}{\max\limits_i \pi_i(t)} \varepsilon \frac{|P_j|}{\max\limits_j |P_j|}$$

Since the long-term behaviour of the $\pi_i(t)$ factor cannot be accurately predicted, the long-term behaviour of the probability above is unknown. Some very rough short-term predictions can be made based on this equation, after all the original PageRank and outdegree values are known. The predictions are very limited however. The more general version of the formula is given by

$$P_{ij}(t + n\Delta t) = \alpha_{ij}(t + (n-1)\Delta t) P_{ij}(t + (n-1)\Delta t) + \beta_{ij}(t + n - 1)\Delta t)[1 - P_{ij}(t + (n-1)\Delta t)]$$

This again is a Markov chain, however this time it is time inhomogeneous, meaning the transition probabilities are time-dependent. Using the formula above one can find the general n^{th}-step formula based on the original probability. Using starting time $t = 0$ and writing $n = n\Delta t$ for simplicity, the following formula is obtained.

[6] By letting $|P_j|$ be time dependent, $|P_j|(t)$ being the number of outlinks at time t, β gets bigger the more outlinks are added, which in turn increases the probability of adding an outlink, resulting in a cycle. So a fixed value for $|P_j|$ is used.

Proposition 2. The n^{th}-step ($n \geq 1$) probability of there being a link from P_j to P_i is given by

$$P_{ij}(n) = P_{ij}(0)\prod_{k=0}^{n-1}[\alpha_{ij}(k) - \beta_{ij}(k)] + \sum_{p=0}^{n-1}\left(\beta_{ij}(p)\prod_{m=p+1}^{n-1}[\alpha_{ij}(m) - \beta_{ij}(m)]\right)$$

Proof. For $n = 1$ once again formula 1 is obtained.
Now assume the formula holds for some $s \in \mathbb{N}$. Then:

$$\begin{aligned}
P_{ij}(s+1) &= \alpha_{ij}(s)P_{ij}(s) + \beta_{ij}(s)[1 - P_{ij}(s)] \\
&= \alpha_{ij}(s)\left[P_{ij}(0)\prod_{k=0}^{s-1}[\alpha_{ij}(k) - \beta_{ij}(k)] + \sum_{p=0}^{s-1}\left(\beta_{ij}(p)\prod_{m=p+1}^{s-1}[\alpha_{ij}(m) - \beta_{ij}(m)]\right)\right] \\
&+ \beta_{ij}(s)\left[1 - P_{ij}(0)\prod_{k=0}^{s-1}[\alpha_{ij}(k) - \beta_{ij}(k)] - \sum_{p=0}^{s-1}\left(\beta_{ij}(p)\prod_{m=p+1}^{s-1}[\alpha_{ij}(m) - \beta_{ij}(m)]\right)\right] \\
&= (\alpha_{ij}(s) - \beta_{ij}(s))P_{ij}(0)\prod_{k=0}^{s-1}[\alpha_{ij}(k) - \beta_{ij}(k)] \\
&+ (\alpha_{ij}(s) - \beta_{ij}(s))\sum_{p=0}^{s-1}\left(\beta_{ij}(p)\prod_{m=p+1}^{s-1}[\alpha_{ij}(m) - \beta_{ij}(m)]\right) + \beta_{ij}(s) \\
&= P_{ij}(0)\prod_{k=0}^{s}[\alpha_{ij}(k) - \beta_{ij}(k)] + \sum_{p=0}^{s-1}\left(\beta_{ij}(p)\prod_{m=p+1}^{s}[\alpha_{ij}(m) - \beta_{ij}(m)]\right) + \beta_{ij}(s) \\
&= P_{ij}(0)\prod_{k=0}^{s}[\alpha_{ij}(k) - \beta_{ij}(k)] + \sum_{p=0}^{s}\left(\beta_{ij}(p)\prod_{m=p+1}^{s}[\alpha_{ij}(m) - \beta_{ij}(m)]\right)
\end{aligned}$$

This concludes the proof. □

The use for this formula is limited, since at each step the α and β still have to be computed, because they are time-dependent. What is known is that $|\alpha - \beta| < 1$, if $|P_j| > 0$. If $|P_j| = 0$, then $\alpha = 1$ and $\beta = 0$, however, in that case $P_{ij}(0) = 0$ for all pages P_i since then P_j would have no outlinks. So in both cases when $n \to \infty$ the first term on the right converges to 0, which means that the probability converges to the same value independent of there being a link at the start.

4.4 Results of the objective extended model

The subset of 9914 webpages in the `*.cs.stanford.edu` domain will be used to illustrate the behaviour of the extended model. The values for the damping factors used are $\gamma = 0.95$, $\delta = 0.05$ and $\varepsilon = 0.1$. The value for γ is chosen, because its main use it to prevent the α to become 1 if P_i is the webpage with maximum PageRank. It seems realistic to suppose that there is a non-zero probability of removing the outlink to the most popular webpage on the web, but it will be quite small, in this case $\delta/20 = 1/400$ for the most active webpage. Both δ and ε are chose quite small, since the goal of the extended model was to get more realistic behaviour. Using $\delta = 0.05$ and $\delta = 0.1$ results in realistic behaviour. This means that not the whole link structure will be lost in one time step, but instead the results will show relatively

small changes, which (more) realistically models weekly or monthly changes in a network. However the values, as will be indicated by the results, are not too small. This is to illustrate the model's behaviour. Too small values for the damping factors would result in barely any change at all, which does not really show the behaviour of the model.

Using these values the following results are obtained. First the results in change in link structure.

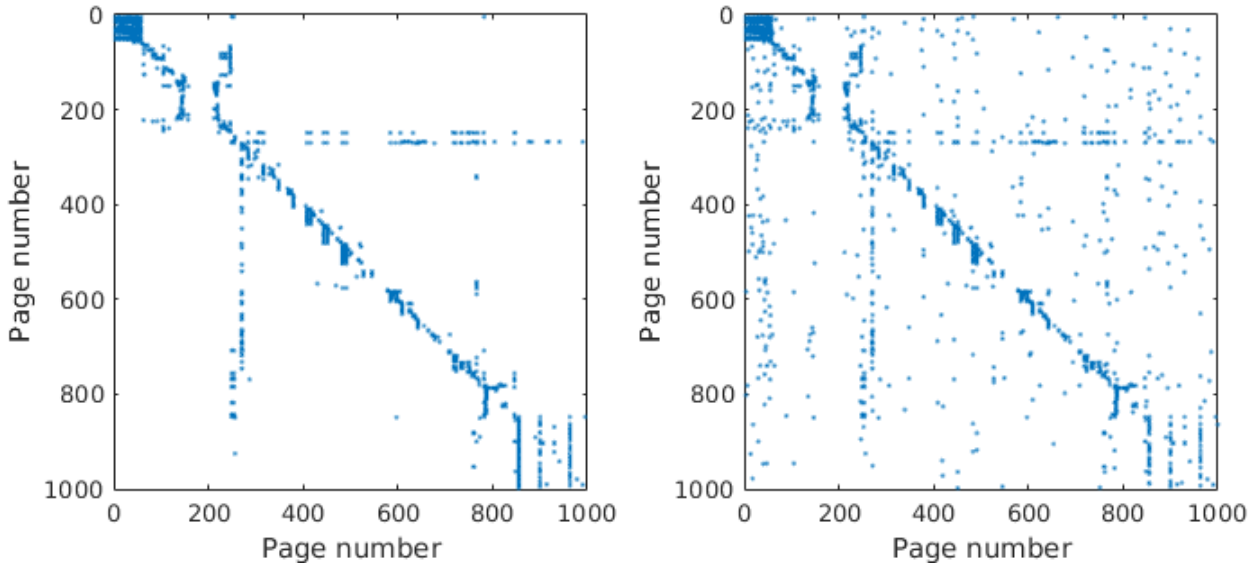


Figure 14: The links between the first 1000 pages in the original matrix and after 50 steps.

Figure 14 shows the changes in links of the first 1000 pages after 50 steps. As expected, the original link structure is mostly intact. The figure on the right shows that some pages seem to have more inlinks and outlinks than others. This is also what is to be expected, since the probability of there being a link now depends on the PageRank of P_i and outdegree of P_j. This behaviour is even better illustrated by the change in the whole network.

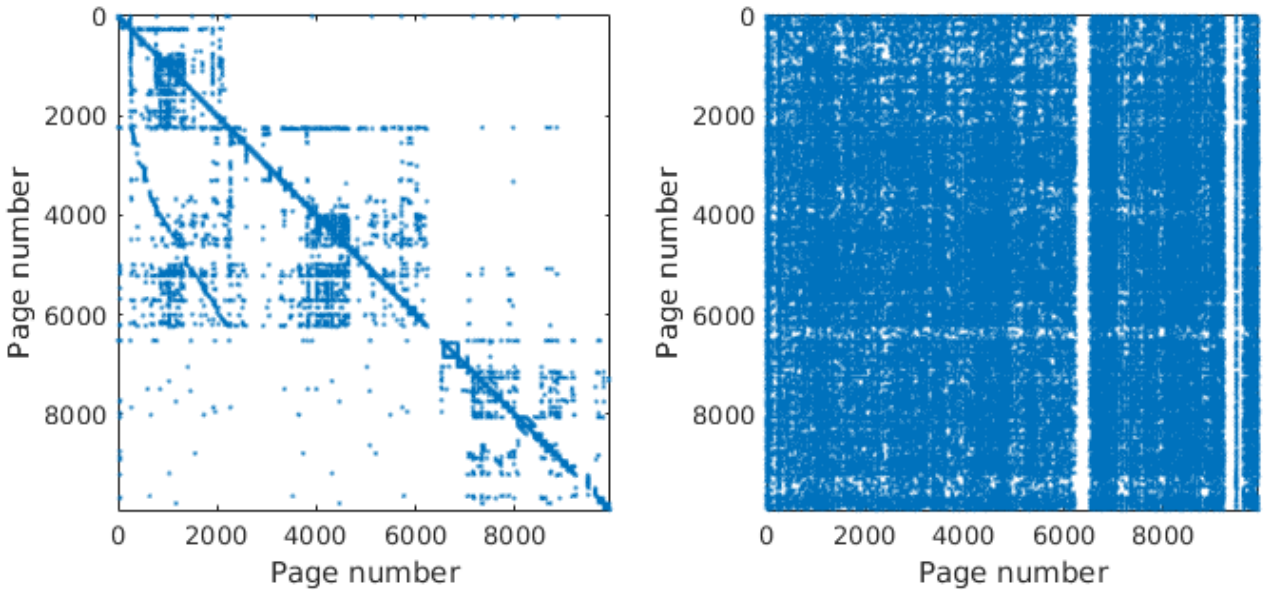


Figure 15: The links between all the pages in the original matrix and after 50 steps.

As can be seen in figure 15 in the blank columns, the webpages with no outlinks in the beginning remain this way. In the definitions of α and β, if $|P_j| = 0$, then $\beta = 0$ for every P_i and at every time step, hence figure 15 shows the expected and intended behaviour. Some rows look more dense than others, which means that some webpages have more inlinks than others. In this case, P_{2264} has both the highest PageRank and number of inlinks at the start, namely about 0.0079 and 340 respectively. After 50 iterations P_{2264} still has the highest PageRank and number of inlinks, now about 0.0064 and 873 respectively. The next figure shows the progression of the total number of links in the matrix.

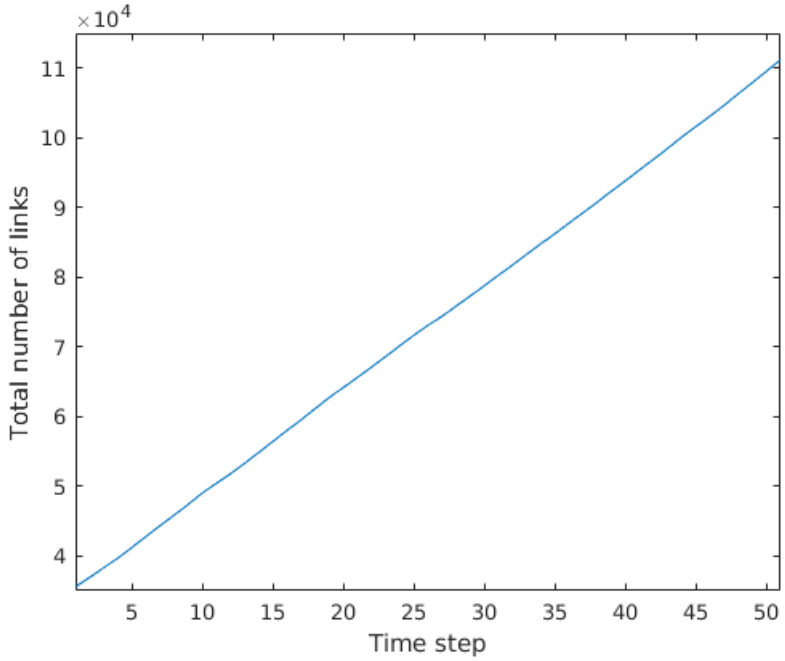


Figure 16: Progression of the total number of links over time.

Figure 16 shows a stable increase in the number of links over time. This means that the α and β must stay relatively stable over time. This in turn means that the outdegree and PageRank stay relatively stable over time. This is however quite surprising, since the number of outlinks has risen from 35555 to 62672, which means that the average outdegree has almost doubled. So then the ratio between the outdegree and the maximum outdegree must have stayed relatively stable. This could certainly be the case since activity is proportional to outdegree at the start. As will be shown in figure 18 the PageRank has changed quite a bit. The small fluctuation in the α and β can mainly be attributed to the chosen values for the damping factors δ and ε. They are quite small and any fluctuations in outdegree and PageRank is multiplied by this value to form the α and β, hence the impact of the PageRank and outdegree is indeed rather low.

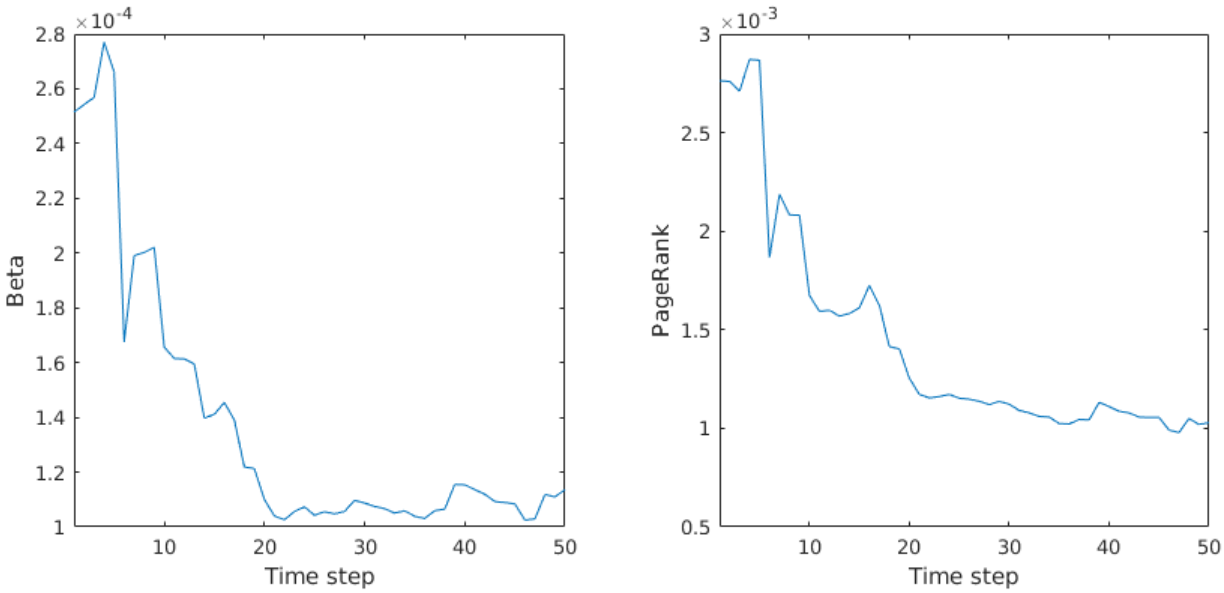
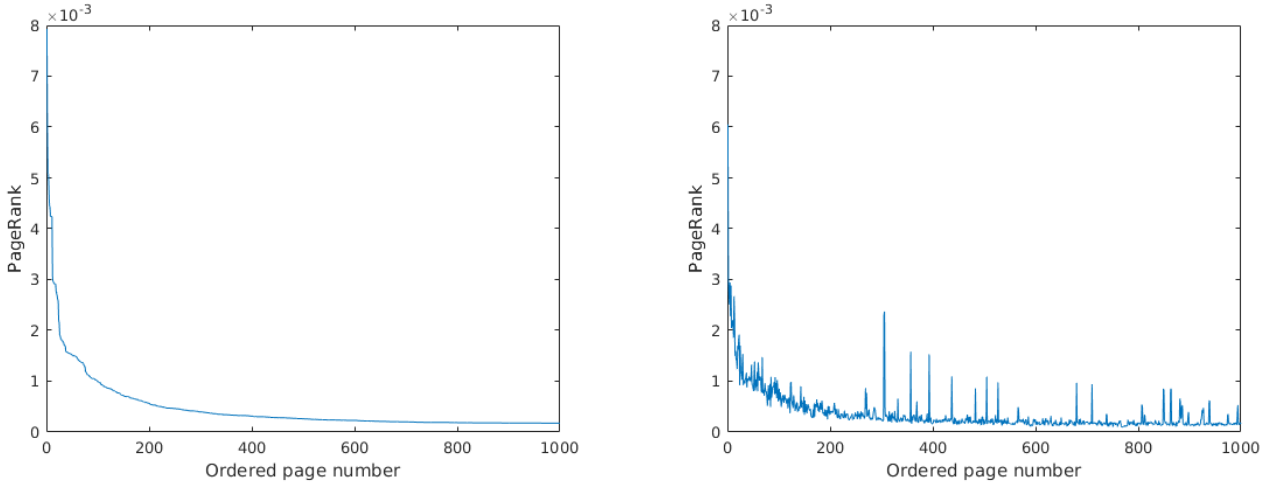


Figure 17: On the left the $\beta_{8227,5288}$ over time and on the right the π_{8227} over time.

In figure 17 the probability of adding a new link from P_{5288} to P_{8227} is being shown. At the start P_{5288} only has 2 outlinks. One can see the impact the PageRank of the webpage the potential link is directed at has on β.



(a) (b)

Figure 18: The average PageRank over 25 simulations of the first 1000 ordered pages (a) before and (b) after 50 steps.

Figure 18a is sorted using a permutation matrix $\mathbf{P}$. The same matrix $\mathbf{P}$ is used to order figure 18b. To correct for probabilistic noise the new PageRank values are the average PageRank values over 25 simulations.

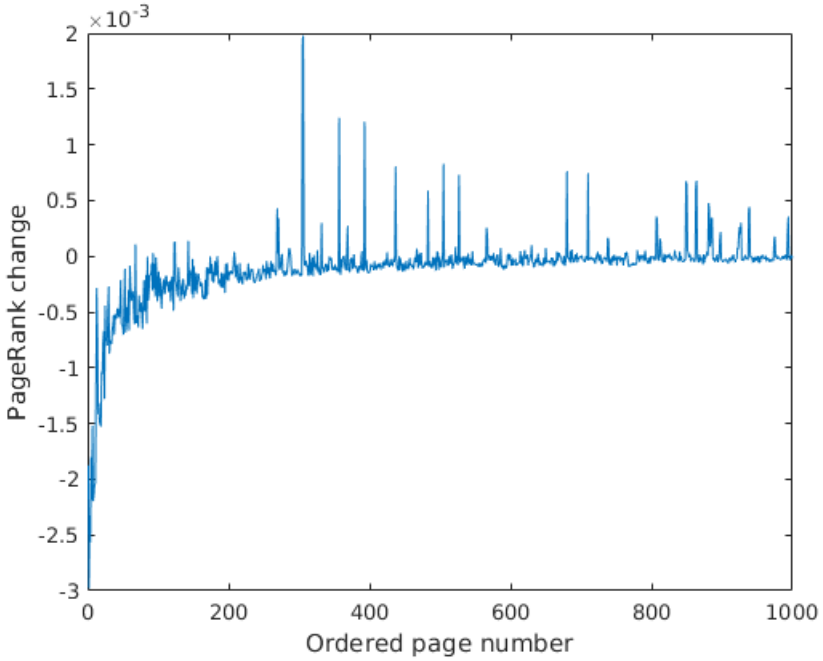


Figure 19: The change in PageRank after 50 steps averaged over 25 simulations.

Figure 19 gives a clearer view of the changes in PageRank. Some webpages have had a significant increase in PageRank, while the PageRank of the 200 pages with the highest PageRank almost all seem to have decreased. Notably the page with ordered number 306 has seen an increase of about 0.002, which is the biggest increase over all pages. P_{2264} still has the maximum PageRank, even though it lost about 0.002 in PageRank. One would certainly expect the page with high PageRank at the start to retain that position in the short-term. The highly ranked pages already have inlinks from other highly ranked pages. Since the probabilities for inlink removal are so low, the inlinks will remain for the most part. This means that, unless some inlink from a page with high PageRank is removed, the PageRank will remain more or less stable in the short-term. However, in the long-term the other pages will gain inlinks over time, possibly from highly ranked pages as well, hence this will increase their PageRank, likely at the expense of the highly ranked pages. The highly ranked pages already have inlinks from highly ranked pages at the start, while lower ranked pages probably do not. So the inlinks from active highly ranked pages to other highly ranked pages are likely to be more or less stable, while new inlinks from highly ranked pages to lower ranked pages will appear over time. This means a net loss of PageRank for pages with high PageRank at the start.

Looking at the correlation between the PageRank at the start and the change in PageRank a value of -0.8490 is obtained. This concurs with the observation that lower ranked pages will increase in PageRank at the expense of high ranked pages by gaining inlinks from higher ranked pages. Although it does not say anything about the quality of the inlinks, the correlation between the number of inlinks at the start and the change in PageRank is -0.7467. A high number of inlinks is associated with high PageRank, there being a correlation of 0.8278 between starting PageRank and number of inlinks, so indeed higher ranked pages seem to already have lots of inlinks from other pages while lower ranked pages do not. There is also a weak correlation of 0.3994 between starting outdegree and PageRank. In one of the next paragraphs it will be shown that there are some really active highly ranked pages.

Is there something that can be said about what determines which pages increase in PageRank? Since the values are an average over 25 simulations, it seems likely that there is something about those pages that showed a significant increase in PageRank that explains their success.

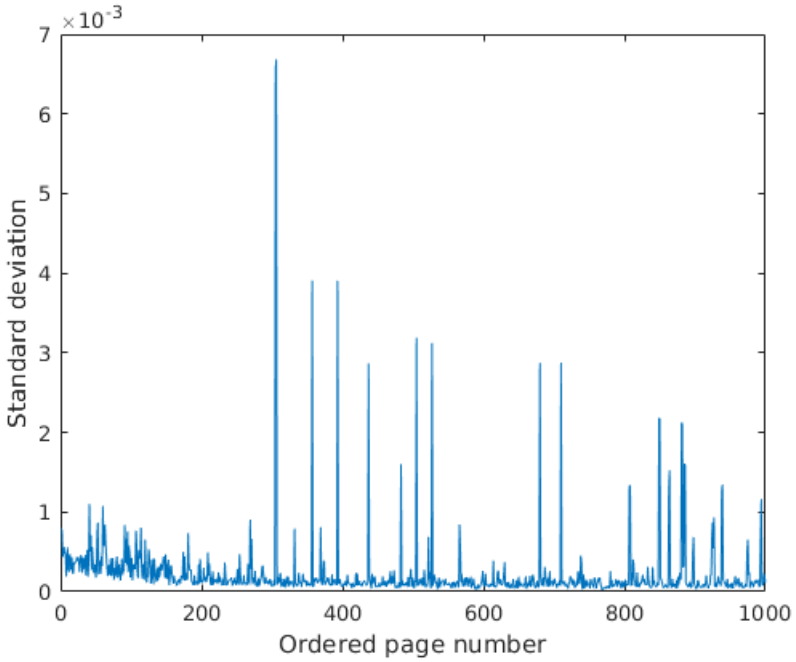


Figure 20: The standard deviation in PageRank of the first 1000 ordered pages over 25 simulations.

The standard deviation in figure 20 gives a good indication of how consistent the results of figure 19 are. The decline in PageRank in the highly ranked pages seems rather consistent. There are some pages with an extremely high standard deviation indicating that the results are not very consistent, although there could be something about those pages that increases the likelihood of big change in PageRank. Adding outlinks in a rational way according to each page's level of activity then leads to some high, albeit unstable, increase in PageRank for some pages.

Page	PageRank change	Standard Deviation
P_{4103}	0.0020	0.0067
P_{4102}	0.0020	0.0066
P_{6551}	0.0012	0.0039
P_{6540}	0.0012	0.0039
P_{2311}	0.0008	0.0032
P_{3325}	0.0008	0.0029
P_{5095}	0.0008	0.0029
P_{5094}	0.0007	0.0029
P_{2312}	0.0007	0.0031
P_{2512}	0.0007	0.0015

Table 3: The top 10 pages with the highest change in PageRank.

The first thing that stands in table 3 out is that there are three pairs of direct neighbours. This is likely because of the structure of the dataset. In the downward diagonal of figure 5 it can be seen that neighbours tend to be interconnected, hence any change in one of the pages will result in change in both pages. From the top 10 pages with the highest PageRank, 9 are in the top 10 of the pages with highest standard deviation in PageRank. So the high increase in PageRank only seems to happen in only some percentage of the simulations, but when it happens, the PageRank increases by a lot. For example, in one simulation P_{4103}'s PageRank increases to 0.0251, which is about 1/40 of the total available PageRank for the 9914 pages. The correlation between activity (number of outlinks) and change in PageRank is a rather weak -0.4136 over all pages. By looking at specific cases, however one can see that inactivity tends to pay off the most. Without exception, the pages of table 3 have only $1, 2$ or 3 outlinks, while the maximum over all pages is 277. Hence those page are not very active themselves and thus do not 'give away' much of their PageRank to other pages. There is also a strong positive correlation of 0.9558 between the number of inlinks at the end and the PageRank change, so pages like P_{4103} tend to gain a lot of inlinks. The quality of the inlinks determine of course the increase in PageRank. By looking at which pages tend to inlink to the pages of table 3, it may be possible to give a coherent explanation for their success. Table 3 shows that these pages have a rather high standard deviation in PageRank, so one should look for links that happen only a small percentage of the time. Looking at which new links occur in more than 20% of the simulations, the following table is obtained.

Page	Inlinking pages
P_{4103}	$P_{6837}, P_{7032}, P_{7033}$
P_{4102}	$P_{6837}, P_{7032}, P_{7034}$
P_{6551}	$P_{7032}, P_{7035}, P_{8057}$
P_{6540}	P_{6562}
P_{2311}	P_{9468}
P_{3325}	P_{6840}
P_{5095}	
P_{5094}	
P_{2312}	
P_{2512}	

Table 4: The new inlinks which occur at least in 6 simulations from the top 10 pages with the highest change in PageRank.

The inlinks for the last 4 pages in table 4 are are apparently not stable enough to show up, because there are no inlinks which are created in more than 20% of the simulations. Looking for example at P_{4103}, the inlinks are from pages which are highly ranked before as well as after 50 steps. The PageRanks of these pages are all in the top 20 before, and all in the top 30 after 50 steps. The consistent inlinks from the other pages almost all have a similarly high PageRank before and after. These pages must be rather active for these links to be created frequently, which is indeed the case. P_{6837} even has the maximal number of outlinks, while most of the other pages have outdegrees close to it. So the changes in PageRank can be explained by there being a relatively high probability of gaining an inlink from highly ranked active webpages, while the pages being linked to are themselves rather inactive. The probability for gaining an inlink is relatively high because the original PageRank of these pages is still in the top 1000 and not likely to decrease significantly over time, since these pages are not very active.

4.5 Subjective extended model

The extended model assumed that page owners made a rational choice about which webpages to add an outlink to. In reality however, page owners do not always make rational choices. Page owners may simply look at what outlinks other webpages have. If P_i has a high number of inlinks, the owner of page P_j will probably notice that many pages have an outlink to P_i. This may be an incentive to create an outlink from P_j to P_i as well. The reason for webpages to link to another webpage is then not determined by an objective factor like PageRank, but by a more subjective determinant of the quality of a page, namely the number of inlinks. This behaviour can be modelled by letting the probabilities depend on $|B(P_i)|(t)$ instead of $\pi_i(t)$. The more inlinks a webpage has at some point in time, the more likely it is that webpage owners will notice the hyperlinks and will add one on their own page as well. Similarly, the higher the number of inlinks P_i has, the higher the probability that P_j will keep its outlink to P_j. Still the activity of the webpage owner plays a big role. As was explained in the section about the extended model based on PageRank, some webpages are more active in adding outlinks than others. So $|P_j|$ also needs to be considered in the definition of α and β. This leads to the following definition.

Definition 20. The probabilities for keeping and adding an outlink from P_j to P_i at time t are defined as respectively

$$\begin{aligned} \alpha_{ij}(t) &:= 1 - \left(1 - \gamma \frac{|B(P_i)|(t)}{\max\limits_i |B(P_i)|(t)}\right) \delta \frac{|P_j|}{\max\limits_j |P_j|} \\ \beta_{ij}(t) &:= \frac{|B(P_i)|(t)}{\max\limits_i |B(P_i)|(t)} \varepsilon \frac{|P_j|}{\max\limits_j |P_j|} \end{aligned}$$

Similar to definition 19, α and β are well-defined. The same damping factors are being used to reduce the impact of the page owner activity and the subjective popularity on the probability of keeping a link. Notice that as the number of inlinks of P_i increases, α increases and β decreases.

In this case equation 1 becomes

$$P_{ij}(t + \Delta t) = \left[1 - \left(\delta + (\varepsilon - \gamma\delta) \frac{|B(P_i)|(t)}{\max\limits_i |B(P_i)|(t)}\right) \frac{|P_j|}{\max\limits_j |P_j|}\right] P_{ij}(t) + \frac{|B(P_i)|(t)}{\max\limits_i |B(P_i)|(t)} \varepsilon \frac{|P_j|}{\max\limits_j |P_j|}$$

Proposition 2 also holds for the subjective extended model. It is still necessary to calculate the α and β at each step however. One could use a formula for the expected number of inlinks for this, but that does not really simplify things.

4.6 Results of the subjective extended model

The `*.cs.stanford.edu` dataset will be used to illustrate the subjective extended model. The same values for the damping factors will be used ($\gamma = 0.95$, $\delta = 0.05$ and $\varepsilon = 0.1$).

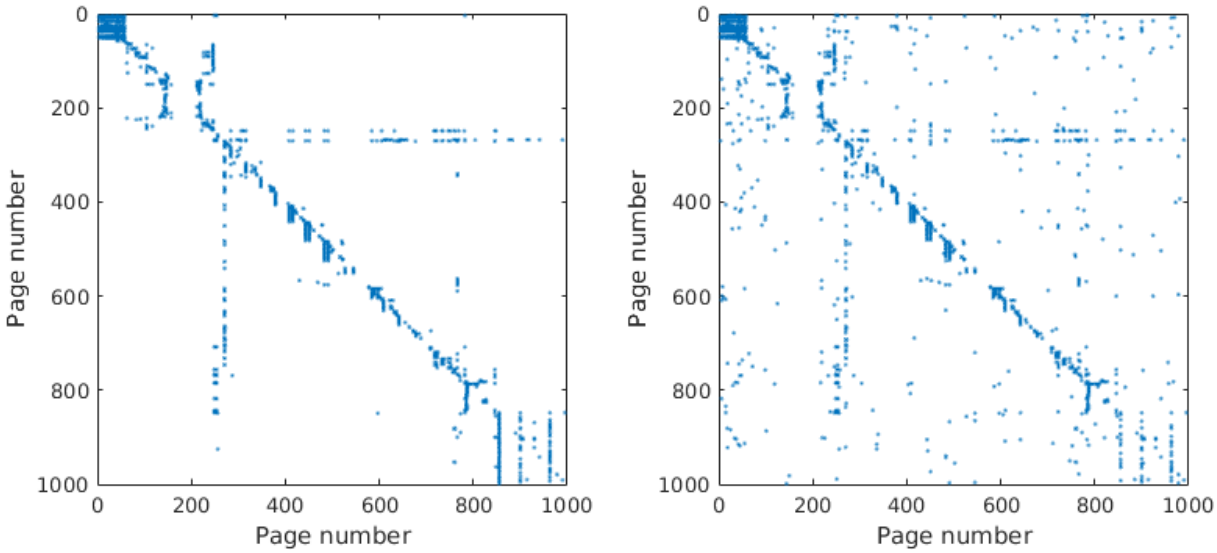


Figure 21: The links between the first 1000 pages before and after 50 iterations.

Figure 21 shows the new adjacency matrix for the first 1000 pages. It is not really possible to see whether or not pages with lots of inlinks in the original matrix have gained more inlinks.

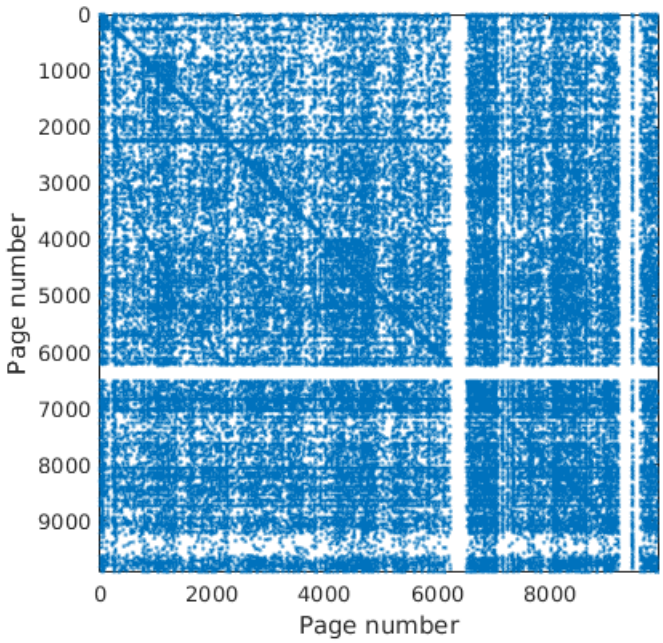


Figure 22: The adjacency matrix after 50 iterations.

Figure 22 shows that some pages have no inlinks. In the sparsity plot of the original matrix it can already be seen that some pages have no inlinks (e.g. P_{6300}). Normally these pages would not show up in the dataset, since Google explores the internet by following links, hence if a page has no inlink, it cannot be discovered. These pages are present in the Stanford dataset, however. The model is constructed such that $\beta_{ij} = 0$ if P_i has no inlinks, hence the pages still

have no inlinks after 50 iterations. After this particular iteration the matrix contains 84950 links, which is more than double the starting value of 35555.

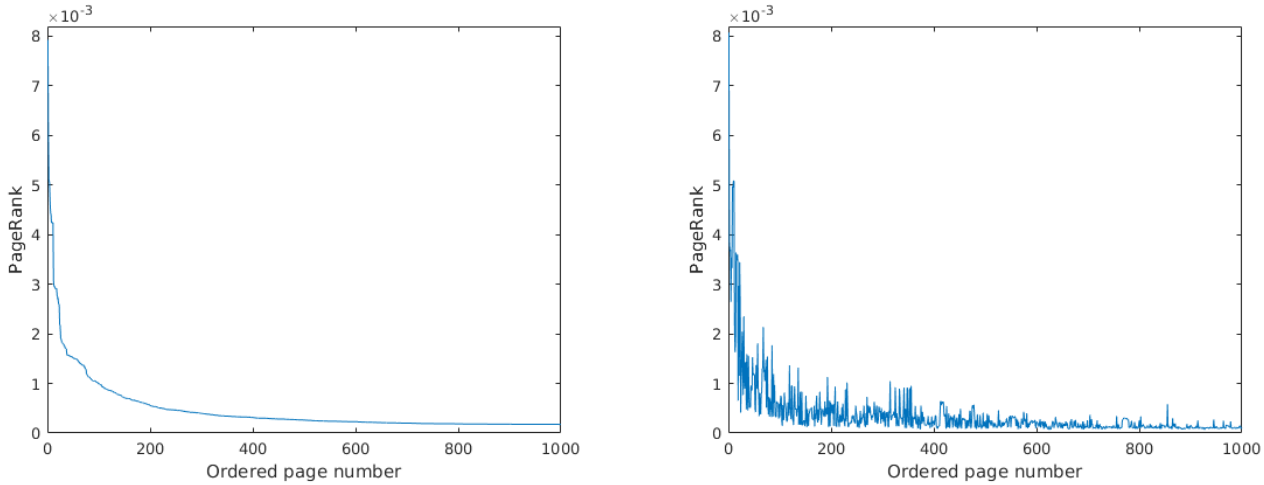


Figure 23: The average PageRank over twenty-five simulations of the first 1000 ordered pages (a) before and (b) after 50 steps.

Figure 23 shows the average PageRank over 25 simulations for the first 1000 pages ordered by the permutation matrix $\mathbf{P}$. The number of outlinks has more than doubled over these 50 iterations, averaging about 87487 links. P_{2264} still has the maximum PageRank, its PageRank having increased by about 0.0001.

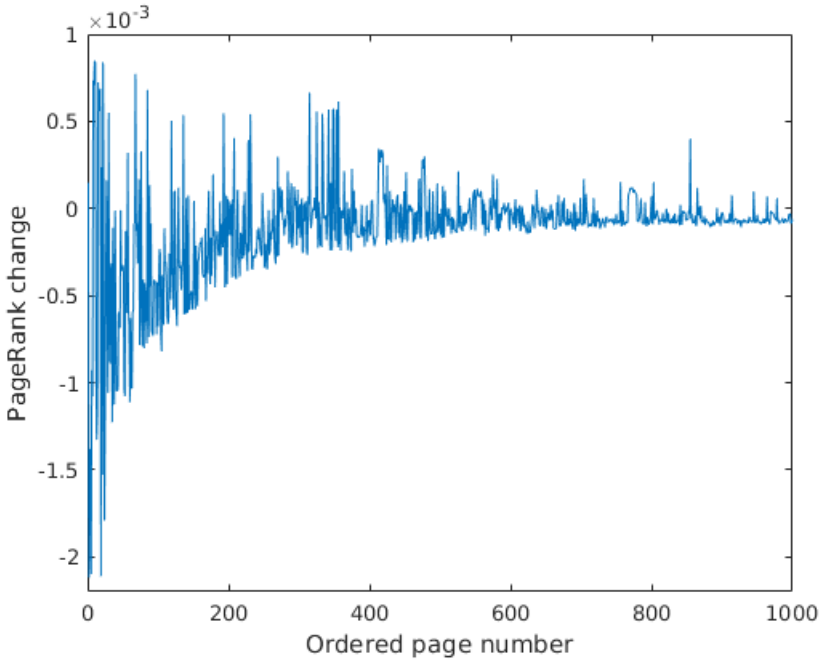


Figure 24: The change in PageRank after 50 steps averaged over 25 simulations.

The changes in PageRank are even better illustrated by figure 24. Most highly ranked pages

have decreased in PageRank (165 of the top 200). The change seems to be more random with respect to the original PageRank than in figure 19 of the objective extended model. This is also backed up by the data. The correlation between the PageRank change and the starting PageRank is only -0.5091, which is not all that strong. For the same reason as with the objective model, the highly ranked pages are likely to decrease in PageRank. However, the correlation between PageRank at the start and number of inlinks at the start is 0.8278. So higher ranked pages tend to have more inlinks, which means that they generally tend to gain more inlinks. The amount of new inlinks could compensate slightly for the observation that highly ranked pages must already have inlinks from other highly ranked pages, hence the weaker correlation. Keeping this in mind, it is no surprise to see that the top 10 pages with the highest positive change in PageRank have on average about 191 inlinks at the start. The correlation between number of inlinks at the start and PageRank change is a mere 0.0391. Indeed gaining lots of inlinks does not mean gaining quality inlinks. One tends to gain inlinks from the more active pages, but there is only a weak 0.3994 correlation between activity (number of outlinks) and PageRank, so the quality of those inlinks is likely not so great. The change in PageRank still depends on the activity of the page itself as well.

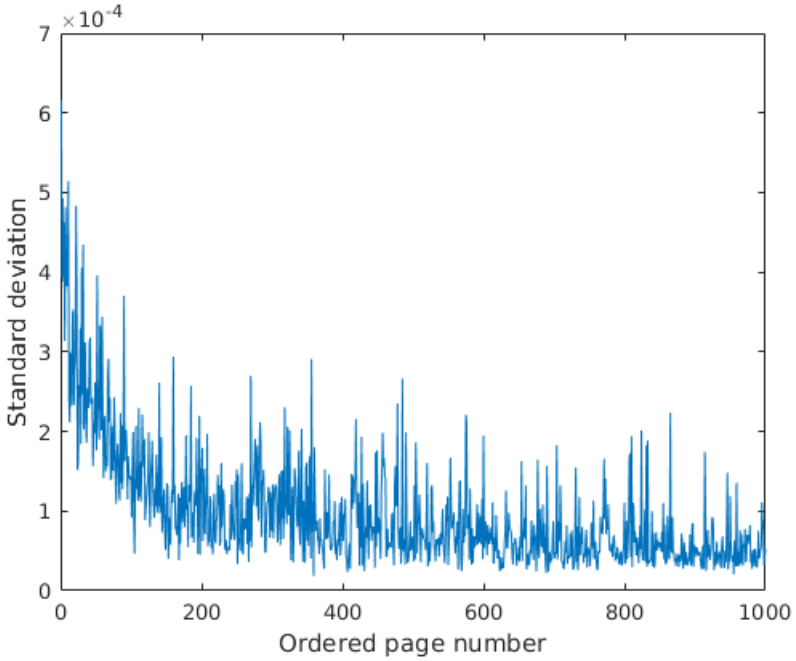


Figure 25: The standard deviation in PageRank of the first 1000 ordered pages over 25 simulations.

The standard deviation in the subjective model as seen in figure 25 is overall much lower than in the objective model. The maximal standard deviation is now about a factor 10^{-1} lower. So overall the values are more consistent. This is what can be expected. Most pages with a low number of inlinks will not gain many inlinks, so it is likely that their PageRank is relatively stable. The new inlinks are likely from active webpages, so in general the new inlinks come from a subset of active webpages. These links will for the most part be directed at pages with already a high number of inlinks, hence the PageRank will not be affected as much. In this model the behaviour of the number of inlinks of a page is more predictable than its PageRank, since it will likely stay roughly proportional to the starting situation. The stability of which links are

created is illustrated by the following fact. For the objective model the number of links which are created more than 20% of the time, equals 3515. For the subjective model this number is 4893.

4.7 Owner activity and PageRank change

Search engine optimization, the practice of trying to get one's webpage highly ranked in the search results, has been a hot topic ever since the birth of search engines. Search engine optimization concerning Google's PageRank is no exception. Using the subjective and objective extended model, what can be said about the influence of the changes in links on the PageRank for a particular page? The behaviour of the webpage is determined by the behaviour of the page owner and/or administrators. Owners or administrators choose to be active (or not), hence the PageRank can be seen as a function of owners'/administrators' behaviour. This presents an opportunity to investigate the PageRank using the models developed, by looking at what behaviour leads to what change in PageRank. The central question concerns the influence of webpage behaviour on its PageRank. The role of activity in PageRank will be looked into for both the subjective and objective model to see if it is better to be rational (PageRank based) in adding and removing links or to follow the crowd (inlink based). It will also be investigated whether activity should focus on adding or removing outlinks in order to consolidate the PageRank.

4.7.1 PageRank-based activity

The most influential factor that the page owner or administrator has control over that can affect the PageRank is activity. In the models discussed the activity was measured by the outdegree at the start. The obvious way to measure the effect of activity on PageRank is by artificially setting all activity to a low value and only setting the value of one page to some specific number. Then look at what happens to the PageRank of that one page. The activity parameter used for this is defined as

$$|P_j'| = \begin{cases} 1, & \text{if } |P_j| > 0 \text{ and } P_j \text{ is not a test page} \\ 0, & \text{if } |P_j| = 0 \text{ and } P_j \text{ is not a test page} \\ \eta, & \text{if } P_j \text{ is a test page} \end{cases}$$

One test page is picked to research the influence of activity on PageRank. To not let the activity of the other pages interfere too much with the results, the activity parameter is not divided by the maximum outdegree, but by 1000. Zero activity for the other pages would not be a realistic situation, since PageRank is heavily influenced by inlinks, which is indirectly influenced by activity as well. In the objective model a higher PageRank results in a higher probability for gaining inlinks, so activity of other pages should be taken into account. $0 \leq \eta \leq 1000$ is used as a parameter to control the activity of the test page. Note that this will not interfere with the links existing between pages, it is only used as an activity parameter. Now the α and β are defined as

$$\begin{aligned} \alpha_{ij}(t) &:= 1 - \left(1 - \gamma \frac{\pi_i(t)}{\max_i \pi_i(t)}\right) \delta \frac{|P_j'|}{1000} \\ \beta_{ij}(t) &:= \frac{\pi_i(t)}{\max_i \pi_i(t)} \varepsilon \frac{|P_j'|}{1000} \end{aligned}$$

For the damping factors the same values as before are used. As test pages the following pages were chosen:

Page	Inlinks	Outlinks	PageRank
P_{3718}	14	9	0.001296
P_{7485}	3	2	0.000065

Table 5: Test pages.

P_{3718} has a relatively high PageRank (top 100), and P_{7485} has an intermediate PageRank (rank 4004). The following table shows the average results for both pages over 10 simulations.

Page	η	PageRank	PageRank new	Difference	St. dev.	inlinks	outlinks
P_{3718}	0	0.001296	0.001190	-0.000106	0.000430	21.2	9
	500	0.001296	0.000567	-0.000729	0.000249	17	179.6
	1000	0.001296	0.000778	-0.000518	0.000251	17.6	205.6
P_{7485}	0	0.000065	0.000061	-0.000004	0.000004	3	2
	500	0.000065	0.000060	-0.000005	0.000011	3.2	184.6
	1000	0.000065	0.000067	0.000002	0.000010	3.6	225.8

Table 6: The PageRank before and after 50 steps using different η's averaged over 10 simulations and rounded off to 6 decimal places.

For the high-ranked page P_{3718} the results in table 6 are unanimous. Activity leads to a lower PageRank compared to no activity. Activity creates sort of a circle in the PageRank-based model. An active page likely creates a number of outlinks in the first few iterations. This leads to a PageRank drop. This lowers the probability of pages adding an outlink to the active page, hence the active page generally loses out on PageRank compared to the inactive case. And so the circle continues, more or less. This can be seen in the table as well. In the inactive test P_{3718} has on average more inlinks than in both active tests. What further stands out is that the PageRank loss is lower in the hyperactive test ($\eta = 1000$) than in the moderately active test ($\eta = 500$). The standard deviation is also about the same, so it does not seem to be caused by some extreme values. It may be accounted for by the fact that the average number of inlinks is slightly higher in the hyperactive test. A quality inlink could significantly boost the PageRank. The number of outlinks does not seem to matter that much beyond a certain number. It is hard to draw any definite conclusions from this though, since, for example, this also depends on whether or not the page has an outlink to your webpage as well. The activity that can be directly controlled by a webpage, adding or removing outlinks, seems to lead to lower PageRank. The main factor which is uncontrollable, inlinks, seems to have the most positive impact on PageRank for the tested pages. The effect of different strategies therefore seems to be overruled by uncontrollable factors, which explains why each strategy results in PageRank loss for the tested pages.

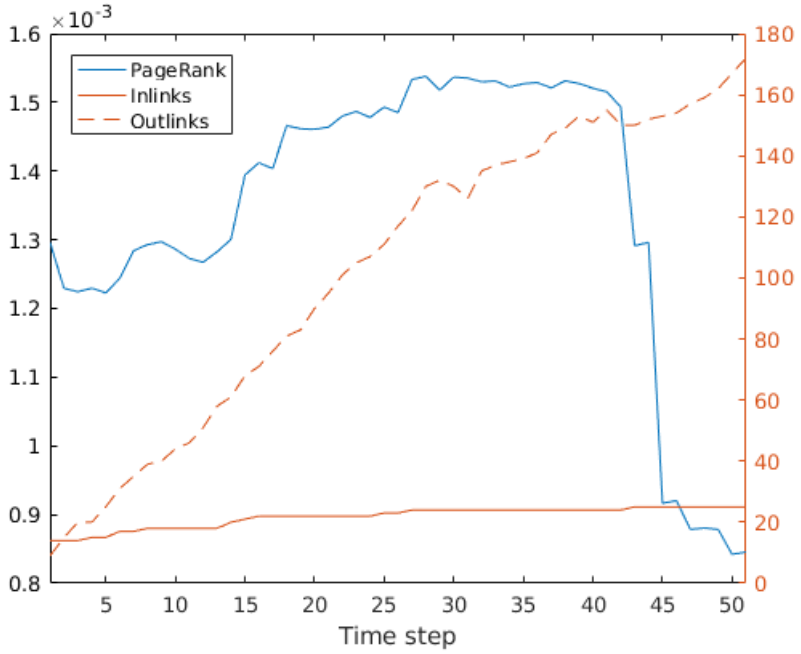


Figure 26: The progress of the number of inlinks and outlinks compared to the PageRank of P_{3718} over time in one realization using $\eta = 500$.

Note that figure 26 only shows the result of one particular simulation. For this figure $\eta = 500$ was used. One can see that each time an inlink is added there is a bump in PageRank. The steady growth of outlinks does not seem to impact the PageRank very much, or at least it seems overruled by the new inlinks. Around step 42 the PageRank plummets, while there is no significant change in inlinks or outlinks to be seen that could have caused it. It might have been caused by some important webpage removing its outlink to P_{3718}, although it seems unlikely that it would happen more than once. The number of inlinks stays the same, so at the same time a new inlink should have been created. The more probable explanation is that it has been caused by some major change outside of P_{3718} itself. An important inlink from a webpage P_k that has an outlink to P_{3718} could have been removed. This would cause P_k to decrease in PageRank, and hence decrease P_{3718}'s PageRank.

The results for P_{7485} are not that clear. The activity of lower ranked pages does not seem to make much of a difference. In the inactive case barely anything changes. In the active cases there is slightly more change. The indifference to activity could be explained by the fact that in some of the active simulations P_{7485} gained an inlink. P_{7485} only has 3 inlinks at the start, so a new inlink could mean a big increase in PageRank. One such inlink could compensate for all the new outlinks.

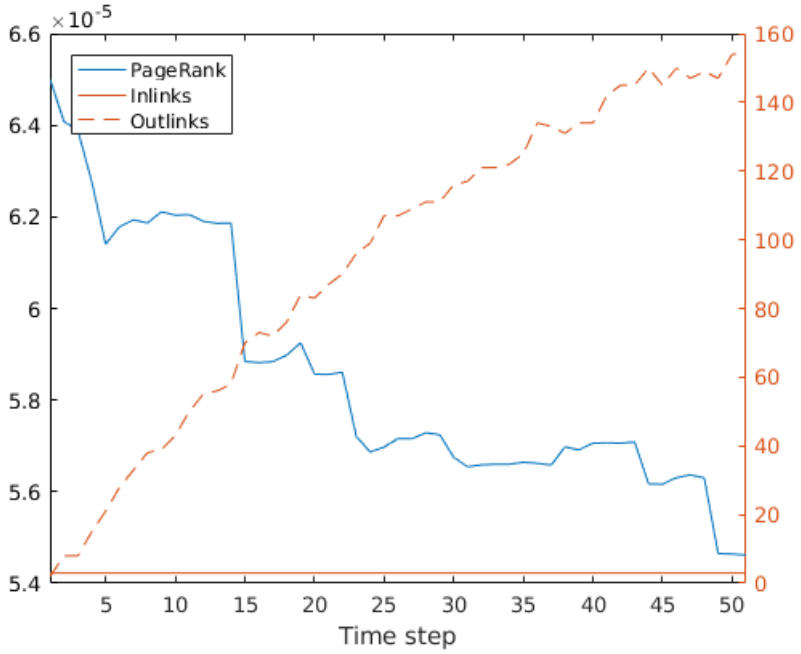


Figure 27: The progress of the number of inlinks and outlinks compared to the PageRank of P_{7485} over time in one realization using $\eta = 500$.

One simulation is pictured in figure 27. The number of inlinks stay stable over time. P_{7485} has a low starting PageRank which diminishes over time, hence there is a very low probability of a new inlink being added. There seems to be a significant negative correlation between change in PageRank and change in the number of outlinks. This makes it safe to say there is some sort of causal relation between the two, which is in accordance with the overall PageRank model.

In general it can be said that activity diminishes PageRank. Providing recommendations to other webpages diminished ones own PageRank, which also makes it less likely for a page to be backlinked to, hence activity is a double-edged sword. Especially webpages with a high PageRank seem to lose out. What was also observed is that after a certain number of outlinks, the PageRank is not influenced as much any more by new outlinks. On the other hand, highly ranked pages with a low number of outlinks are especially sensitive to activity. Big pages with already lots of outlinks, seem not to be influenced too much, which makes it possible for pages like Wikipedia to have a high PageRank.

4.7.2 Inlink-based activity

The influence of activity on PageRank will be investigated in a similar way using the subjective model. The same pages, P_{3718} and P_{7485}, will be used so as to be able to make a fair comparison between both models. The same definitions and values will be used to model the activity. In this case the probabilities will be defined as

$$\begin{aligned} \alpha_{ij}(t) &:= 1 - \left(1 - \gamma \frac{|B(P_i)|(t)}{\max_i |B(P_i)|(t)}\right) \delta \frac{|P'_j|}{1000} \\ \beta_{ij}(t) &:= \frac{|B(P_i)|(t)}{\max_i |B(P_i)|(t)} \varepsilon \frac{|P'_j|}{1000} \end{aligned}$$

For clarity, again the test pages used:

Page	Inlinks	Outlinks	PageRank
P_{3718}	14	9	0.001296
P_{7485}	3	2	0.000065

Table 7: Test pages.

The results are shown in the following table.

Page	η	PageRank	PageRank new	Difference	St. dev.	inlinks	outlinks
P_{3718}	0	0.001296	0.000950	-0.000346	0.000146	15.2	9
	500	0.001296	0.000931	-0.000365	0.000274	15.8	149
	1000	0.001296	0.000838	-0.000458	0.000246	15.4	176.8
P_{7485}	0	0.000065	0.000076	0.000011	0.000034	3.2	2
	500	0.000065	0.000056	-0.000009	0.000017	3	150.8
	1000	0.000065	0.000063	-0.000002	0.000012	3.4	182.4

Table 8: The PageRank before and after 50 steps using different η's averaged over 10 simulations and rounded off to 6 decimal places.

The results for P_{3718} seem mostly in accordance with the results from the objective model. More activity generally results in a lower PageRank. The difference between $\eta = 0$ and $\eta = 500$ seems rather small, but this can be explained by the difference in the average number of inlinks. With such a low number of inlinks, one additional inlink could compensate for a lot of outlinks. Since both test pages have a low number of inlinks, as can be expected from the subjective model, there is not much change in inlinks.

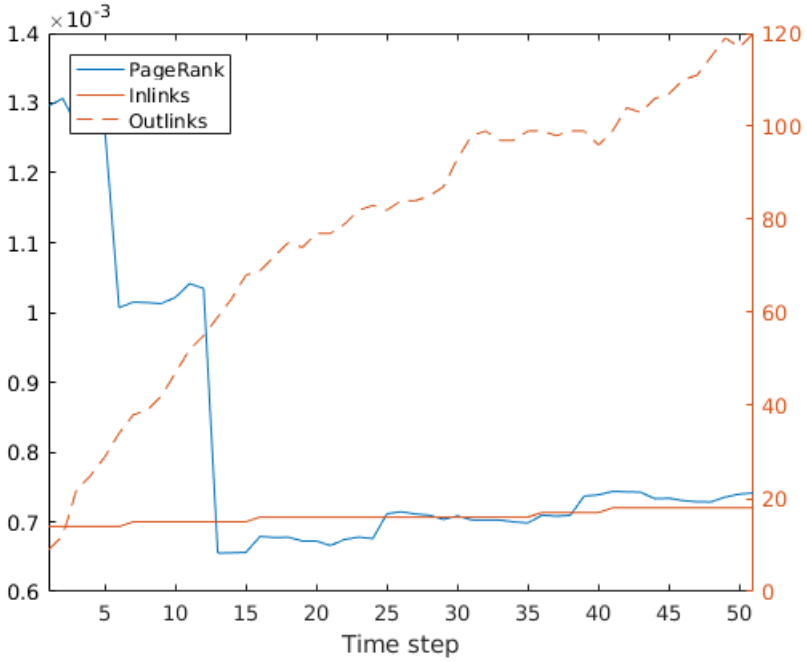


Figure 28: The progress of the number of inlinks and outlinks compared to the PageRank of P_{3718} over time in one realization using $\eta = 500$.

Figure 28 clearly shows the correlation between number of inlinks and change in PageRank in this particular simulation. Each time a page is added, there is a bump in PageRank. Just as in the previous figures, the sudden and steep declines in PageRank are likely to have been caused by change in PageRank from pages inlinking to P_{3718}.

With no activity, P_{7485} even shows an increase in PageRank. This is evidently caused by the two simulation in which a new inlink was created. A new inlink could mean a significant increase in PageRank, thereby creating two extreme PageRank values in the simulations. The relatively high standard deviation further supports this explanation. Again the decline in PageRank due to additional outlinks seems evident, albeit the only marginal decrease in PageRank for $\eta = 1000$ compared to $\eta = 500$. This, again, could be explained by the on average more inlinks that P_{7485} has.

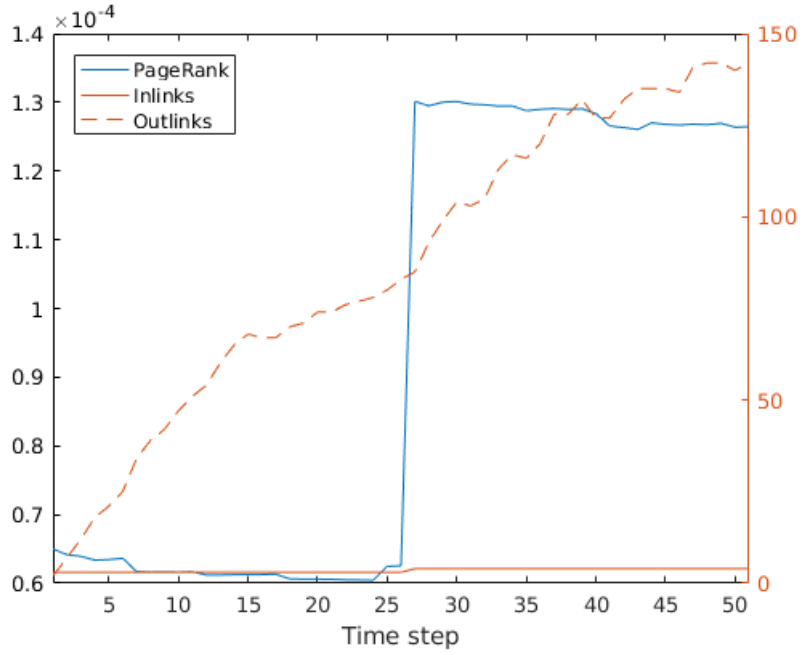


Figure 29: The progress of the number of inlinks and outlinks compared to the PageRank of P_{7485} over time in one realization using $\eta = 500$.

The particular simulation in figure 29 shows the enormous increase a low ranked webpage with a low number of inlinks endures when a new inlink is added. In step 26 a new link is added and the PageRank more than doubles in value. Figure 29 also illustrates the marginal impact outlinks have on a low ranked webpage. Only a minor decrease in PageRank is seen while the outdegree steadily increases.

All in all, the activity in the subjective model seems to have the same impact as in the objective case. The only real difference here is that the tested pages both have a low number of inlinks, hence not many inlinks are being added each simulation. This results in more PageRank on average, however it is not something a webpage itself can directly influence.

4.7.3 Adding or removing links

With the information from the previous sections a cautious prediction can be made about what activity a webpage should focus on to consolidate its PageRank. The results in the previous sections showed that a page's PageRank is negatively impacted by a higher number of outlinks. Hence increasing PageRank is achieved by minimizing the number of outlinks on the page, although the number of outlinks does not matter as much for pages with high numbers of outlinks. A small test will be devised to test whether or not the expectations will hold. The same setup will be used as in the previous sections. Only now the η factor will be split in η_1 and η_2, for α and β respectively. Now $|P'_j$ is defined as

$$|P'_j| = \begin{cases} 1, & \text{if } |P_j| > 0 \text{ and } P_j \text{ is not a test page} \\ 0, & \text{if } |P_j| = 0 \text{ and } P_j \text{ is not a test page} \\ \eta_1, & \text{if } P_j \text{ is a test page (used in } \alpha) \\ \eta_2, & \text{if } P_j \text{ is a test page (used in } \beta) \end{cases}$$

So the test page can use different activity values for outlink removal and link addition activity, denoted by η_1 and η_2 respectively. The test pages used can be seen in the following table.

Page	Inlinks	Outlinks	PageRank
P_{3718}	14	9	0.001296
P_{7485}	3	2	0.000065

Table 9: Test pages.

First the objective model will be tested. The results can be found in the following table.

Page	η_1	η_2	PageRank	PageRank new	Difference	St. dev.	inlinks	outlinks
P_{3718}	0	1000	0.001296	0.000839	-0.000457	0.000442	18.9	545.4
	500	500	0.001296	0.000567	-0.000729	0.000249	17	179.6
	1000	0	0.001296	0.000881	-0.000415	0.000322	17.9	0.9
P_{7485}	0	1000	0.000065	0.000067	0.000002	0.000019	3.2	564.1
	500	500	0.000065	0.000060	-0.000005	0.000011	3.2	184.6
	1000	0	0.000065	0.000060	-0.000005	0.000008	3.2	0.1

Table 10: The PageRank before and after 50 steps using different η's averaged over 10 simulations and rounded off to 6 decimal places.

The first thing that really stands out is the enormous amount of outlinks in the first and fourth row. In table 6 with $\eta = 1000$ – which means $\eta_1 = \eta_2 = 1000$ in this test – the number of outlinks is a little over 200, hence active link removal makes a big difference in that aspect. For P_{3718} there is not much difference in only focusing either on outlink removal or on addition. This could be caused by the 1 more outlink that occurs when only focusing on outlink addition. The standard deviation in that case is also higher, which means that it seems more of a hit-or-miss strategy than when solely focusing on link removal. The removal strategy seems in its own turn a better choice than the balanced one.

P_{7485} also shows a higher standard deviation when using the addition strategy. This reinforces the claim that it seems more like a hit-or-miss strategy. The removal strategy shows more consistency for both pages. Looking at both pages the removal strategy seems the best choice when trying to improve one's PageRank. The strategy is the most consistent overall

and for pages with a relatively high PageRank already it seems to be the best strategy to consolidate their rank. For pages with a relatively low PageRank there is not much difference in PageRank change among the strategies, however, based on the consistency of the removal strategy it still seems like the best choice. This also affirms what one would expect. Removing outlinks increases PageRank, which in turn increases the probability of gaining inlinks.

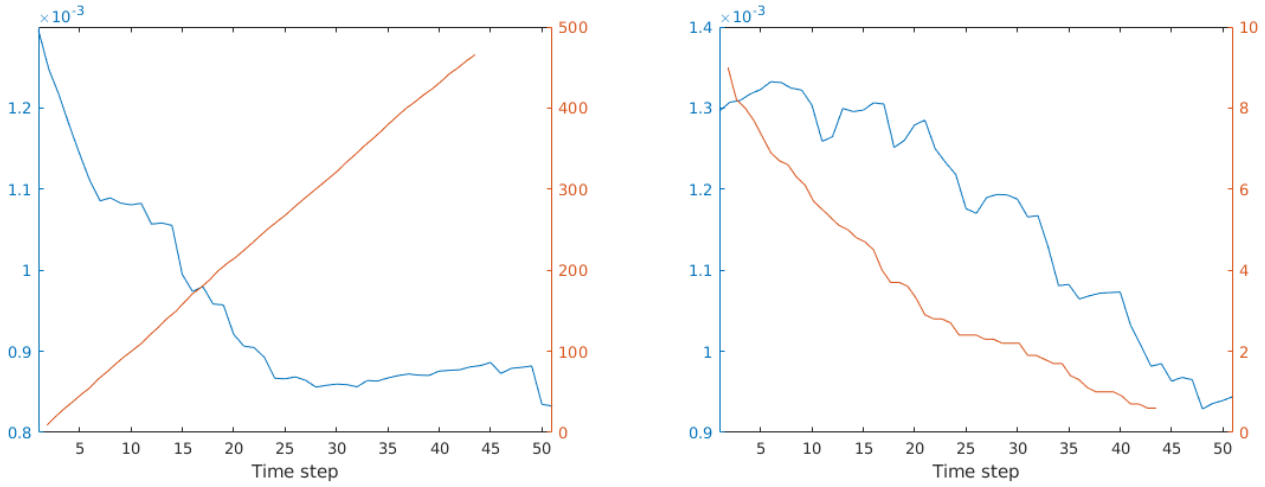
The results for the subjective model are shown in the following table, using the same activity parameters and test pages as before.

Page	η_1	η_2	PageRank	PageRank new	Difference	St. dev.	inlinks	outlinks
P_{3718}	0	1000	0.001296	0.000832	-0.000464	0.000222	15.1	465.8
	500	500	0.001296	0.000931	-0.000365	0.000274	15.8	149
	1000	0	0.001296	0.000944	-0.000352	0.000210	15.4	0.6
P_{7485}	0	1000	0.000065	0.000062	0.000003	0.000009	3.1	452.6
	500	500	0.000065	0.000056	-0.000009	0.000017	3	150.8
	1000	0	0.000065	0.000063	-0.000002	0.000011	3.5	0.1

Table 11: The PageRank before and after 50 steps using different η's averaged over 10 simulations and rounded off to 6 decimal places.

The consistency of the removal strategy in the objective model does not show up in the subjective model. For P_{3718} the removal strategy still seems like the best choice. The balanced strategy is very close, however, this may well be caused by the slightly higher number of inlinks which compensate for the outlinks. In the subjective model the probability for new inlinks is based on the number of inlinks present, hence one would expect the balanced strategy to be lower in PageRank. The higher number of inlinks just happened by chance.

For the lower ranked page, P_{7485}, the removal strategy seems the best one as well, however the average number of inlinks is higher than for the link addition strategy. So the addition strategy seems the best strategy here. In one of the previous sections the observation was made that additional outlinks did not seem to have much impact on low ranked pages. The results in this test seem to confirm this. One thing that is certain is that the balanced strategy loses out here.



(a) (b)

Figure 30: The correspondence between PageRank and outdegree for (a) the link addition strategy and (b) the link removal strategy.

Figure 30 nicely illustrates the short-term behaviour of the link addition and removal strategy. In figure 30a the link addition shows that in the short term the PageRank decreases due to outlink addition. In the long turn it still decreases, but this is likely to be governed by other factors. The removal strategy in figure 30b shows that in the short-term the PageRank increases, while in the long run it is probably mainly governed by other factors.

To conclude it can be said that if one wants to be active, one should generally focus on outlink removal. For low-ranked pages focusing on link addition is a viable possibility as well. At the end of section 4.7.2 it was concluded that activity generally was bad for PageRank, however the tests in that section all included activity in link addition as well as removal. In this section balanced activity was in almost all cases the worst option. Comparing these results it seems that no activity at all is overall the best choice, albeit not being a very satisfying result for a PageRank optimization strategy. For P_{3718} the average change over both models with no activity is about -0.000226 while for the removal strategy it is about -0.000384. For P_{7485} the changes are 0.000004 and -0.000004 respectively. Both results speak for the inactive strategy. The results achieved here do point in a certain direction, however, to really take any hard conclusions from this one should devise a more extensive test with a wider variety of pages and more simulations.

5 Conclusion and discussion

Change in the internet or a subset thereof was modelled by webpage behaviour due to owner/administrator activity. Owners or administrators have direct control over which links are on their webpage and thus can influence their page's PageRank. It was observed that the only actions webpage owners can do is add or delete an outlink. This was modelled in a probabilistic manner by using the formula

$$P_{ij}(t + n\Delta t) = \alpha P_{ij}(t + (n-1)\Delta t) + \beta[1 - P_{ij}(t + (n-1)\Delta t)]$$

for the probability of there being a link from P_j to P_i. First in the basic model, which assumed that activity and thus the α and β are equal for each webpage. All models were tested using numerical methods, since it was not possible to use analytic methods for extensively testing the models. The results for the basic model were rather uninteresting, since the links were added and removed randomly. The extended models took more information into account. Activity was then based on the activity level of a webpage itself, measured by the ratio between the number of outlinks on the page and the maximum number of outlinks over all pages. In the objective extended model this was combined with the assumption that owners/administrators are rational. They would not just randomly add or remove a link, but they would base their decisions on the importance of the other webpage, measured by its PageRank. In the subjective model the decisions were made based on the number of inlinks a webpage has. The more inlinks a webpage has, the likelier it is that the owner/administrator would notice that page and the more popular it would appear to be to him. This in turn makes it more likely that the owner/administrator would add a link to that page, and less likely to remove an existing one to that particular page. So in the extended models the changes were based on choices made by the owner/administrator of a webpage.

Both the extended models were used to test what the best strategy would be to optimize the PageRank of a webpage. First results indicated that no activity would be the best choice. In later tests a different value for addition and removal activity was used. In this test the best strategy was only being active in removing links. Combining these results it was concluded that the inactive strategy was still better than the removal strategy, since it usually leads to a higher PageRank. The tests showed a connection between outlinks and PageRank, the first negatively influencing the latter, although low ranked pages seemed not to be affected as much by additional outlinks. The tests used a very limited set of pages and activity values as well as a modest number of simulations. For more reliable test results a more extensive testing is necessary.

In order to more realistically model webpage activity and hence network change, one could look at a model which incorporates hype-based activity. A hype is a short-lived phenomenon in which a certain webpage or webpages are extremely popular, hence gaining lots of inlinks over a short time span. This could be modelled by basing the probabilities on change in the number of inlinks for a webpage. A webpage with lots of new inlinks in a short time is a hyped webpage and hence even more webpages will add a link to that webpage, until the hype wears off. Another addition could be static pages. What happens often is that pages are active at the start, but stay the same afterwards. Possibly some criterion can be found to identify these pages and model their behaviour realistically. Another possible addition to the model is the identification of *hubs*. A hub is a webpage with lots of outlinks, which could indicate that the webpage is a sort of catalogue of information, meaning that this webpage is the go-to page for all information on a certain topic. For instance, a Wikipedia entry on

a certain topic contains lots of outlinks. These outlinks are (mostly) to pages whose content has a direct connection with the topic of the Wikipedia entry. It is possible to group the webpages in a network around hubs. Then, for example, the probability of adding a link to a page connected to the same hub (or hubs) is higher than adding one to a page which does not have a hub in common, because these pages likely contain similar or related content. Similarly *authoritative* pages can be identified. An authoritative page is a page with lots of inlinks. The inlinks indicate that the page is an authority on a topic. Pages linking to this authoritative page likely contain similar or related content, so groupings can be made around authoritative pages in the same way as hubs

6 References

References

[1] A.N. Langville, and C.D. Meyer. *Google's PageRank and Beyond*. Princeton University Press. 2006.

[2] S. Brin, and L. Page. *The Anatomy of a Large-scale Hypertextual Web Search Engine*. Stanford University, Stanford.

[3] C.D. Meyer. *Sensitivity of the Stationary Distribution of a Markov Chain*. SIAM Journal on Matrix Analysis and Applications, vol. 15, no. 3. 1994.

[4] A.N. Langville and C.D. Meyer. *Updating Markov Chains with an Eye on Google's PageRank*. SIAM Journal on Matrix Analysis and Applications, vol. 27, no. 4. 2005.

[5] K. Avrachenkov and N. Litvak. *The Effect of New Links on Google PageRank*. Stochastic models, vol. 22, issue 2. 2006.

[6] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. *Resilience of the Internet to Random Breakdowns*. Physical Review Letters, vol. 85, no. 21. 2000.

[7] J. Gao, B. Barzel, A.L. Barabási. *Universal Resilience Patters in Complex Networks*. Nature, vol. 530. 2016.

[8] R. Larson, B.C. Edwards, and D.C. Falvo. *Elementary Linear Algebra*. Fifth Edition. Houghton Mifflin. 2003.

[9] A.Y. Ng, A.X. Zheng, and M.I. Jordan. *Link Analysis, Eigenvectors and Stability*. International Joint Conference on Artificial Intelligence. 2001.

[10] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web*. John Wiley & Sons. 2003.

[11] R. Pastor-Satorras, and A. Vespignani. *Evolution and Structure of the Internet*. Cambridge University Press. 2007.

[12] R. Lempel, and S. Moran. *Rank-Stability and Rank-Similarity of Link-Based Web Ranking Algorithms in Authority-Connected Graphs*. Information Retrieval, vol. 8, issue 2. 2005.

[13] D. Gleich. *The `*.cs.stanford.edu` Matrix*. `http://www.cise.ufl.edu/research/sparse/matrices/Gleich/wb-cs-stanford.html`

A Appendix – Matlab code

A.1 General PageRank files

Matlab functions used for computations concerning the PageRank algorithm and building a dataset.

A.1.1 surfer.m

```
function [U,G] = surfer(root,n) % by Cleve Moler
% SURFER  Create the adjacency graph of a portion of the Web.
%    [U,G] = surfer(root,n) starts at the URL root and follows
%    Web links until it forms an adjacency graph with n nodes.
%    U = a cell array of n strings, the URLs of the nodes.
%    G = an n-by-n sparse matrix with G(i,j)=1 if node j is linked
    to node i.

clf
shg
set(gcf,'doublebuffer','on')
axis([0 n 0 n])
axis square
axis ij
box on
set(gca,'position',[.12 .20 .78 .78])
uicontrol('style','frame','units','normal','position',[.01 .09 .98
    .07]);
uicontrol('style','frame','units','normal','position',[.01 .01 .98
    .07]);
t1 = uicontrol('style','text','units','normal','position',[.02 .10
    .94 .04], ...
    'horiz','left');
t2 = uicontrol('style','text','units','normal','position',[.02 .02
    .94 .04], ...
    'horiz','left');
slow = uicontrol('style','toggle','units','normal', ...
    'position',[.01 .24 .07 .05],'string','slow','value',0);
quit = uicontrol('style','toggle','units','normal', ...
    'position',[.01 .17 .07 .05],'string','quit','value',0);

U = cell(n,1);
hash = zeros(n,1);
G = logical(sparse(n,n));
m = 1;
U{m} = root;
hash(m) = hashfun(root);

j = 1;
while j < n && get(quit,'value') == 0

```

```
37    % Try to open a page.
38    try
39       set(t1,'string',sprintf('%5d %s',j,U{j}))
40       set(t2,'string','');
41       drawnow
42       page = urlread(U{j});
43    catch
44       set(t1,'string',sprintf('fail: %5d %s',j,U{j}))
45       drawnow
46       continue
47    end
48    if get(slow,'value')
49       pause(.25)
50    end
51
52    % Follow the links from the open page.
53    for f = findstr('http:',page);
54       % A link starts with 'http:' and ends with the next quote.
55
56       e = min([findstr('"',page(f:end)) findstr('''',page(f:end))]);
57       if isempty(e), continue, end
58       url = deblank(page(f:f+e-2));
59       url(url<' ') = '!';   % Nonprintable characters
60       if url(end) == '/', url(end) = []; end
61
62       % Look for links that should be skipped.
63
64       skips = {'.gif','.jpg','.pdf','.css','lmscadsi','cybernet',
             ...
65                'search.cgi','.ram','www.w3.org', ...
66                'scripts','netscape','shockwave','webex','fansonly','
                     ogp.me',...
67                'youtube.com/embed','xmlrpc','\','<'};
68       skip = any(url=='!') | any(url=='?') | any(url(end)=='/');
69       k = 0;
70
71       while ~skip && (k < length(skips))
72          k = k+1;
73          skip = ~isempty(findstr(url,skips{k}));
74       end
75       if skip
76          if isempty(findstr(url,'.gif')) && isempty(findstr(url,'.
                jpg'))
77             set(t2,'string',sprintf('skip: %s',url))
78             drawnow
79             if get(slow,'value')
80                pause(.25)
81             end
82          end
```

```
83            continue
84        end
85
86        % Check if page is already in url list.
87        i = 0;
88        for k = find(hash(1:m) == hashfun(url))';
89            if isequal(U{k},url)
90                i = k;
91                break
92            end
93        end
94
95        % Add a new url to the graph if there are fewer than n.
96        if (i == 0) && (m < n)
97            m = m+1;
98            U{m} = url;
99            hash(m) = hashfun(url);
100           i = m;
101       end
102
103       % Add a new link. j->i
104       if i > 0
105           G(i,j) = 1;
106           set(t2,'string',sprintf('%5d %s',i,url))
107           line(j,i,'marker','.','markersize',6)
108           drawnow
109           if get(slow,'value')
110               pause(.25)
111           end
112       end
113   end
114   j = j+1;
115 end
116 delete(t1)
117 delete(t2)
118 delete(slow)
119 set(quit,'string','close','callback','close(gcf)','value',0)
120
121 %-------------------------
122
123 function h = hashfun(url)
124 % Almost unique numeric hash code for pages already visited.
125 h = length(url) + 1024*sum(url);
```

A.1.2 danglingnode.m

```
function [ y ] = danglingnode( H )
%   Calculates the danglingnode vector for a row substochastic
    matrix H
y= (sum(H,2)==0);
end
```

A.1.3 hyperlinkmatrix.m

```
function [ y ] = hyperlinkmatrix( A )
%   Gives row-substochastic matrix H based on a matrix A such that
    A(i,j)=1 if there is a link from j to i
n=length(A);

%compute outdegrees of nodes
x=sum(A,1);

% build matrix H
H=sparse(double(A));
for j=1:n
    if x(j)~=0
        H(:,j)=H(:,j)./x(j);
    end
end
y=transpose(H);
end
```

A.1.4 powermethod.m

```
function [ pi ] = powermethod(H,alpha,v,pi0,error)
%   Calculates the PageRank vector using the power method
n=length(H);
e=ones(n,1);
d=danglingnode(H);

if(nargin<5)
    error =1e-5;
end
if(nargin<3)
    v=e/n;
end
if (nargin <4)
    pi0 =v;
end
if(nargin<2)
    alpha=0.85;
end
```

```
K=alpha*H';
pi=pi0;
steps=100;
for i=1:steps
    temp=pi;
    pi=K*pi+(1-alpha+alpha*sum(d.*pi))*v;
    if(sum(abs(pi-temp))<error)
        break;
    end
end
end
```

A.1.5 permutationmatrix.m

```
function [ P ] = permutationmatrix( vector )
%   Computes the permutation matrix P such that P*vector is a row
    vector sorted in descending order
%   Convert to column vector
vector=vector(:);
m=length(vector);

%   Obtain indices of pages in sorted vector
[~, index]=sort(vector,'descend');

%   Use these indices to compute P
P=sparse(m,m);
    for i=1:m
        P(i,index(i))=1;
    end
end
```

A.2 Probabilistic model files

Matlab functions used for computations concerning the probabilistic model for link changes.

A.2.1 basicmodel.m

```
function [ A ] = basicmodel( A,alpha,beta )
%   Compute new matrix at the next time step for the basic model
%   Each element: P_ij(t+delta t)= alpha*P_ij(t)+beta*[1-P_ij(t)]
if(nargin<2)
    alpha=0.95;
    beta=0.05;
end

m=length(A);
M1=binornd(1,alpha,m,m);
M0=binornd(1,beta,m,m);

A=A.*M1+(A==0).*(M0-diag(diag(M0)));
%   Display the number of links in the new matrix
display(['#outlinks: ',num2str(sum(sum(A)))]);
end
```

A.2.2 objectivemodel.m

```
function [ A ] = objectivemodel( A,tempalphas,tempbetas,gamma)
%   Calculates matrix at the next time step using alpha_ij(t) and
    beta_ij(t) which are PageRank based
%   This function is used for the extended objective model
n=length(A);

if (nargin<4)
    gamma=0.95;
end

%   Calculate PageRank first
H=hyperlinkmatrix(A);
pi=powermethod(H);

%   Calculate the relative PR for each row
temp=pi./max(pi);

%   Calculate P(1->1) for the whole matrix
alphas=1-(1-gamma*temp)*tempalphas;

%   Calculate the P(0->1) for the whole matrix
betas=temp*tempbetas;

%   Compute changes P(1->1)
```

```
OnetoOne=binornd(1,alphas.*A,n,n);
%    Compute P(0->1)
ZerotoOne=binornd(1,betas,n,n);

%    Compute changes in A
A=A.*OnetoOne+(A==0).*(ZerotoOne-diag(diag(ZerotoOne)));

%    Display new number of links in matrix
display(['#outlinks: ',num2str(sum(sum(A)))]);
end
```

A.2.3 subjectivemodel.m

```
function [ A ] = subjectivemodel( A,tempalphas,tempbetas,gamma)
%    Calculates matrix at the next time step using alpha_ij(t) and
    beta_ij(t) which are inlink based
%    This function is used for the extended subjective model
n=length(A);

if (nargin<4)
    gamma=0.95;
end

%    Calculate relative number of inlinks of each page
temp=sum(A,2);
temp=temp./max(temp);

%    Calculate P(1->1) for each column
alphas=1-(1-gamma*temp)*tempalphas;

%    Calculate the P(0->1) for each column
betas=temp*tempbetas;

%    Compute changes P(1->1)
OnetoOne=binornd(1,alphas.*A,n,n);
%    Compute P(0->1)
ZerotoOne=binornd(1,betas,n,n);

%    Compute changes in A
A=A.*OnetoOne+(A==0).*(ZerotoOne-diag(diag(ZerotoOne)));

%    Show new number of links in matrix
display(['#outlinks: ',num2str(sum(sum(A)))]);
end
```