# The Second Eigenvalue of Google Matrix

**Anna Maria K V**
**Anurag Dey**
**Anuraj Kashyap**
**Anusha R**

**Submitted to:**
Dr. Priyavrat Deshpande,
Associate Professor,
Chennai Mathematical Institute

cmi
CHENNAI
MATHEMATICAL
INSTITUTE

## Abstract

We determine analytically the modulus of the second eigenvalue for the web hyperlink matrix used by Google for computing PageRank. Specifically, we prove the following statement:
"For any matrix $A = [cP + (1-c)E]^T$, where $P$ is an $n \times n$ row-stochastic matrix, E is a nonnegative $n \times n$ rank-one row-stochastic matrix, and $0 \leq c \leq 1$, the second eigenvalue of $A$ has modulus $|\lambda_2| \leq c$. Furthermore, if $P$ has at least two irreducible closed subsets, the second eigenvalue $\lambda_2 = c$."
This statement has implications for the convergence rate of the standard PageRank algorithm as the web scales, for the stability of PageRank to perturbations to the link structure of the web, for the detection of Google spammers, and for the design of algorithms to speed up PageRank. We perform a brief study about applications of the second eigenvalue on the convergence of PageRank and spam detection.

# 1 Some Important Definitions

**Markov chain -** A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
**Stochastic matrix -** A stochastic matrix is a matrix used to characterize transitions for a finite Markov chain, Elements of the matrix must be real numbers in the closed interval $[0, 1]$.
**Row-stochastic matrix -** A row-stochastic matrix is a stochastic matrix where the sum of all the elements in a row is equal to 1.

# 2 Theorem

**Theorem 1.** *Let $P$ be an $n \times n$ stochastic matrix. Let $c$ be a real number such that $0 \leq c \leq 1$. Let $E$ be the rank-one row-stochastic matrix $E = ev^T$, where $e$ is the n-vector whose elements are all $e_i = 1$, and $v$ is an n-vector that represents a probability distribution.*
*Define the matrix $A = [cP + (1-c)E]^T$. Its eigenvalue $|\lambda_2| \leq c$.*

**Theorem 2.** *Further if $P$ has at least two irreducible closed subsets $(which\ is\ the\ case\ for\ web\ hyperlink\ matrix)$, then the second eigenvalue is given by, $\lambda_2 = c$.*

## 3 Notations and Preliminaries

$P$ is an $n \times n$ row-stochastic matrix; $E$ is an $n \times n$ rank-one row-stochastic matrix, such that $E = ev^T$, where $e$ is the *n-vector* whose elements are $e_i = 1$. $A$ is the column-stochastic matrix:

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

We denote the $i^{th}$ eigenvalue of A as $\lambda_i$, and the corresponding eigenvector as $x_i$.

$$Ax_i = \lambda_i x_i \tag{2}$$

We choose eigenvectors $x_i$ such that $||x_i||_1 = 1$(by convention).
Since, $A$ is column-stochastic,

$$\lambda_1 = 1, 1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$$

Similarly, we can denote $i^{th}$ eigenvalue of $P^T$ as $\gamma_i$, and its corresponding eigenvector as $y_i$.
Since, $P^T$ is column-stochastic,

$$\gamma_1 = 1, 1 \geq |\gamma_2| \geq \cdots \geq |\gamma_n| \geq 0$$

Similarly, we can denote $i^{th}$ eigenvalue of $E^T$ as $\mu_i$, and its corresponding eigenvector as $z_i$.
Since, $E^T$ is rank-one and column-stochastic,

$$\mu_1 = 1, \mu_2 = \cdots = \mu_n = 0$$

It can be noted that for any row-stochastic matrix, $M$, $Me = e$.
$S$ is a *closed subset* corresponding to $M$ if and only if $i \in S$ and $j \notin S \implies M_{ij} = 0$.
$S$ is a *irreducible closed subset* corresponding to $M$ if and only if $S$ is a closed subset and no proper subset of $S$ is a closed subset.

## 4 Proof of Theorem 1

We first show that Theorem 1 is true when $c = 0$ and $c = 1$.

- *Case* 1: $c = 0$
  If $c = 0$, then from equation 1, we get $A = E^T$. Since $E$ is rank-one matrix, therefore $\lambda_2 = 0$. Thus, Theorem 1 is proved for $c = 0$.

- *Case* 2: $c = 1$
  If $c = 1$, then from equation 1, we get $A = P^T$. Since $P^T$ is column-stochastic matrix, therefore $\lambda_2 \leq 1$. Thus, Theorem 1 is proved for $c = 1$.

- *Case* 3: $0 < c < 1$
  We will prove this by a series of lemmas.

  **Lemma 1.** The second eigenvalue of $A$ has modulus, $|\lambda_2| < 1$.

  *Proof:* Consider the Markov chain corresponding to $A^T$. Now, if the Markov chain corresponding to $A^T$ has only one irreducible closed subchain $S$, and if $S$ is aperiodic, then the chain corresponding to $A^T$ must have a unique eigenvector with eigenvalue 1, by *Ergodic Theorem* in **??** Lemma 1.1 shows that $A^T$ has a single irreducible closed subchain S, and Lemma 1.2 shows this subchain is aperiodic.

  *Lemma 1.1:* There exists a unique irreducible subset S of the Markov chain corresponding to $A^T$.

  *Proof:* We split the proof into proof of existence and proof of uniqueness.
  *Existence*:
  Let the set $U$ be the states with nonzero components in $v$. Let $S$ consist of the set of all states reachable from $U$ along nonzero transitions in the chain. Observe that $S$ trivially forms a closed subset. Since every state has a transition to $U$, no subset of $S$ can be closed. Therefore, $S$ forms an irreducible closed subset.

  *Uniqueness*:
  Every closed subset must contain U, and every closed subset containing U must contain S. Therefore, S must be the unique irreducible closed subset of the chain.

  *Lemma 1.2:* The unique irreducible closed subset S is an aperiodic subchain.
  *Proof*: From Theorem 3 in Appendix, all members in an irreducible closed subset have the same period. Therefore, if at least one state in S has a self-transition, then the subset S is aperiodic. Let u be any state in U. By construction, there exists a self-transition from u to itself. Therefore, S must be aperiodic.

  **Lemma 2.** The second eigenvector $x_2$ of A is orthogonal to e: $e^T x_2 = 0$.
  *Proof:* Since $|\lambda_2| < |\lambda_1|$ (by Lemma 1), the second eigenvector $x_2$ of A is orthogonal to the first eigenvector of $A^T$ by Theorem 2 in the Appendix. The first eigenvector of AT is e. Therefore, $x_2$ is orthogonal to e.

  **Lemma 3.** $E^T x_2 = 0$.

*Proof:* By definition,

$$\begin{aligned}
&E = ev^T \\
\implies &E^T = ve^T \\
\implies &E^T x_2 = ve^T x_2 \\
\implies &E^T x_2 = v(e^T x_2) = v0 \\
\implies &E^T x_2 = 0
\end{aligned}$$

Hence, Proved.

**Lemma 4.** The second eigenvector $x_2$ of $A$ must be an eigenvector $y_i$ of $P^T$, and the corresponding eigenvalue is $\gamma_i = \lambda_2/c$.
*Proof:* From equation 1 and 2:

$$cP^T x_2 + (1-c)E^T x_2 = \lambda_2 x_2 \tag{3}$$

From Lemma 3 and equation 3:

$$cP^T x_2 = \lambda_2 x_2 \tag{4}$$

Putting $y_i = x_2$ and $\gamma_i = \lambda_2/c$:

$$P^T y_i = \gamma_i y_i \tag{5}$$

Therefore, the second eigenvector $x_2$ of $A$ is an eigenvector $y_i$ of $P^T$, and the corresponding eigenvalue is

$$\gamma_i = \lambda_2/c \tag{6}$$

**Lemma 5.** $|\lambda_2| \leq c$
*Proof:* From Lemma 4, we see that $\lambda_2 = \gamma_i c$. But, since $P$ is stochastic, therefore $|\gamma_i| \leq 1$. Hence $|\lambda_2| = |\gamma_i|c \leq c$. Hence, Theorem 1 is proved.

# 5 Proof of Theorem 2

Theorem 2 states that if $P$ has at least two irreducible closed subsets, $\lambda_2 = c$.
*Proof:*

- *Case* 1: $c = 0$.
  This case is proven in Theorem 1.

- *Case* 1: $c = 1$.
  From Ergodic Theorem, the multiplicity of eigenvalue 1 equals the number of irreducible closed subsets of the chain. Since P has at least two irreducible closed subsets, $\lambda_2 = 1$

- *Case* 3: $0 < c < 1$. We will prove this using two lemmas.

  **Lemma 1.** Any eigenvector $y_i$ of $P^T$ that is orthogonal to $e$ is an eigenvector $x_i$ of $A$.
  The relationship between eigenvalues is $\lambda_i = c\gamma_i$

  *Proof:*

  It is given that

  $$e^T y_i = 0 \tag{7}$$

  Therefore,

  $$E^T y_i = ve^T y_i = 0 \tag{8}$$

  By definition,

  $$P^T y_i = \gamma_i y_i \tag{9}$$

  From equations 1, 8, and 9,

  $$Ay_i = cP^T y_i + (1 - c)E^T y_i = c\gamma_i y_i \tag{10}$$

  Therefore, $y_i$ is an eigenvector of $A$ with eigenvalue $c\gamma_i$. Hence Lemma 1 is proved.
  **Lemma 2.** There exists an eigenvector $x_i$ of A such that the corresponding $\lambda_i = c$.

  *Proof:*

  From Ergodic theorem, the multiplicity of eigenvalue 1 is at least two for $P^T$.
  Therefore, we can find two linearly independent eigenvectors $y_1$ and $y_2$ of $P^T$ corresponding to the dominant eigenvalue 1. Let,

  $$k_1 = y_1^T e \tag{11}$$

  $$k_2 = y_2^T e \tag{12}$$

  If $k_1 = 0$ or $k_2 = 0$,choose $x_i$ to be $y_1$ $y_2$ respectively.

  If $k_1 > 0$ and $k_2 > 0$, Choose,

  $$x_i = \frac{y_1}{k_1} - \frac{y_2}{k_2} \tag{13}$$

  $x_i$ is an eigenvector of $P^T$ with eigenvalue 1

  $$\begin{aligned} P^T x_i &= P^T(\frac{y_1}{k_1} - \frac{y_2}{k_2}) \\ &= \frac{y_1}{k_1} - \frac{y_2}{k_2} \\ &= x_i \end{aligned} \tag{14}$$

and, $x_i$ is orthogonal to $e$

$$\begin{aligned} e^T x_i &= e^T(\frac{y_1}{k_1} - \frac{y_2}{k_2}) \\ &= \frac{e^T y_1}{k_1} - \frac{e^T y_2}{k_2} \\ &= 0 \end{aligned} \tag{15}$$

From Lemma 1 $x_i$ is an eigenvector of $A$ corresponding to eigenvalue $\lambda_i = c$.

$$|\lambda_2| \geq |\lambda_i| = c \tag{16}$$

From Theorem 1,

$$|\lambda_2| \leq c \tag{17}$$

Hence

$$\lambda_2 = c$$

# 6 Implications

**Convergance of Google Page Rank:**
The World Wide Web is a vast network of interconnected web pages, where links between pages are used to navigate and explore information. The importance of a web page can be determined by its popularity, which is typically measured by the number of incoming links it has. In this report, we describe the PageRank algorithm, a technique for analyzing the importance of web pages based on their link structure. The algorithm views the web as a directed graph, with nodes representing web pages and edges representing links between them.
*Probabilistic View:*
The PageRank algorithm models the process of a random surfer on the web who follows hyperlinks from one page to another. The importance of a page i can be viewed as the probability that a random surfer on the Internet that opens a browser to any page and starts following hyperlinks, visits the page i. The process can be modeled as a random walk on the web graph. A smaller, but positive percentage of the time, the surfer will dump the current page and choose arbitrarily a different page from the web and "teleport" there. The damping factor p reflects the probability.
*Page Rank Vector:*
The PageRank vector for a web graph with transition matrix A and damping factor p is the unique probabilistic eigenvector of the matrix M, corresponding to the dominant eigenvalue 1. The PageRank algorithm computes the PageRank vector by iteratively applying the power method to the transition matrix A until convergence. The power method starts with an initial probability distribution over the nodes, typically uniform or proportional to the number of

incoming links, and updates the probabilities based on the link structure of the web.
*Convergence:*
By Perron-Frobenius theorem, if the underlying web-graph is connected and aperiodic, then the power-iteration algorithm used to compute the page ranks is guaranteed to converge to a steady-state vector, which is precisely the vector with the page ranks of all the nodes. Therefore, the PageRank algorithm can be used to analyze the importance of web pages and provide insights into the structure of the web.
The second eigen value of the google page matrix helps us determine the rate at which the page converges to the steady state Rank vector. The smaller the second eigen value faster will be the rate of convergence.
*Conclusion:*
The PageRank algorithm is a powerful tool for analyzing the importance of web pages based on their link structure. The algorithm models the process of a random surfer on the web and computes a unique probabilistic eigenvector of the transition matrix corresponding to the dominant eigenvalue 1. The power-iteration algorithm used to compute the page ranks is guaranteed to converge to a steady-state vector, which is precisely the vector with the page ranks of all the nodes. The algorithm has numerous applications in web search, information retrieval, and network analysis and has revolutionized the way we navigate and explore the vast web of information.

**Spam Detection:**
Link spam is a type of spamming technique used to manipulate search engine rankings by creating a large number of low-quality links pointing to a website. Link spam intends to increase the number of links pointing to a website, which can lead to higher search engine rankings. A typical technique to increase the PageRank of a group of websites is to create many inlinks to the group, and to remove all outlinks. This makes it easy for random surfer to enter the group, but difficult to leave.
Each pair of leaf nodes in P corresponds to an eigenvector of A having eigenvalue c. These leaf nodes may have incoming edges but no outgoing edges. Link spammers often generate such structures in attempts to hoard ranks. Analysis of the non-principal eigenvectors of A may lead to strategies for combating link spam.
The second eigenvalue of the PageRank matrix can be an indicator of the presence of link spamming because it represents the stability of the PageRank distribution. If the second eigenvalue is close to zero, it indicates that the distribution of PageRank scores is close to the expected distribution, and there is no significant deviation from it.
However, if the second eigenvalue is negative, it indicates that there is a significant deviation from the expected distribution, which is often caused by link spamming.

# 7 Appendix

**Theorem 1: (The Ergodic Theorem)**
If $P$ is the transition matrix for the finite Markov chain, then the multiplicity of the eigenvalue 1 equals the number of irreducible closed subsets of the chain.

**Theorem 2:**
If $x_i$ is an eigenvector of $A$ corresponding to the eigenvalue $\lambda_i$, and $y_j$ is an eigenvector of $A^T$ corresponding to $\lambda_j$, then $x_i^T y_j = 0$ (if $\lambda_i \neq \lambda_j$).

**Theorem 3:**
Two distinct states belonging to the same class (irreducible closed subset) have the same period. In other words, the property of having period $d$ is a class property.

# 8 References

1. M. Iosifescu. *Finite Markov Processes and Their Applications.* John Wiley and Sons, Inc., 1980.

2. 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.

3. A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In IJCAI, pages 903–910, 2001

4. A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In *IJCAI*,pages 903–910, 2001.

5. `https://www.sciencedirect.com/science/article/pii/S0377042714004105`

6. `http://networksciencebook.com/chapter/5introduction5`

| **Team Member** | **Contribution** |
|---|---|
| Anna Maria K V | Worked on the theorems of the research paper<br>Prepared the slides and the report |
| Anurag Dey | Worked on the implementations of the research paper<br>Prepared the slides and the report |
| Anuraj Kashyap | Worked on the theorems of the research paper<br>Prepared the slides and the report |
| Anusha R | Worked on the implementations of the research paper<br>Prepared the slides and the report |