LAA Project Report

Nandini Jaiswal
Narendra C
Nooh Ali
Om Ambaye

cmi

CHENNAI
MATHEMATICAL
INSTITUTE

Submitted to: Prof. Priyavrat Deshpande

Contents

1 Introduction 2

2 Some Preliminaries and Notations 2

3 Results 2

4 Applications 5

4.1 Link Spamming 5

4.1.1 Introduction 5

4.1.2 Energy in PageRank 5

4.1.3 Methods to boost PageRank 5

4.1.4 Irreducible closed subsets and link spamming 6

4.1.5 Example 7

4.2 The second eigenvector and its relation to link spamming 7

4.3 Algorithm to compute second eigenvector 8

4.3.1 Block Power Method 8

4.3.2 Simple Power Method 9

5 Work Distribution 9

6 Appendix 10

7 References 10

1 Introduction

This a report on the paper **The Second Eigenvalue of the Google Matrix** by Taher H. Haveliwala and Sepandar D. Kamvar. 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 non-negative $n \times n$ 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 algorithms to detect the second eigenvalue and link spamming detection.

2 Some Preliminaries and Notations

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 $e_{\mathrm{i}} = 1$.
A is the column-stochastic matrix:

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

We denote the i^{th} eigenvalue of A ss λ_i, and the corresponding eigenvector as x_{i}.

$$Ax_i = \lambda_1 x_i \tag{2}$$

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

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

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

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

Similarly, we can denote i^{th} eigenvalue of E^T as μ_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

$$i \in S \text{ and } j \notin S \Longrightarrow M_{ij} = 0$$

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.

3 Results

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 v is an n-vector that represents a probability distribution. Define the matrix $A = [cP + (1 - c)E]^T$. Its eigenvalue $|\lambda_2| \leq c$.

Proof of Theorem 1:
We proceed case wise.

- 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 subset S, and if S is aperiodic, then the chain corresponding to A^T must have a unique eigenvector with eigenvalue 1, by Ergodic Theorem. Lemma 1.1 shows that A^T has a single irreducible closed subset S, and Lemma 1.2 shows this subset 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 markov chain.
  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 A^T is e. Therefore, x_2 is orthogonal to e.

- **Lemma 3**: $E^T x_2 = 0$
  Proof: By definition,
  $$\begin{aligned} & E = ev^T \\ \implies & E^T = ve^T \\ \implies & E^T x_2 = ve^T x_2 \\ \implies & E^T x_2 = v\left(e^T x_2\right) = v0 \\ \implies & E^T x_2 = 0 \end{aligned}$$
  Hence, Proved.

- **Lemma 4**: The second eigenvector x_2 of A must be an eigenvector y_1 of P^T, and the corresponding eigenvalue is $\gamma_i = \lambda_2/c$.
  Proof: From equations 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 $\gamma_{\mathrm{t}} = \lambda_2/c$
  $$P^T y_i = \gamma_i y_i \tag{5}$$
  Hence, the second eigenvector x_2 of A is an eigenvector y_i of P^T, and the corr. eigenvalue is
  $$\gamma_i = \lambda_2/c \tag{6}$$

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

Hence, Theorem 1 is proved.

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 $\lambda_2 = c$

Proof of Theorem 2:
Again we proceed case wise.

- Case 1: $c = 0$. This case is already proven in Theorem 1.
- Case 2: $c = 1$. 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 6**: 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,
    $$E^T y_{\mathrm{i}} = ve^T y_1 = 0 \tag{8}$$
    By definition,
    $$P^T y_1 = \gamma_1 y_1 \tag{9}$$
    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 $c\gamma_t$. Hence Lemma 1 is proved.

    - **Lemma 7**: 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 and y_2 respectively.
    If $k_1 > 0$ and $k_2 > 0$, Choose,
    $$x_i = \frac{y_1}{k_1} - \frac{y_2}{k_2}$$
    x_i is an eigenvector of P^T with eigenvalue 1
    $$\begin{aligned} P^T x_1 &= P^T \left(\frac{y_1}{k_1} - \frac{y_2}{k_2} \right) \\ &= \frac{y_1}{k_1} - \frac{y_2}{k_2} \\ &= x_1 \end{aligned}$$
    and x_i is orthogonal to e
    $$\begin{aligned} e^T x_i &= e^T \left(\frac{y_1}{k_1} - \frac{y_2}{k_2} \right) \\ &= \frac{e^T y_1}{k_1} - \frac{e^T y_2}{k_2} \\ &= 0 \end{aligned}$$
    From Lemma 6, x_i is an eigenvector of A corresponding to eigenvalue $\lambda_i = c$.
    $$|\lambda_2| \geq |\lambda_i| = c$$

From Theorem 1,

$$|\lambda_2| \leq c$$

Hence

$$\lambda_2 = c$$

Hence, we have proved Theorem 2.

4 Applications

4.1 Link Spamming

4.1.1 Introduction

Link spamming involves manipulating the PageRank algorithm, originally designed by Google to rank web pages based on their importance. Companies, often referred to as link farms, specialize in boosting the PageRank of client websites through strategic hyperlinking from other sites. This practice is deemed undesirable by search engines because it can degrade the quality of search results. As Google continues to evolve its algorithm to counteract such tactics, the exact mechanisms of PageRank today remain undisclosed, making it a continuous battle against link spamming.

4.1.2 Energy in PageRank

The concept of 'Energy' as introduced by Bianchini offers a new perspective on PageRank, suggesting that websites with more energy have higher PageRank values. The energy of a website group (W_I) is calculated as:

$$E_I = |I| + E_I^{\text{in}} - E_I^{\text{out}} - E_I^{dn}$$

where $|I|$: amount of pages in W_I,
E_I^{in} : energy from outside W_I going inside W_I,
E_I^{out} : energy from inside W_I going outside W_I,
E_I^{dn} : energy going to dangling nodes inside W_I.

4.1.3 Methods to boost PageRank

These are two methods to increase the PageRank of a particular website. Suppose for example we have the initial Markov Chain as:

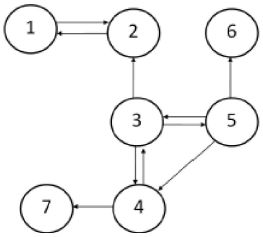

Figure 1: The initial Markov Chain

- **Method 1**
This method involves the strategic addition of promotion nodes, which are specific types of nodes within a network that link back and forth exclusively to the target page. This configuration is aimed at maximizing the internal energy within the target group, thereby enhancing the PageRank of the target node. For example, replacing a dangling node (a node with no outgoing links that distributes energy outward) with a promotion node helps retain more energy within the target group.

- **Method 2**
The best way to increase your website's page rank is to create an irreducible closed subset. To do this, first remove all dangling nodes and external hyperlinks and then add sufficient promotion nodes. This setup ensures that all energy remains within the target group, thus maximizing the PageRank. The key advantage of this method is the minimal external energy leakage, making it highly efficient for boosting PageRank.

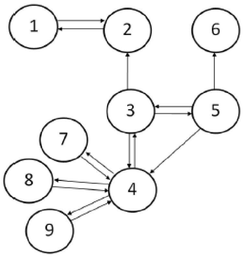

Figure 2: The Markov Chain after applying method 1 for node 4

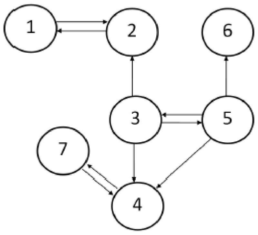

Figure 3: The Markov Chain after applying method 2 for node 4

4.1.4 Irreducible closed subsets and link spamming

Clearly,in the Method 2, node 4 now has the highest PageRank. To analyse this we will recall some well known defintions.

Definition 1. *A set of states S is a closed subset of the Markov chain corresponding to $\boldsymbol{P}^T$ if and only if $i \in S$ and $j \notin S$ implies that $p_{ji} = 0$.*

Definition 1 tells us that a Markov chain is closed if it is not possible to get out of subset S as soon as you are in it. This means that any subset containing a dangling node cannot be closed, and in particular, any dangling node cannot be a a closed subset.

Definition 2. *A set of states S is an irreducible closed subset of the Markov chain corresponding to $\boldsymbol{P}^T$ if and only if S is a closed subset, and no proper subset of S is a closed subset.*

Let l be the number of irreducible closed subsets of $\mathbf{P}$. Then we can rewrite $\mathbf{P}$ in canonical form[5] by renumbering the nodes:

$$\mathbf{P} \sim \begin{pmatrix} \mathbf{T_{11}} & \mathbf{T_{12}} \\ \mathbf{0} & \mathbf{T_{22}} \end{pmatrix} = \left(\begin{array}{cccc|cccc} \mathbf{P_{11}} & \mathbf{P_{12}} & \cdots & \mathbf{P_{1r}} & \mathbf{P_{1,r+1}} & \mathbf{P_{1,r+2}} & \cdots & \mathbf{P_{1m}} \\ \mathbf{0} & \mathbf{P_{22}} & \cdots & \mathbf{P_{2r}} & \mathbf{P_{2,r+1}} & \mathbf{P_{2,r+2}} & \cdots & \mathbf{P_{2m}} \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P_{rr}} & \mathbf{P_{r,r+1}} & \mathbf{P_{r,r+2}} & \cdots & \mathbf{P_{rm}} \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P_{r+1,r+1}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{P_{r+2,r+2}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P_{mm}} \end{array}\right),$$

Let $l = m - r$ and each $P_{11}, \ldots, P_{rr}$ is either irreducible or $[0]_{1\times1}$, and $P_{r+1;r+1}, \ldots, P_{mm}$ are irreducible and closed. First, note that each P_{ij} is a submatrix of the n-by-n matrix P. Let us call the dimension of the block T_{11} $\tilde{r}$-by-$\tilde{r}$, and thus the dimension of the block T_{22} is $(n - \tilde{r})$-by-$(n - \tilde{r})$.

4.1.5 Example

We illustrate the theory by the graph displayed in Figure 3. Firstly, we will renumber the nodes to get the canonical form as shown. For a graphical representation of the renumbering, we refer to Figure 4.

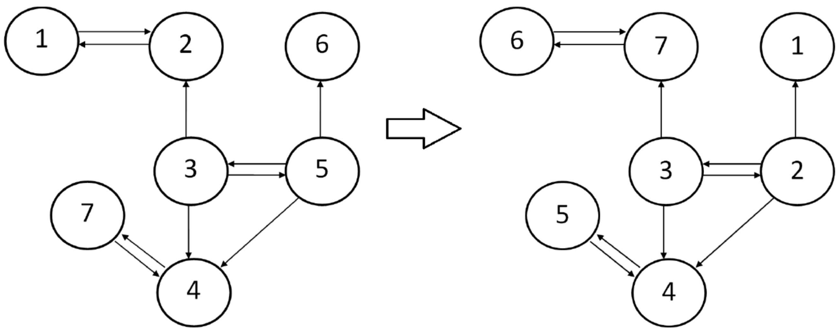

Figure 4: Renumbering the nodes of Figure 3 to canonical form.

Thus, rewriting **P** to **Pcanon**:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\sim \left(\begin{array}{ccc|cccc} \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \hline 0 & 0 & 0 & 0 & \multicolumn{1}{c|}{1} & 0 & 0 \\ 0 & 0 & 0 & 1 & \multicolumn{1}{c|}{0} & 0 & 0 \\ \cline{4-7} 0 & 0 & 0 & 0 & \multicolumn{1}{c|}{0} & 0 & 1 \\ 0 & 0 & 0 & 0 & \multicolumn{1}{c|}{0} & 1 & 0 \end{array}\right) = \mathbf{P_{canon}}.$$

Let us take a closer look at **Pcanon**. Firstly, we recognize the block on the lower left side of all zeros. Also, it is clear that we have two irreducible closed subsets (**P22** and **P33**), which can be reached by **T12**. However, **T12** includes all other nodes that are not in **T22** and thus, **P11** is the only block in the upper left side of **Pcanon** (i.e., there are no nodes that do not refer to one of the irreducible closed subsets). Note that **P11** is irreducible, but not closed. **P22** and **P33** are irreducible and closed.

4.2 The second eigenvector and its relation to link spamming

The second eigenvector in link analysis algorithms like PageRank helps detect link spamming by examining the link graph's structure, particularly within irreducible closed subsets. Irreducible closed subsets represent tightly interconnected clusters of web pages, often indicative of natural linking patterns. However, when link spamming occurs within these subsets, the flow of link authority is disrupted, leading to abnormal patterns in PageRank scores. Pages engaged in spamming often exhibit inflated PageRank scores or unnatural link distributions within these subsets, which can skew the second eigenvector. By analyzing deviations from expected patterns within irreducible closed subsets, search engines can identify and penalize sites attempting to manipulate rankings through spammy link practices, maintaining the integrity of their algorithms and providing users with more accurate search results.

4.3 Algorithm to compute second eigenvector

4.3.1 Block Power Method

Here is the algorithm (written in Python) for **blockpower** and **Ptwithoutdangling** method.

```python
import numpy as np
from numpy.linalg import eig
from scipy.linalg import orth
from scipy.sparse import csr_matrix

def blockpower(G, r):
    """
    Google's PageRank with block power method.

    Args:
        G (numpy.ndarray): Connectivity matrix.
        r (int): Size of the block, creates an n by r block.

    Returns:
        tuple:
            x (list): List of eigenvectors corresponding to the r largest eigenvalues.
            y (numpy.ndarray): Eigenvalues.
            iter (int): Number of iterations.
    """
    p = 0.85
    maxiter = 3000
    n, _ = G.shape
    delta = (1 - p) / n
    e = np.ones((n, 1))

    # Constructing Pt without dangling nodes
    Pt, k = Ptwithoutdangling(G)  # Unpack k as well

    # Choose Q(n,r) such that Q^T * Q = I
    Z = np.random.rand(n, r)
    iter = 0
    y2 = np.random.rand(r, 1)
    y = np.random.rand(r, 1)

    # QR decomposition
    while np.max(np.abs(np.sort(y2) - np.sort(y))) > 10 ** (-6) and iter < maxiter:
        Q = orth(Z)
        Z = p * Pt @ Q + delta * e @ (e.T @ Q) + p / n * e @ (k.T @ Q)
        if Q.shape[1] != r:
            print("Found at least one eigenvalue equal to zero. Try smaller r.")
            return
        s, D = eig(Q.T @ Z)
        y = np.sort(s)[-r:].reshape(-1, 1)
        iter += 1

    # Determining and scaling x
    x = [Q @ D[:, -i].reshape(-1, 1) for i in range(1, r + 1)]

    # Checking iteration amount
    if iter == maxiter:
        print(f"Maximum iterations ({maxiter}) reached: solution probably inaccurate!")

    return x, y, iter

def Ptwithoutdangling(G):
    """
    Returns the row-stochastic matrix P representing the possibility of
    transitioning between nodes in a graph without dangling nodes.

    Args:
        G: A numpy array representing the adjacency matrix of the graph.

    Returns:
        tuple:
        P: A scipy.sparse.csr_matrix representing the row-stochastic matrix.
        k: A numpy array representing the vector with values indicating whether a node is dangling.
    """
    n, _ = G.shape
    c = np.sum(G, axis=1)
    k = np.zeros((n, 1))
    Pt = csr_matrix((n, n), dtype=float)
    L = []

    for j in range(n):
        L.append(np.where(G[:, j])[0])
        c[j] = len(L[j])

    for j in range(n):
        if c[j] == 0:
            Pt[L[j], j] = 1 / n  # Assign 1/n instead of 0
            k[j] = 1
        else:
            Pt[L[j], j] = 1 / c[j]

    return Pt, k
```

4.3.2 Simple Power Method

The power method offers an alternative approach to identifying irreducible closed subsets within a matrix P^T, by directly analyzing its first eigenvector. While iterating $u_{k+1} = P^T u_k$ with an initial random stochastic vector, convergence is not guaranteed due to potential multiple irreducible closed subsets with eigenvalue $\gamma = 1$. Unlike scenarios where $\lambda_1 = 1$ is unique, here, the eigenvectors oscillate towards a linear combination corresponding to $\gamma = 1$. The convergence rate is determined by the second most dominant term, $c_{i+1}\gamma^{k_{i+1}}y_{i+1}$, with the rate equal to $|\gamma_{i+1}|$. However, the convergence rate remains unknown due to uncertainty regarding the largest eigenvalue distinct from one. Source code to find the same is given below, we are not importing any libraries, since we have imported for the above code and also we are using auxiliary function defined above.

```
def SimplePowerMethod(G,iter):
    G = np.array(G)
    n = G.shape[0]
    P_transi, k = Ptwithoutdangling(G)
    e = np.ones(n,dtype = int)
    x = np.random.rand(n)
    while (iter<=0):
        x = P_transi*x + 1/n*e*(k.T*x)
        iter -= 1
    return x/sum(x)
```

5 Work Distribution

Nandini Jaiswal	Block power method to compute second eigenvector
Narendra C	Link Spamming detection
Nooh Ali	Simple power method to compute second eigenvector
Om Ambaye	Theorems and proofs given in the paper

6 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.
(Ref. [1])

Theorem 2:
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 $\lambda_i \neq \lambda_j$)[2]
(Ref. [2])

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.[3]
(Ref. [3])

7 References

1. D. L. Isaacson and R. W. Madsen. *Markov Chains: Theory and Applications*, chapter IV, pages 126–127. John Wiley and Sons, Inc., New York, 1976.

2. J. H.Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, Oxford, 1965.

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

4. A. Sangers. *The second eigenvector of the Google matrix and its relation to link spamming.* Bachelor's Thesis, 2012. Link

5. Carl D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra.* Chapter 7 and 8. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.