Web Graph and PageRank algorithm

Danil Nemirovsky

Department of Technology of Programming,
Faculty of Applied Mathematics and Control Processes,
St. Petersburg State University,
Universitetskii prospekt 35,
Peterhof, St Petersburg 198504, Russia
nemd@rambler.ru

Abstract

The pages and hyperlinks of the World-Wide Web may be viewed as nodes and arcs in a directed graph. This graph has about a billion nodes today, several billion links, and appears to grow exponentially with time. Known facts about macroscopic structure, diameter and in-degree and out-degree distributions of the graph are reviewed. The PageRank as another way of characterizing structure of the Web graph is considered. Power method, decomposition, and aggregation/disaggregation method computing of PageRank are recalled.

Key words: Web, Web graph, Markov process, PageRank, Decomposition, Aggregation, Disaggregation.

1 Introduction

The World-Wide Web has spawned a sharing and dissemination of information on an unprecedented scale. Hundreds of millions - soon to be billions - of individuals are creating, annotating, and exploiting hyperlinked content in a distributed fashion. These individuals come from a variety of backgrounds and have a variety of motives for creating the content. The pages and hyperlinks of the World-Wide Web may be viewed as nodes and arcs in a directed graph. There are many reasons for studying the evolution of this graph. We review a number of measurements and properties of the graph, such macroscopic structure, diameter and in-degree and out-degree distributions. The another way to characterize structure of the Web graph is PageRank. PageRank is the method of finding page authorities produced by the Web graph structure. Power method, decomposition, and aggregation/disaggregation method are recalled.

In **Section 2** we review the Web graph and its characteristics. Basic terminology of graph theory is placed. Macroscopic structure, diameter, and in-out-degree distributions are described. In **Section 3** we introduce Markov theory needed for defining PageRak. In **Section 4** we define PageRank and place its properties. In **Section 5** we consider decomposition of PageRank and different special cases. In **Section 6** we describe aggregation/disaggregation method.

2 Web graph

2.1 Basic terminology of graph theory

The reader familiar with basic notions from graph theory may skip this primer.

Definition 1 (Directed graph) *A **directed graph** G is a pair $G = (V, E)$, where V is a set of any nature, elements of which are called nodes, E is a set of ordered pairs (u, v) called arcs.*

Definition 2 (In-degree and out-degree) *The **out-degree** of a node u is the number of distinct arcs $(u, v) \in E$, and the **in-degree** is the number of distinct arcs $(v, u) \in E$.*

Definition 3 (Path) *A **path** from node u to node v is a sequence of arcs $(u, u_1), (u_1, u_2), \ldots, (u_k, v)$, where $(u, u_1), (u_i, u_{i+1}), (u_k, v) \in E, \forall i = \overline{1, k-1}$.*

Definition 4 (Strongly connected component) *A **strongly connected component** (strong component for brevity) of a graph $G = (V, E)$ is a set of nodes such that for any pair of nodes u and v in the set there is a path from u to v.*

Definition 5 (Diameter) *A **diameter** of a graph $G = (V, E)$ is the maximum over all ordered pairs (u, v) of the shortest path from u to v.*

2.2 Definition of the Web graph

Lets define a directed graph corresponding to the Web. Pages represent nodes and hyperlinks between pages represent arcs. Hence, we defined a directed graph called the Web graph. There are many pages in the Web and, therefore, the Web graph is large and has complex structure.

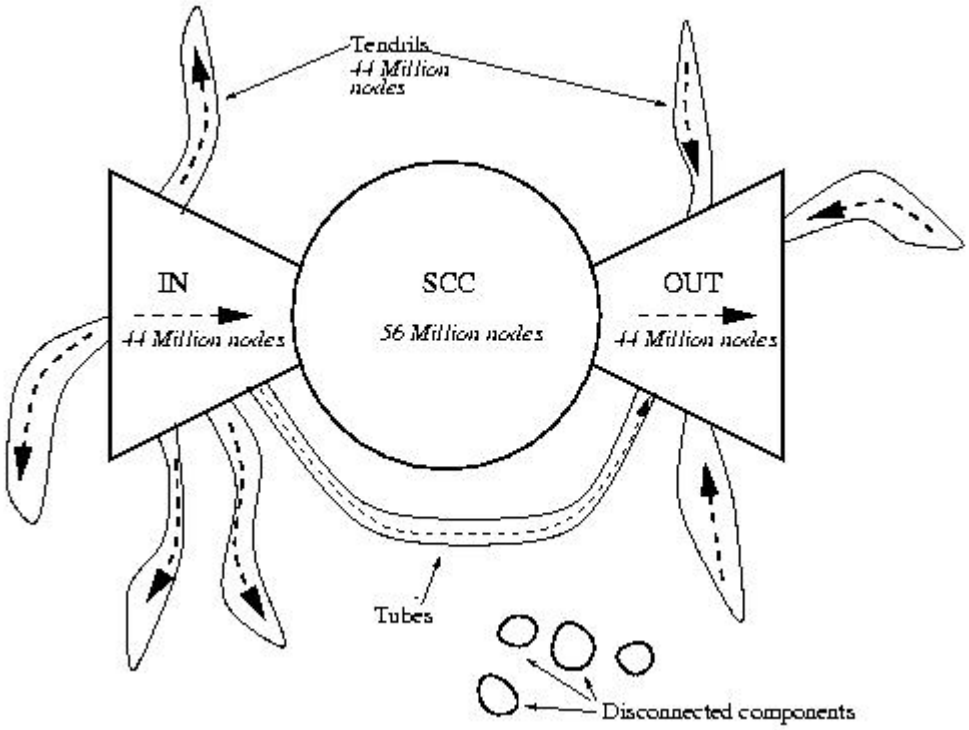

Figure 1. The macroscopic structure of the Web.

2.3 Structure of the Web graph

The structure of the Web graph is considered in [5]. The web macroscopic structure is represented on Figure 1. This connected web breaks naturally into four pieces. The first piece is a central core, all of whose pages can reach one another along directed hyperlinks – this "giant strongly connected component" (SCC) is at the heart of the web. The second and third pieces are called IN and OUT. IN consists of pages that can reach the SCC, but cannot be reached from it - possibly new sites that people have not yet discovered and linked to. OUT consists of pages that are accessible from the SCC, but do not link back to it, such as corporate websites that contain only internal links. Finally, the TENDRILS contain pages that cannot reach the SCC, and cannot be reached from the SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE – a passage from a portion of IN to a portion of OUT without touching SCC.

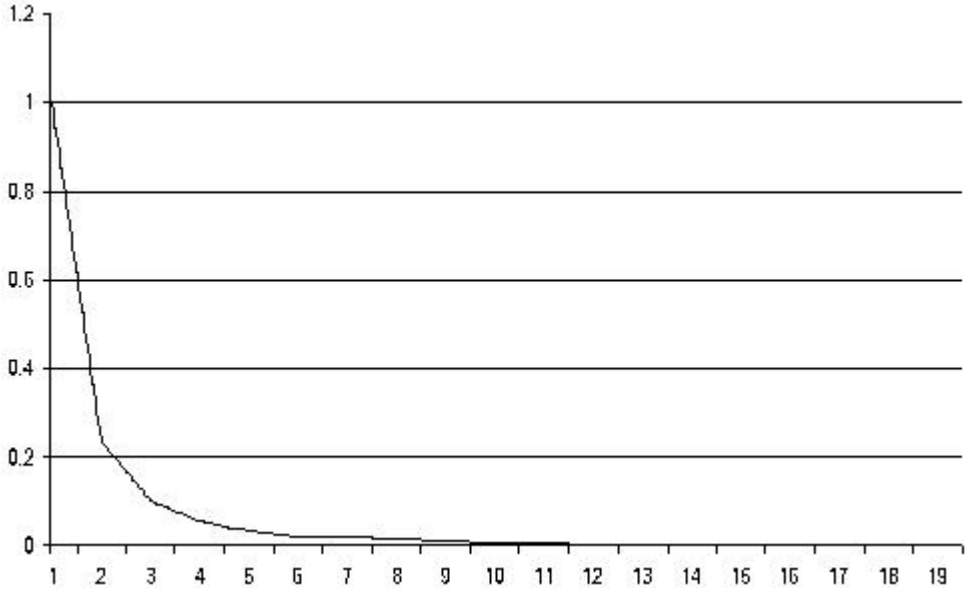

Figure 2. In-degree distribution.

2.4 Diameter of the Web graph

The diameter of SCC is at least 503, but is probably substantially more than this: unless a short tube connects the most distant page of IN to the most distant page of OUT without passing through the SCC, the maximum finite shortest path length is likely to be close to 905.

2.5 In- and out-degree distributions

In- and out-degree distributions on the Web are distributed according to power laws, e.g. the probability that a node has in-degree (out-degree) i is proportional to $(\frac{1}{i})^x$, for some positive $x > 1$. In the case of in-degree, the exponent of the power law is around 2.1 (Figure 2), and, in the case of out-degree, the exponent of the power law is 2.72.

3 Markov theory

We will consider a PageRank algorithms below and and we introduce Markov theory in order to fix notation.

3.1 Markov processes

Definition 6 (Markov process) *An S-valued **Markov process** is an infinite sequence of random variables $X_k = X_0, X_1, \ldots \in S$ if S is finite and the probability function P satisfies: $P(X_{k+1} = b | X_0 = a_0, \ldots, X_k = a_k) = P(X_{k+1} = b | X_k = a_k)$ is the same for all $k \geqslant 0$.*

*Its **transition function** is $\omega(a, b) = P(X_{k+1} = b | X_k = a)$.*

*Its **initial distribution** is $\sigma(a) = P(X_0 = a)$.*

In the Stochastic processes literature, this is technically called a homogeneous, discrete time, finite space Markov process. In applications of the theory, they are often simply called Markov processes or Markov chains.

3.2 Convergence of Markov processes

In this section, we review the conditions under which $\lim_{k\to\infty} P(X_k = a)$ converges.

Most of the Markov processes we will discuss have a nice property called ergodicity. To define this, we need to define the period of a state and irreducibility first. Intuitively, if the only way from a state back to itself is through a cycle, then that state is periodic. If every state has the same period, then everything moves 'in sync', affecting its convergence properties.

Definition 7 (Period of state) *Let $\{X_k\}$ be an S-valued Markov process. The **period** of a state $a \in S$ is the largest d satisfying: $(\forall k, n \in \mathbb{N})$*

$$P(X_{n+k} = a | X_k = a) > 0 \Rightarrow d \text{ divides } n$$

*If $d = 1$, then the state a is **aperiodic**.*

Definition 8 (Closed subset) *Let $\{X_k\}$ be an S-valued Markov process. The subset $\widetilde{S} \subseteq S$ is called **closed subset** if $\forall a \in \widetilde{S}, \forall b \notin \widetilde{S} \Rightarrow \omega(a, b) = 0$.*

Definition 9 (Irreducible closed subset) *Let $\{X_k\}$ be an S-valued Markov process. The subset $\widetilde{S} \subseteq S$ is called **irreducible closed subset** iff $\widetilde{S}$ is a closed subset, and no proper subset of $\widetilde{S}$ is closed subset.*

Definition 10 (Irreducible Markov process) *Let $\{X_k\}$ be an S-valued Markov process. The Markov process is called **irreducible Markov process** iff S is a irreducible closed subset.*

Definition 11 (Ergodic Markov process) *An **ergodic** Markov process is a Markov process $\{X_k\}$ that is both*

- ***irreducible**: every state is reachable from every other state.*
- ***aperiodic**: the greatest common divisor of the states periods is 1.*

Lemma 1 (Ergodic Condition) *An irreducible S-valued Markov process with transition function ω that has $\omega(a, a) > 0$ for some state $a \in S$ is aperiodic, and hence ergodic. Proof can be found in [6]*

Theorem 1 (Ergodic Convergence) *If $\{X_k\}$ is an ergodic S-valued Markov process, then the probability function converges for all $a \in S$:*

$$\lim_{k \to \infty} P(X_k = a) = p_a$$

Proof can be found in [2].

3.3 Transition matrix and stationary distribution

Lets the set of states is finite and n is the number of states. Let us number all states, e.g. $\forall a_i \in S, i = \overline{1, n}$. Now, the transition function ω can be rewritten in form of transition matrix P, e.g.

$$P_{ij} = \omega(a_i, a_j), \forall a_i, a_j \in S$$

Definition 12 (Row-stochastic matrix) *An matrix P is called **row stochastic matrix** iff $Pe = e$, where e is a vector of appropriate dimension whose all entries equal one.*

The transition matrix is a row-stochastic matrix.

Definition 13 (Ergodic matrix) *An matrix P is called **ergodic** matrix iff it is a transition matrix of the Markov process $\{X_k\}$ and the Markov process is ergodic.*

Definition 14 (Probability distribution) *An row vector π having dimension n is called **probability distribution** over a set of states S if $\pi e = 1$.*

Definition 15 (Stationary probability distribution) *An row vector π is called **stationary probability distribution** over a set of states S if π is probability distribution and*

$$\pi P = \pi. \tag{1}$$

As it follows from Theorem 1, ergodic transition matrix has unique stationary distribution.

3.4 Power method

One way to compute the stationary distribution of a Markov process is by explicitly computing the distribution using $\pi^{(k+1)} = \pi^{(k)}P$, until the distribution converges. This leads us to the Power Method for computing the stationary distribution of P.

$function\ \pi^{(m)} = PowerMethod(P, v, \varepsilon)\{$
$\pi^{(0)} = v;$
$k = 1;$
repeat
    $\pi^{(k)} = \pi^{(k-1)}P;$
    $\delta = \|\pi^{(k)} - \pi^{(k-1)}\|_1;$
    $k = k + 1;$
until $\delta < \varepsilon;$
$\},$

where $\|\pi\|_1 = \pi e$, v is the first approximation, and ε is accuracy. The rate of convergence of the power method is given by $\frac{|\lambda_2|}{|\lambda_1|}$, where $\lambda_1,\ \lambda_2$ are eigenvalues of P, if all eigenvalues are ordered by modulus, [7,15]. If P is row-stochastic matrix then $\lambda_1 = 1,\ 1 \geqslant |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n| \geqslant 0$ [8].

4 Defining of PageRank

Lets consider how authors of the method, Brin and Page [14], define PageRank. Lets A is a web page and $T_1, ..., T_m$ are pages linking to the page A. Lets $l(A)$ is the number of outgoing links from the page A. The parameter c is a damping factor which can be set between 0 and 1 (typically 0.85). Assume that n is the number of pages. Then PageRank of the page A, $\pi(A)$, define as

$$\pi(A) = \frac{(1-c)}{n} + c(\pi(T_1)/l(T_1) + \ldots + \pi(T_m)/l(T_m)) \qquad (2)$$

If we number all page then we call the row vector $\pi = (\pi_1, \ldots, \pi_n)$, where π_i is the PageRank of i-th page, the PageRank vector. The vector is a distribution of probability on the set of the pages. The PageRank of a page, as it consequents from (2), depends from PageRank of pages refering to the page, hence, computing of the PageRank is iterative process.

The PageRank vector is a stationary probability distribution of special formed Markov process. Lets consider a surfer, who stochastic rambling over Internet.

Assume that there are n web pages in Internet and the surfer is on i-th page having k_i outgoing links. Lets the surfer chooses either one of the outgoing link of the i-th page with probability c or arbitrary page with probability $(1-c)$. Hence, we obtain Markov chain where pages are states and transition function depends on outgoing links of pages. The Markov chain has finite number of states, therefore, we can define transition matrix. Lets the matrix P define with the rule: $P_{ij} = 0$, if page i has not links to the page j, and $P_{ij} = 1/l(i)$, if the page i has a link to the page j, and $P_{ii} = 0$ in any case. If the page i has no outgoing links then $P_{ij} = 1/n, \forall j = \overline{1,n}$. The P matrix is row-stochastic. The transition matrix for the Markov chain is

$$G = cP + (1-c)1/nE, \tag{3}$$

where E is a matrix whose all entries equal one. The matrix G is called a Google matrix. A Google matrix is row-stochastic and ergodic. The PageRank vector satisfies the equation

$$\pi = \pi G \tag{4a}$$
$$\pi e = 1, \tag{4b}$$

where e is a vector of appropriate dimension whose all entries equal one. Equation (4b) is used for normalization of PageRank vector. One can find the PageRank vector from (4) [4,11,13] as

$$\pi = \frac{1-c}{n} e^t (I - cP)^{-1}, \tag{5}$$

Because of there are the large number of web pages in the Internet (Google reports about 8 billions of pages), calculation of $(I - cP)^{-1}$ is computation expensive. Different methods of approximative computation of PageRank were developed. Power iteration method helps to find the PageRank vector. $PowerMethod(G, \frac{1}{n}e^t, \varepsilon)$. It is known that for a Google matrix $|\lambda_1(G)| = 1$ and $|\lambda_2(G)| \leqslant c$ [8,11], so power method converges to stationary distribution with rate c.

5 Decomposition

The PageRank vector is very long and it is good idea try to divide it on several components and find each component separately and, after that, find the whole PageRank vector. The matrix P is represented in block structure

for the purpose.

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1N} \\ P_{21} & P_{22} & \dots & P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{N1} & P_{N2} & \dots & P_{NN} \end{pmatrix}, \tag{6}$$

where $N < n$. The PageRank vector is

$$\pi = (\pi_1, \pi_2, \dots, \pi_N), \tag{7}$$

where π_I is row vector with $dim(\pi_I) = n_I$ and

$$\sum_{I=1}^{N} n_I = n$$

5.1 Block-diagonal case

Lets consider the case when the matrix P is block-diagonal [1], e.g.

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 \\ 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_N \end{pmatrix}. \tag{8}$$

For block I define the perturbed matrix

$$G_I = cP_I + (1-c)1/n_I E, \tag{9}$$

and let vector π_I be the PageRank of P_I such that

$$\pi_I = \pi_I G_I \tag{10a}$$
$$\pi_I e = 1, \tag{10b}$$

Theorem 2 *The PageRank π is given by*

$$\pi = \left(\frac{n_1}{n}\pi_1, \frac{n_2}{n}\pi_2, \dots, \frac{n_N}{n}\pi_N\right) \tag{11}$$

Proof can be found in [1].

So, we can compute part of the PageRank vector independently and find whole PageRank with formula (11). Each block in the matrix P represents

disconnected component in the Web graph, but, as it was shown in Figure 1, there are the large number of pages in giant connected component, hense, although the method lets decrease dimension of unknown vector but in small degree only.

5.2 2×2 *case*

Let us consider the case

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \ \pi = (\pi_1, \pi_2)$$

where P_{11} and P_{22} are square. The equation (1) can be rewritten as

$$\pi(I - P) = 0.$$

The partition of P was considered in [9]. Assume that P is irreducible. Hence, $I-P$ is an singular and irreducible, the non-trivial leading principal submatrix $I - P_{11}$ is non-singular [3]. Hence, we can factor $I - P = LDU$, where

$$L = \begin{pmatrix} I & 0 \\ -P_{21}(I - P_{11})^{-1} & I \end{pmatrix}, \tag{12}$$

$$D = \begin{pmatrix} I - P_{11} & 0 \\ 0 & I - S \end{pmatrix}, \tag{13}$$

$$U = \begin{pmatrix} I & -(I - P_{11})^{-1}P_{12} \\ 0 & I \end{pmatrix}, \tag{14}$$

and

$$S = P_{22} + P_{21}(I - P_{11})^{-1}P_{12}$$

The matrix S is the stochastic complement of P_{22} in P [12]. Since U is non-singular we have $\pi(I - P) = 0$ iff $\pi LD = 0$. Hence,

$$\pi_2 S = \pi_2 \qquad \pi_1 = \pi_2 P_{21}(I - P_{11})^{-1}, \tag{15}$$

which means that π_2 is a stationary distribution for the smaller matrix S. Since P is irreducible and stochastic, so is S. Hence, S has a unique stationary distribution σ,

$$\sigma S = \sigma, \qquad \sigma e = 1$$

Therefore, we can determine π_2 from the stationary distribution σ of S and then $\pi_2 = \rho\sigma$ where the factor ρ is responsible for the normalization $\pi e = 1$.

Unfortunately, there are cases when $|\lambda_2(S)| > |\lambda_2(P)|$, so power method for S converges slower then one for P. But if the partitioning is considered for a Google matrix G there is such stochastic complement S that $|\lambda_2(S)| < |\lambda_2(G)|$ [9].

6 Aggregation/Disaggregation method

When power method is used for finding PageRank different components of the PageRank vector can converges with different speed. And while the appropriate accuracy is achieved for some components we have to continue computation to reach a good accuracy for components converging slowly. It is useful if we do as many iteration for each component of the PageRank vector as need to achieve appropriate accuracy. Aggregation/disaggregation methods are based on the idea. Partitioning of a Google matrix is used

$$G = \begin{pmatrix} G_{11} & G_{12} & \dots & G_{1N} \\ G_{21} & G_{22} & \dots & G_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ G_{N1} & G_{N2} & \dots & G_{NN} \end{pmatrix}. \tag{16}$$

6.1 Blockrank method

Lets π_i, $i = \overline{1, N}$, is a PageRank vector for G_{ii} if G_{ii} represents entire Web, and if links to pages in other groups did not exist. The vector π_i is called local PageRank for i-th group, and found approximately with $\pi_i = PowerMethod(G_{ii}, \frac{1}{n}e^t, \varepsilon)$. Power method can converge with different speed for different groups and we do as many iterations for each group as we need to get appropriate accuracy. An aggregation matrix A is defined as

$$A_{ij} = \pi_i G_{ij} e, \tag{17}$$

where e is a vector of appropriate dimension whose all entries equal one. Lets ν is a PageRank vector for the matrix A. Dimension of ν is equal to N. Now we can approximate global PageRank vector as local PageRank vectors weighted by ν, e.g.

$$\widetilde{\pi} = (\nu_1\pi_1, \dots, \nu_N\pi_N), \tag{18}$$

and use $\widetilde{\pi}$ for power method with matrix G, $\pi = PowerMethod(G, \widetilde{\pi}, \varepsilon)$. The approach was considered in [10].

7 Conclusion

The World Wide Web is a very complex entity, and individuals find sophisticated approaches to discover its properties.

Acknowledgements

I would like to thank Vladimir Dobrynin for useful discussions and editing the paper.

References

[1] K.Avrachenkov and N.Litvak. Decomposition of the Google PageRank and Optimal Linking Strategy. Inria Sophia Antipolis, University of Twente, 2004.

[2] E.Behrends. Introduction to Markov Chains (with Special Emphasis on Rapid Mixing). Vieweg Verlag, 1999.

[3] A.Berman and R.J.Plemmons. Nonnegative Matrices in the Mathematical Sciences. SIAM Classics In Applied Mathematics, SIAM, Philadelphia, 1994.

[4] M.Bianchini, M.Gori, and F.Scarselli. Inside PageRank. ACM Trans, Internet Technology, In press, 2002.

[5] A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, J.Wiener. Graph structure in the web. Proc. WWW9 conference, 309-320, May 2000. http://www9.org/w9cdrom/160/160.html

[6] A.Clausen. Online Reputation Systems: The Cost of Attack of PageRank. 2003

[7] G.H.Golub and C.F.V.Loan. Matrix Computations. The Johns Hopkins University Press, Baltimore, 1996.

[8] T.H.Haveliwala and S.D.Kanvar. The second eigenvalue of the Google matrix. Tech. Rep. 2003-20, Stanford University, March 2003. http://dbpubs.stanford.edu/pub/2003-20

[9] C.F.Ipsen and S.Kirklad. Convergence analysis of the Langville-Meyer PageRank algorithm.

[10] S.Kamver, T.Haveliwala, C.Manning, and G.Golub. Exploiting the block structure of the web for computing PageRank. Tech. Rep. SCCM03-02, Stanford University, http://www-sccm.stanford.edu/nf-publications-tech.html, 2003.

[11] A.N.Langville and C.D.Meyer. Deeper Inside PageRank. Preprint, North Carolina State University, 2003.

[12] C.Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. SIAM Rev., 31 (1989), pp. 240-72.

[13] C.D.Moler and K.A.Moler. Numerical Computing with MATLAB. SIAM, 2003.

[14] L.Page, S.Brin, R.Motwani, and T.Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[15] J.H.Wilkinson. The Algebraic Eigenvalue Problem. Oxford University Press, Oxford, 1965.