The Condition Number of the PageRank Problem

Sepandar D. Kamvar and Taher H. Haveliwala

Stanford University
{sdkamvar,taherh}@cs.stanford.edu

Abstract. We determine analytically the condition number of the PageRank problem. Specifically, we prove the following statement:
"Let P be an $n \times n$ row-stochastic matrix whose diagonal elements $P_{ii} = 0$. Let c be a real number such that $0 \leq c < 1$. Let E be the $n \times n$ rank-one row-stochastic matrix $E = \boldsymbol{e}\boldsymbol{v}^T$, where $\boldsymbol{e}$ is the n-vector whose elements are all $e_i = 1$, and $\boldsymbol{v}$ is an n-vector that represents a probability distribution.
Define the matrix $A = [cP + (1-c)E]^T$. The problem $A\boldsymbol{x} = \boldsymbol{x}$ has condition number $\kappa = (1+c)/(1-c)$."
This statement has implications for the accuracy to which PageRank can be computed, currently and as the web scales. Furthermore, it provides a simple proof that, for values of c that are used by Google, small changes in the link structure of the web do not cause large changes in the PageRanks of pages of the web.

1 Theorem

Theorem 1. *Let P be an $n \times n$ row-stochastic matrix whose diagonal elements $P_{ii} = 0$.. Let c be a real number such that $0 \leq c \leq 1$. Let E be the $n \times n$ rank-one row-stochastic matrix $E = \boldsymbol{e}\boldsymbol{v}^T$, where $\mathbf{e}$ is the n-vector whose elements are all $e_i = 1$, and $\boldsymbol{v}$ is an n-vector that represents a probability distribution[1].*
Define the matrix $A = [cP + (1-c)E]^T$. The problem $A\boldsymbol{x} = \boldsymbol{x}$ has condition number $\kappa = (1+c)/(1-c)$.

2 Notation and Preliminaries

P is an $n \times n$ row-stochastic matrix whose diagonal elements $P_{ii} = 0$. E is the $n \times n$ rank-one row-stochastic matrix $E = \boldsymbol{e}\boldsymbol{v}^T$, where $\boldsymbol{e}$ is the n-vector whose elements are all $e_i = 1$ and $\boldsymbol{v}$ is an n-vector whose elements are all non-negative and sum to 1. A is the $n \times n$ column-stochastic matrix:

$$A = [cP + (1-c)E]^T \tag{1}$$

We let $\boldsymbol{x}$ be the dominant eigenvector of A. By convention, we choose eigenvectors $\boldsymbol{x}$ such that $||\boldsymbol{x}||_1 = 1$. Since A is a non-negative matrix, the dominant eigenvector $\boldsymbol{x}$ is also non-negative. Therefore,

$$\boldsymbol{e}^T\boldsymbol{x} = ||\boldsymbol{x}||_1 = 1 \tag{2}$$

Since A is column-stochastic, it's dominant eigenvalue $\lambda_1 = 1$, $1 \geq |\lambda_2| \geq \ldots \geq |\lambda_n| \geq 0$. That is,

$$A\boldsymbol{x} = \boldsymbol{x} \tag{3}$$

[1] i.e., a vector whose elements are nonnegative and whose L_1 norm is 1.

3 Proof of Theorem 1

We prove this case via a series of lemmas.

Lemma 1. $E^T\boldsymbol{x} = \boldsymbol{v}$.

Proof. By definition, $E = \boldsymbol{e}\boldsymbol{v}^T$. Therefore, $E^T\boldsymbol{x} = \boldsymbol{v}\boldsymbol{e}^T\boldsymbol{x}$. From equation 2, $\boldsymbol{e}^T\boldsymbol{x} = 1$. Therefore, $E^T\boldsymbol{x} = \boldsymbol{v}$, and Lemma 1 is proved.

Lemma 2. The eigenvalue problem $A\boldsymbol{x} = \boldsymbol{x}$ can be rewritten as the nonsingular system of equations $(I - cP^T)\boldsymbol{x} = (1 - c)\boldsymbol{v}$.

Proof. From $A\boldsymbol{x} = \boldsymbol{x}$, we can rearrange terms to get

$$(I - A)\boldsymbol{x} = 0.$$

By the definition of A (equation 1):

$$[I - (cP + (1 - c)E)^T]\boldsymbol{x} = 0.$$

From Lemma 1, $E^T\boldsymbol{x} = \boldsymbol{v}$. Therefore, $(I - cP^T)\boldsymbol{x} - (1 - c)\boldsymbol{v} = 0$. Rearranging terms, we get $(I - cP^T)\boldsymbol{x} = (1 - c)\boldsymbol{v}$, and Lemma 2 is proved.

Lemma 3. $\boldsymbol{x} = (I - cP^T)^{-1}\boldsymbol{v}$.

Proof. Let $M = I - cP^T$. Then $M^T = I - cP$. Since P has zeros on the diagonals and is row-stochastic, and since $c < 1$, $I - cP$ is strictly diagonally dominant and therefore invertible. Since M^T is invertible, M is also invertible. Therefore, we may write $\boldsymbol{x} = (I - cP^T)^{-1}\boldsymbol{v}$ and Lemma 3 is proved.

Lemma 4. $||I - cP^T||_1 = 1 + c$.

Proof. Since the diagonal elements of cP^T are all zero,

$$||I - cP^T||_1 = ||I||_1 + c||P^T||_1 = 1 + c||P^T||_1.$$

Since P^T is a column-stochastic matrix, $||P^T||_1 = 1$. Thus, $||I - cP^T||_1 = 1 + c$ and Lemma 4 is proved.

Lemma 5. $||(I - cP^T)^{-1}||_1 = 1/(1 - c)$.

Proof. Recall from equation 1 that $A = [cP + (1 - c)E]^T$, where $E = \boldsymbol{e}\boldsymbol{v}^T$ and v is some n-vector whose elements are non-negative and sum to 1. Let $\boldsymbol{x}(\boldsymbol{e_i})$ be the n-vector that satisfies the following equations:

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{e_i} \\ A\boldsymbol{x}(\boldsymbol{e_i}) &= \boldsymbol{x}(\boldsymbol{e_i}) \\ ||\boldsymbol{x}(\boldsymbol{e_i})||_1 &= 1. \end{aligned}$$

From Lemma 2, $\boldsymbol{x} = (1-c)(I-cP^T)^{-1}\boldsymbol{v}$. Therefore, $\boldsymbol{x}(\boldsymbol{e_i}) = (1-c)(I-cP^T)^{-1}\boldsymbol{e_i}$. Taking the norm of both sides, $||\boldsymbol{x}(\boldsymbol{e_i})||_1 = (1-c)||(I-cP^T)^{-1}\boldsymbol{e_i}||_1$. Since $||\boldsymbol{x}(\boldsymbol{e_i})||_1 = 1$, we have

$$||(I - cP^T)^{-1}\boldsymbol{e_i}||_1 = 1/(1-c). \tag{4}$$

Notice that $(I-cP^T)^{-1}\boldsymbol{e_i}$ gives the ith column of $(I-cP^T)^{-1}$. Thus, from equation 4, the L1 norm of the matrix $(I - cP^T)^{-1}$ is $||(I - cP^T)^{-1}|| = 1/(1-c)$.

Lemma 6. The 1-norm condition number of $\boldsymbol{x} = (I-cP^T)^{-1}\boldsymbol{v}$ is $\kappa = (1+c)/(1-c)$.

Proof. By definition, the 1-norm condition number κ of the problem $\boldsymbol{y} = M^{-1}\boldsymbol{b}$ is given by $\kappa = ||M||_1||M^{-1}||_1$. From Lemmas 4 and 5, this is $\kappa = (1+c)/(1-c)$.

4 Implications

The matrix A is used by Google to compute PageRank, an estimate of web-page importance used for ranking search results [3]. PageRank is defined as the stationary distribution of the Markov chain corresponding to the $n \times n$ stochastic transition matrix A^T. The matrix P corresponds to the web link graph; in making P stochastic, there are standard techniques for dealing with web pages with no outgoing links [1].

The strongest implication of this result has to do with the stability of PageRank. A proof of stability of PageRank is given in [2], but we show a tighter stability bound here. Imagine that the Google matrix A is perturbed slightly, either by modifying the link structure of the web (by adding or taking away links), or by changing the value of c. Let us call this perturbed matrix $\tilde{A} = A+\epsilon B$, where ϵB is the "error matrix" describing the change to the web matrix A. Let $\boldsymbol{x}$ be the PageRank vector corresponding to the web matrix A, and let $\tilde{\boldsymbol{x}}$ be the vector corresponding to the web matrix $\tilde{A}$. It is known that, for a linear system of equations,

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_1 \leq \kappa\epsilon||B||$$

From Theorem 1, we can rewrite this as:

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}||_1 \leq \epsilon\frac{1+c}{1-c}||B||$$

What this means is, for values of c near to 1, PageRank is not stable, and a small change in the link structure may cause a large change in PageRank. However, for smaller values of c such as those likely used by Google ($.8 < c < .9$), PageRank is stable, and a small change in the link structure will cause only a small change in PageRank.

Another implication of this is the accuracy to which PageRank may be computed. Again, for values of c likely used by Google, PageRank is a *well-conditioned* problem meaning that it may be computed accurately by a stable algorithm. However, for values of c close to 1, PageRank is an *ill-conditioned* problem, and it cannot be computed to great accuracy by any algorithm.

Acknowledgments

We would like to thank Gene Golub for useful conversations.

This paper is based on work supported in part by the National Science Foundation under Grant No. IIS-0085896 and Grant No. CCR-9971010, and in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University (research project on Concept Bases for Lexical Acquisition and Intelligently Reasoning with Meaning).

References

1. S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.
2. A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In *IJCAI*, pages 903–910, 2001.
3. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1998.