Subspaces Preserved by a Matrix: The Internet and Markov Matrices

Karen Fletcher[2], Soham Lakhi[3], Bassam Nima[4], Jason Gilbert[1], Steven Rayan[1], Alex Weekes[1], Curtis Wendlandt[1],

[1]University of Saskatchewan [2]Athabasca University [3]University of British Colombia [4]University of Alberta

VXML

Background and Motivation

Prior to Google, search engines often found and ranked websites based on the incidence of keywords, which could bury important results below less relevant ones, and leave the user to look through pages of results.

Google's PageRank algorithm instead relies on a mathematical model for the structure of the internet

- **Hypothesis:** The number of links to a page reflects the relevance of a page
- **Goal:** Improve the quality of search results
- **Strategy:** Represent the structure of the internet mathematically create a mathematical measure of website significance, separate from user behaviour/website content.
- **Benefit:** Mathematical representation opens the problem to computational analysis, allowing the use of powerful computers.

Representation of the Internet

The structure of the internet can be represented by a directed graph and its associated matrix.

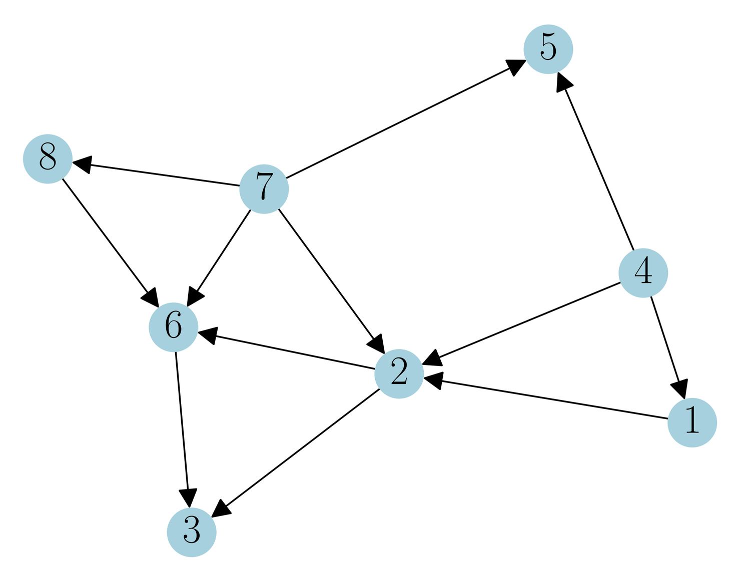

Figure 1. Webpages are nodes, edges are links between them

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

The adjacency matrix of the graph

Questions Posed

- What we can discover about the structure of the internet based on its matrix representation?
- How do the linear algebraic representations compare to corresponding representations in terms of Markov chain?

Random Surfer Model and the Stationary Distribution

$\mathbf{P}(\alpha)$: the new random surfer Markov matrix, α: probability the user accesses a hyperlink, $1-\alpha$: probability the user navigates to another site (searching/typing URL), $\mathbf{P}$: the original Markov matrix, n: the number of web pages, $\mathbf{ee}^T$: an $n \times n$ matrix of ones.

$$\mathbf{P}(\alpha) = \alpha\mathbf{P} + \frac{(1-\alpha)}{n}\mathbf{ee}^T \tag{1}$$

Stationary distribution $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi}^T\mathbf{P}(\alpha) = \boldsymbol{\pi}^T \tag{2}$$

which is a left-eigenvector using the dominant eigenvalue of the Markov matrix, $\lambda_1 = 1$.

Markov Matrix Sampling

The Markov matrix was sampled to build a Markov chain. Each node was counted and normalized to determine an estimate for the stationary distribution.

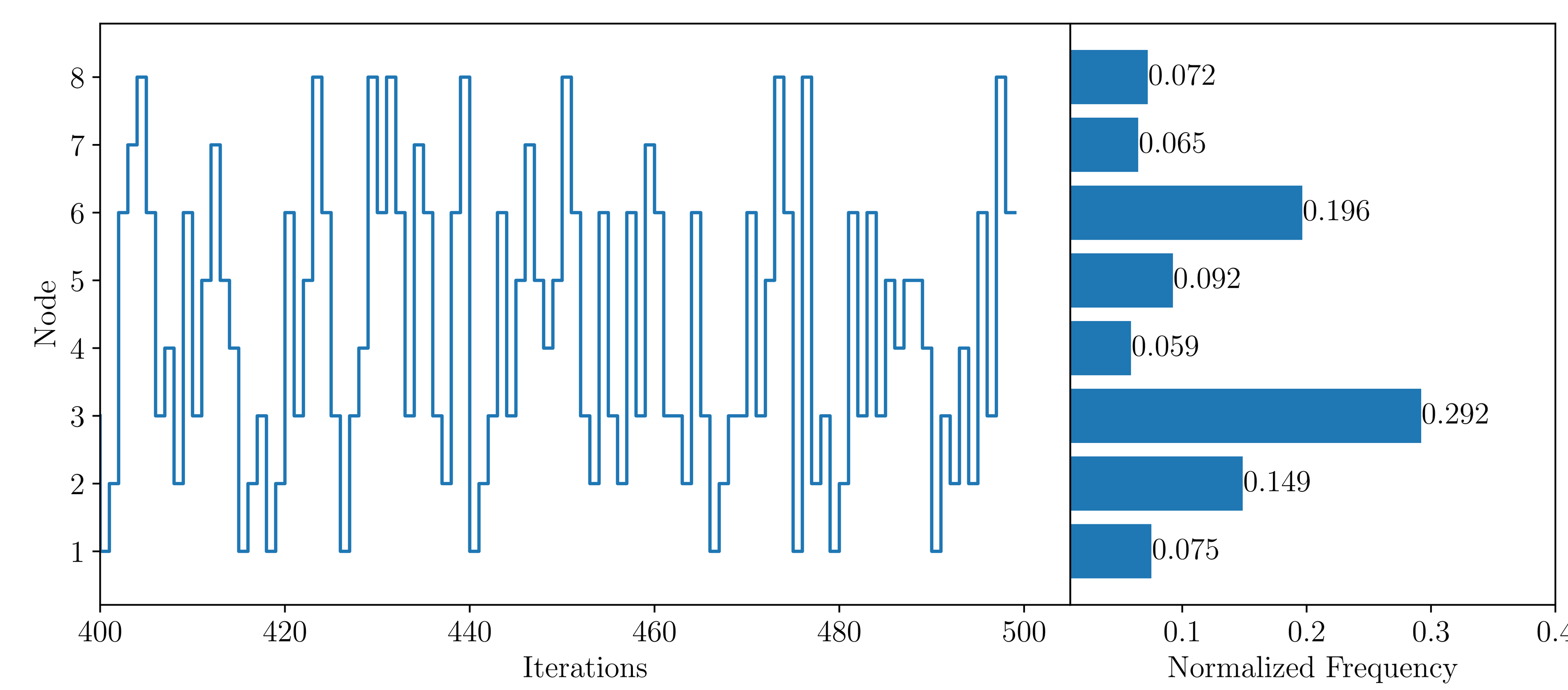

Figure 2. A sample 500 length Markov chain sampled from the Markov matrix, and counts of the nodes.

The normalized frequency of the simulation approaches the dominant eigenvector

$$\boldsymbol{\pi}^T = (0.076, 0.153, 0.293, 0.059, 0.089, 0.198, 0.059, 0.072)$$

Distributions of Markov Chains and Mean Squared Error

What happens when we generate many of these Markov chains with similar conditions?

- A distribution for each node forms (Fig. 3)
- Mean Squared Error: To measure the norm distance from the stationary distribution

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\pi_i - \hat{\pi}_i)^2 \tag{3}$$

where $\hat{\pi}_i$ is the ith node count frequency

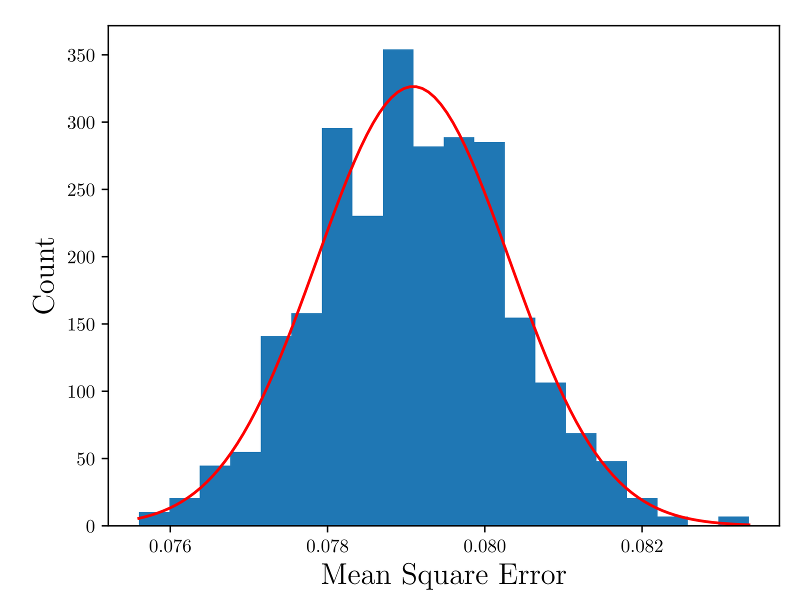

Figure 3. A sample 500 length Markov chain sampled from the Markov matrix, and counts of the nodes.

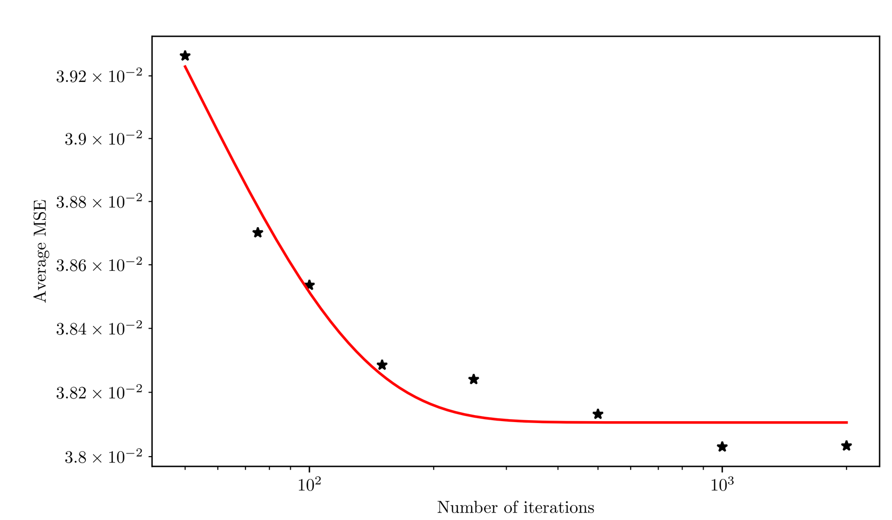

Figure 4. A sample 500 length Markov chain sampled from the Markov matrix, and counts of the nodes.

What happens to the mean and standard deviation of the MSE error?

- Standard deviation decreases with the simulation length linearly on a log-log plot
- Average MSE decreases non-linearly with $\mu = ad^t + c$ (Fig. 4) where t is the simulation time
- All fitting parameters were plotted against various second eigenvalues of networks

Convergence of the Power Method

A common way of computing the dominant eigenvector, in this case the stationary distribution, is using the Power Method, expressed as

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k\mathbf{P}\ , \tag{4}$$

where $\boldsymbol{\pi}$ is stationary vector which approximates the dominant eigenvector, $\boldsymbol{\pi}_k$ is the approximation resulting from the k_{th} iteration of the above procedure, and $\mathbf{P}$ is the transition matrix corresponding to the graph being considered.

In theory, this method should converge like $|\lambda_2|^k$. We tested and verified this prediction, as shown in Figure 5.

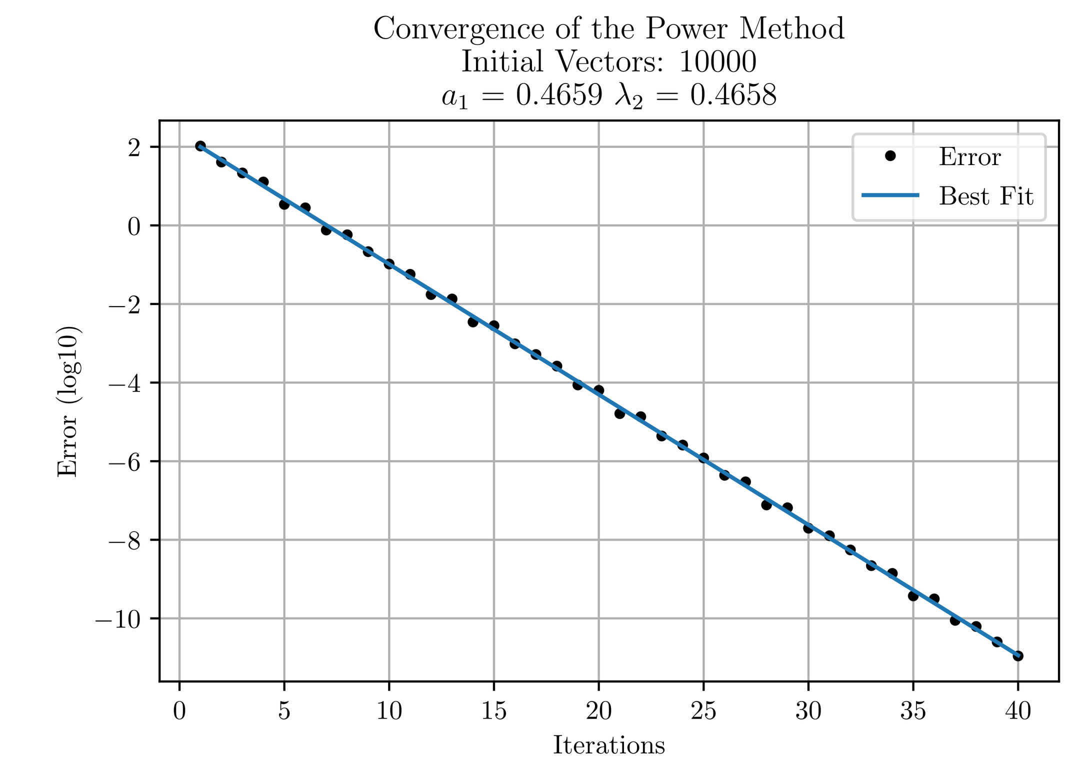

Figure 5. Error between approximate stationary distribution and dominant eigenvector, in terms of the Euclidean norm. Best fit according to the logarithm of a power law, where each value is averaged over 10000 initial vectors ($\boldsymbol{\pi}_0$).

Conclusion and Future Work

- In this work we compared a linear algebra model of the internet to a Markov chain model
- We found that they are consistent with each other, and that the linear algebra model gives considerable computational advantages

Future work may include:

- Studying the relationship between non-dominant eigenvectors and important features of the internet
- Computing the eigenvectors of many Markov matrices and searching for a correlation to any interesting features of the internet
- Using symbolic computation tools to search for symmetries (operations preserving the matrix or its eigenspace) in the Markov matrices representing the internet

References

[1] Ying Bao, Guang Feng, Tie-Yan Liu, Zhi-Ming Ma, and Ying Wang. Ranking websites: a probabilistic view. *Internet Mathematics*, 3(3): 295–320, 2006.

[2] Amy N Langville and Carl D Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[3] Amy N Langville and Carl D Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton university press, 2006.