Second Eigenvalue of Google Matrix

Anna Maria K V (MDS202209)
Anurag Dey (MDS202210)
Anuraj Kashyap (MDS202211)
Anusha R (MDS202212)

April 22, 2023

Motivation

- Theorems on second eigenvalue
- Convergence of PageRank
- Spam Detection

Theorems

Theorem 1

Let P be an $n \times n$ row-stochastic matrix. Let $c \in \mathbb{R}$ such that $0 \leq c \leq 1$. Let E be the $n \times n$ rank-one row-stochastic matrix, such that $E = ev^T$, 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 second eigenvalue $|\lambda_2| \leq c$.

Theorem 2

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

Important Notations and Preliminaries

Definitions

- **Google Matrix:** A Google matrix is a particular stochastic matrix that is used by Google's PageRank algorithm. The matrix represents a graph with edges representing links between pages.
- **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.
- **Column-stochastic Matrix:** A column-stochastic matrix is a stochastic matrix where the sum of all the elements in a column is equal to 1.

Definitions(*Cont.*)

- **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.
- **Closed subset:** A set of states S is a closed subset of the Markov chain corresponding to matrix M if and only if $i \in S$ and $j \notin S$ implies that $M_{ij} = 0$.
- **Irreducible closed subset:** A set of states S is an irreducible closed subset of the Markov chain corresponding to the matrix M if and only if S is a closed subset, and no proper subset of S is a closed subset.

cmi

Important Notations and Preliminaries

- P is an $n \times n$ row-stochastic matrix.
- E is the $n \times n$ rank-one row-stochastic matrix, $E = ev^T$.
- A is the $n \times n$ column-stochastic matrix:

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

- i^{th} eigenvalue of A is denoted as λ_i, and its corresponding eigenvector is x_i, i.e.

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

- Since A is column-stochastic, $\lambda_1 = 1$ and $1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$.
- Also, we choose eigenvectors x_i such that $||x_i||_1 = 1$.

Important Notations and Preliminaries

- i^{th} eigenvalue of P^T is denoted by γ_i, and its corresponding eigenvector is y_i.
- Since P^T is column-stochastic, $\gamma_1 = 1$ and $1 \geq |\gamma_2| \geq \cdots \geq |\gamma_n| \geq 0$.
- Similarly, i^{th} eigenvalue of E^T is denoted by μ_i, and its corresponding eigenvector is z_i.
- Since E^T is rank-one and column stochastic, $\mu_1 = 1$ and $\mu_2 = \cdots = \mu_n = 0$.
- For any row-stochastic matrix M, $Me = e$.
- E, P, and A^T are row stochastic, and can thus be viewed as transition matrices of Markov chains.

Important theorems used in proofs

- **Ergodic Theorem:** If P is the transition matrix for a finite Markov chain, then the multiplicity of the eigenvalue 1 is equal to the number of irreducible closed subsets of the chain.[1]
- If x_i is an eigenvector of A corresponding to the eigenvalue λ_i, and y_j is an eigenvector of A^T corresponding to λ_j, then $x_i^T y_j = 0$ (if $\lambda_i \neq \lambda_j$).[1]
- 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. [1]

Proof of Theorem 1

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

- **Case 1:** $c = 0$
  If $c = 0$, then from equation 1, $A = E^T$. Since E is a rank-one matrix, $\lambda_2 = 0$. Thus, Theorem 1 is proved for $c = 0$.
- **Case 2:** $c = 1$
  If $c = 1$, then from equation 2, $A = P^T$. Since P^T is a column-stochastic matrix, $|\lambda_2| \leq 1$. Thus, Theorem 1 is proved for $c = 1$.
- **Case 3:** $0 < c < 1$
  We prove this case via 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.
- 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 the Ergodic Theorem.
- Now we simply show that the Markov chain corresponding to A^T has a single irreducible closed subchain S, and that this subchain is aperiodic.
- Lemma 1.1 will show that A^T has a single irreducible closed subchain S, and Lemma 1.2 will show that this subchain is aperiodic.

cmi

Lemma 1.1

There exists a unique irreducible closed subset S of the Markov chain corresponding to A^T.

Proof

We will 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.
- But S is an irreducible closed subset.
- Therefore, S must be the unique irreducible closed subset of the chain.

Therefore, there exists a unique irreducible closed subset S of the Markov chain corresponding to A^T.

Lemma 1.2

The unique irreducible closed subset S is an aperiodic subchain.

Proof

- From the third theorem in ITP, we know that 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.
- By construction for any $u \in U$, there exists a self transition from u to itself.
- Therefore the unique irreducible closed subset S is an aperiodic subchain.

From Lemmas 1.1 and 1.2, and the Ergodic Theorem, $|\lambda_2| < 1$and Lemma 1 is proved.

Lemma 2

The second eigenvector x_2 of A is orthogonal to e. That is, $e^T x_2 = 0$.

Proof

- Since $|\lambda_2| < |\lambda_1| = 1$, the second eigenvector x_2 of A is orthogonal to the first eigenvector of A^T by the second theorem in ITP.
- The first eigenvector of A^T 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.

Proof of Theorem 2

Theorem:
P has at least two irreducible closed subsets, $\lambda_2 = c$

- **Case 1:** $c = 0$
  Proved in Theorem 1
- **Case 2:** $c = 1$
  Ergodic theorem states that, *If P is the transition matrix of a finite Markov chain, then the multiplicity of the eigenvalue* 1 *is equal to the number of irreducible closed subsets of the chain.*
  Hence, $\lambda_2 = 1$
- **Case 3:** $0 < c < 1$
  We proof this case 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

Given, $$e^T y_i = 0$$

Therefore,

$$E^T y_i = ve^T y_i = 0 \tag{7}$$

By definition,

$$P^T y_i = \gamma_i y_i \tag{8}$$

From equations 1, 7, and 8,

$$Ay_i = cP^T y_i + (1-c)E^T y_i = c\gamma_i y_i \tag{9}$$

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,

$$k_1 = y_1^T e \tag{10}$$

$$k_2 = y_2^T e \tag{11}$$

- $k_1 = 0$ or $k_2 = 0$:

$$x_i = y_1 \text{ or } x_i = y_2$$

- $k_1, k_2 > 0$:

  Choose,

$$x_i = \frac{y_1}{k_1} - \frac{y_2}{k_2} \tag{12}$$

  x_i is an eigenvector of P^T with eigenvalue 1

$$\begin{aligned} P^T x_i &= P^T(\frac{y_1}{k_1} - \frac{y_2}{k_2}) \\ &= \frac{y_1}{k_1} - \frac{y_2}{k_2} \\ &= x_i \end{aligned} \tag{13}$$

x_i is orthogonal to e

$$\begin{aligned} e^T x_i &= e^T\left(\frac{y_1}{k_1} - \frac{y_2}{k_2}\right) \\ &= \frac{e^T y_1}{k_1} - \frac{e^T y_2}{k_2} \\ &= 0 \end{aligned} \tag{14}$$

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

$$|\lambda_2| \geq |\lambda_i| = c \tag{15}$$

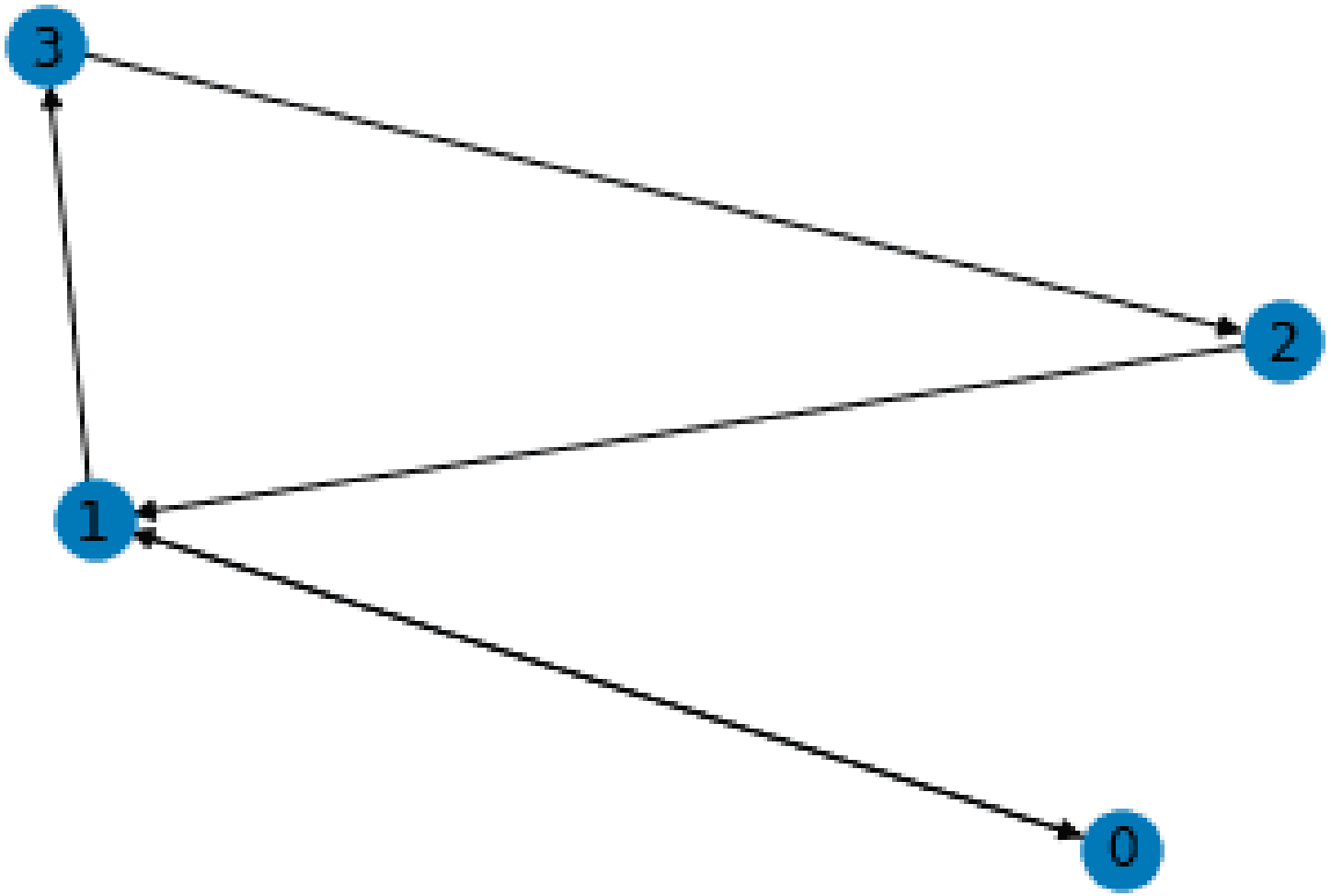
From Theorem 1,

$$|\lambda_2| \leq c \tag{16}$$

Hence, $\lambda_2 = c$

Web as a directed graph

Nodes are pages and edges are hyperlinks.Lets say there are 4 pages and this is the following webgraph.



Adjacency Matrix

Each column of the adjacency matrix represents the out-going edges of each node, and in this example the matrix is:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Degree Matrix

D=diag($\sum_j A_{ij} | \forall i \in v$)

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

Transition Matrix

Normalize the adjacency Matrix A by the degree of each node M. It represents the transitional probability at each node.
M is column stochastic.

$$M = AD^{-1} = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 0 \end{pmatrix}.$$

Random Surfer

Consider a random surfer who explores the web just by clicking on the hyperlinks on the current page uniformly at random. For example, if the random surfer is viewing the page 2, the next page will be the page 0 or 3 at a 50-50 chance.
How often does this random surfer reach each page? Let v be the probability distribution over the 4 pages and initialized as the uniform distribution. We can get to the answer by multiplying the column stochastic transition matrix M from the left iteratively.

$V_k + 1 = M * V_k$

Eigen Decomposition or Iterative Multiplication

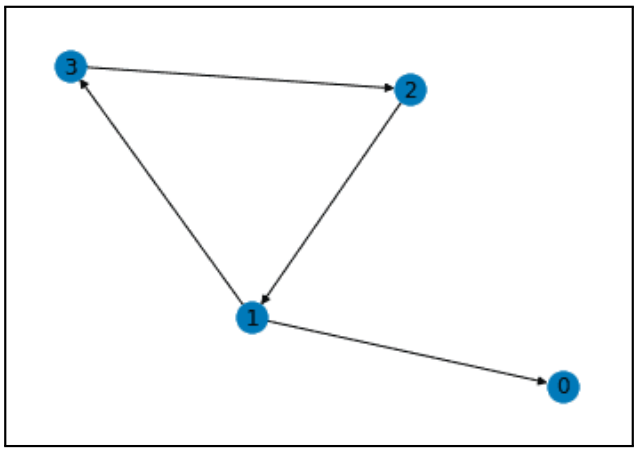
If we do this for 11 times finally, we get to the answer! The random surfer is viewing the page 1 for 40% of the time and page 0, 2, and 3 for 20% of the time. This final probability is called PageRank and serves as an importance measure for web pages.
The iterative multiplication has converged to a constant PageRank. It is equivalent to calculating the eigenvector corresponding to the eigen value 1 by the power method.

$$v=M*v$$

The need for Google Page Matrix?

Sink dangling Nodes

- Nodes with no outgoing edges
- Sink Nodes absorb Random Surfer and set the rank of other pages to 0



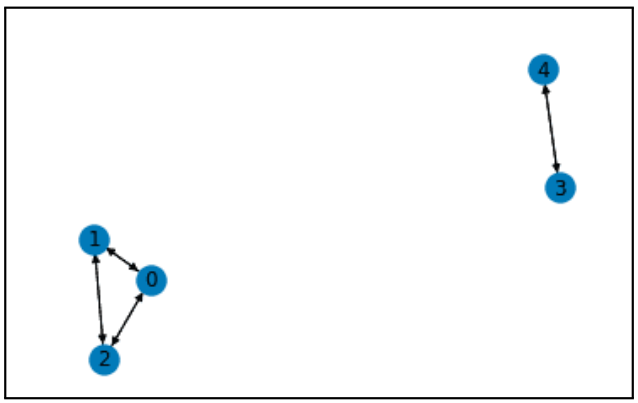
$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \\ 1/3 & 0 & 0 & 1 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

Figure: Transition Matrix

The need for Google Page Matrix?

Disconnected Components

- The Web graph may have disconnected components



$$M = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$v = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/2 \\ 1/2 \end{pmatrix}$$

Figure: Tranistion Matrix and eigen vectors

The need for Google Page Matrix?

To address this issue the damping factor d (=0.15) is introduced and the transition matrix is reformulated as :

$$\tilde{M} = (1 - d) * M + \frac{d}{n} * J_n \ ,$$

where n is the number of nodes and J_n is a matrix of ones.This reformulated transition matrix is also referred to as the Google matrix.

Perron Frobenius Theorem

- M has an eigenvalue of multiplicity 1
- 1 is the largest eigenvalue: all the other eigenvalues have absolute values smaller than 1.
- For the eigenvalue 1, there exists a unique eigenvector with the sum of its entries equal to 1.

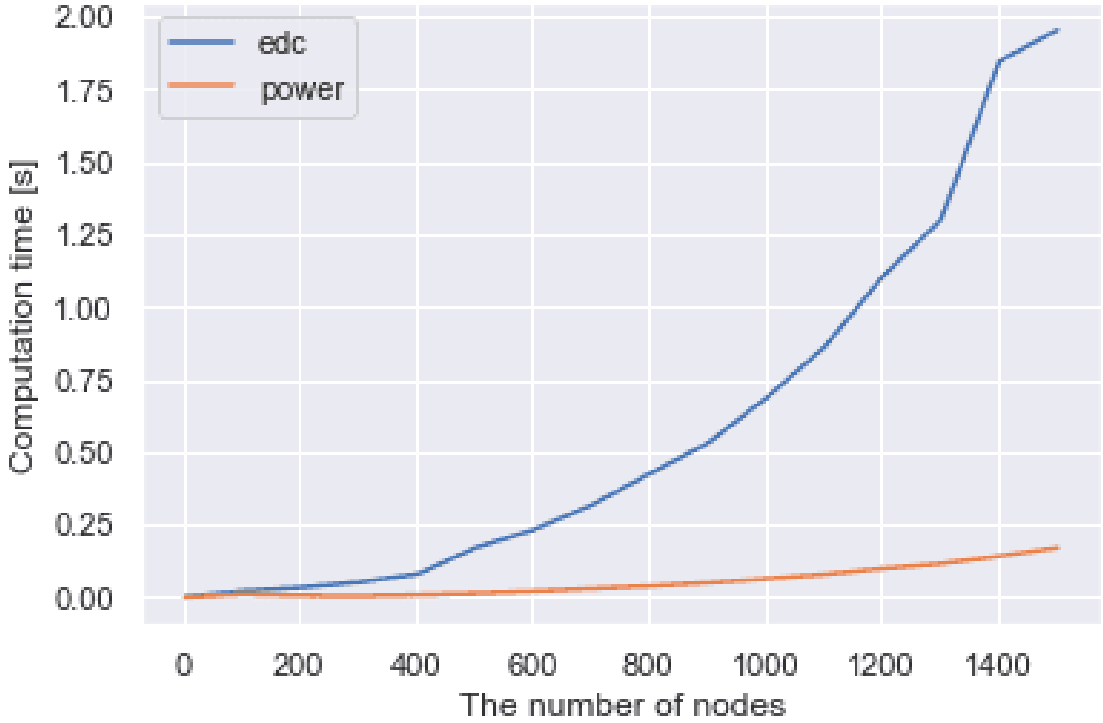
The L1 normalized eigenvector corresponding to the largest eigenvalue is the page-rank vector.

Power Method

It initializes v_0 as the uniform distribution and iteratively multplies the google matrix until it converges or reaches the max no of iterations.
Denoting the greatest and the second greatest (in absolute value) eigenvalue as λ_1,λ_2 respectively, the convergence ratio is:
$\left|\frac{\lambda_2}{\lambda_1}\right| = |\lambda_2|$
The smaller $|\lambda_2|$ is, the faster the algorithm converges.

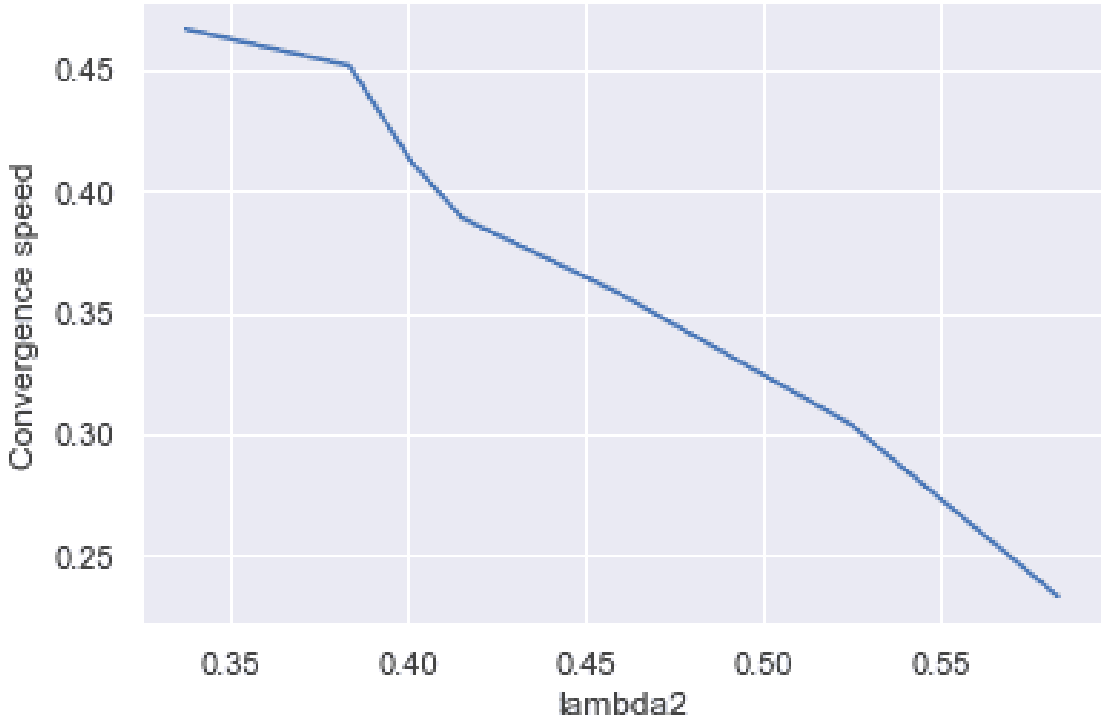
Eigen Decomposition Vs Power Method

Computation time The figure below compares the eigendecomposition and the power method on the computation time of PageRank for Barabási–Albert network with the different number of nodes.



The power method is faster than the eigen decomposition especially when there are many nodes. So, a smaller value of $|\lambda_2|$ ensures a faster computation time of page rank even for a large number of nodes.

Convergence Speed



It is experimentally confirmed that the smaller $|\lambda_2|$ is, the faster the power method converges.

Link Spamming

- 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.
- The intention of link spam is 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.

Example

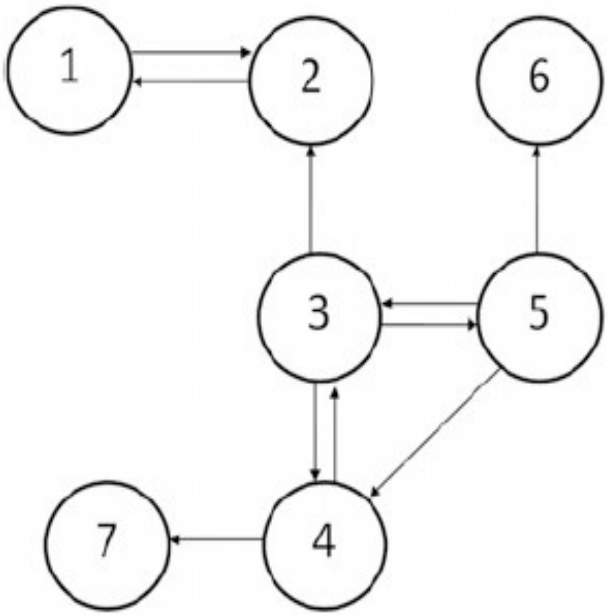


Fig. 1. Simple directed graph.

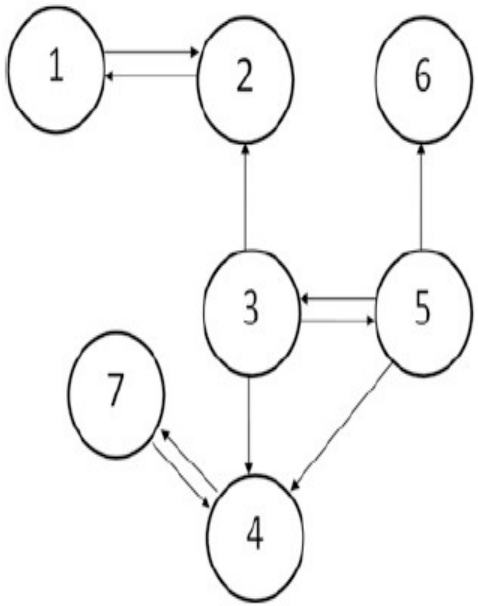


Fig. 2. Changes to improve the PageRank for node 4.

Spam Detection

- Each pair of leaf node 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.

Role of Second Eigenvalue

- 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.

Thank You!