Background and Motivation

Prior to Google, search engines often found and ranked websites based on the incidence of key-words, 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.

\[
A = \begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}
\]

The adjacency matrix of the graph

Questions Posed

• What can we 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

\[
P(\alpha): \text{the new random surfer Markov matrix, } \alpha \text{ probability the user accesses a hyperlink}, \ 0 < \alpha < 1 \text{ probability the user navigates to another site (searching/typing URL)}, \ P \text{ the original Markov matrix, } n \text{ the number of web pages, } e \text{ an } n \times n \text{ matrix of ones.}
\]

Stationary distribution \( \pi \) is given by

\[
\pi^T \cdot P(\alpha) = \pi^T \quad \text{(3)}
\]

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.

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

\[
\text{MSE} = \sum_{i=1}^{n} (\pi_i - \bar{\pi}_i)^2
\]

where \( \pi_i \) is the \( i \)th node count frequency

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