CS492 Special Topics in Computer Science
Distributed Algorithms and Systems

Lecture 4
“The PageRank Citation Ranking: Bringing Order to the Web”

L. Page, S. Brin, R. Motwani, T Winograd

1998
Origin of “Google”

- **Googol**
  - $10^{100}$

- **Motivation behind**
  - Human maintained indices such as Yahoo!
  - Explosive growth

http://news.netcraft.com
Design Goals of Google

- Improved search quality
  - In 1997, 1 out of 4 top search engines found itself
  - High precision in finding relevant document was necessary

- Academic search engine research
  - Search engine technology went commercial: an black art
  - To build systems that a good number of people could use
  - To build an architecture to support novel research on large-scale Web data
Weakness of Existing Approaches

- Calculate similarities
  - Based on flat, vector-space model of each page
  - Prone to cheating (Web spamming or search engine persuasion)
Basic Idea of PageRank

- Exploit the topological structure of hypertextual systems
Simple Example
Related Work

- Academic citation analysis
  - Similarities
    - Graph structure;
      paper = node, web page = node
      citation = link, URL = link
    - “node” authority independent of “node” content
  - Differences
    - Uniform unit of info (paper) versus
great variability in quality, usage, citations, and length
    - Equal link weight vs variable importance
      A backlink from Yahoo! vs. from a friend
Which Page Should Be Ranked Higher?

A

B

John Doe
Simple Expression

\[ R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \]

- \( R(u) \): page rank of \( u \)
- \( B_u \): set of pages pointing at \( u \)
- \( N_u \): out-degree of \( u \)

**Question:** role of \( c \)?

**Answer:** total rank of all web pages constant
Dangling links

- Pages without outgoing pointers
  - Example: Pages not yet downloaded
- Do not affect the calculation much
  - Remove them, calculate ranks, and add them back
Loop

Question: ranks of A, B, and C?
Answer: infinite! (rank sink)
Basic Algorithm

\[ R(u) = d \sum_{v \in B_u} \frac{R(v)}{N_v} + (1 - d) \]

- \( R(u) \): page rank of \( u \)
- \( B_u \): set of pages pointing at \( u \)
- \( N_u \): out-degree of \( u \)
- \( d \): dumping factor
Matrix Representation

\[ R = d W R + (1 - d)I \]

where \( W = \{w_{i,j}\} \) and \( w_{i,j} = \frac{1}{N_j} \)

Question: Where to start?
Iterative Algorithm

\[ R(t) = d \ W \ R(t - 1) + (1 - d)I \]

where \( W = \{w_{i,j}\} \) and \( w_{i,j} = 1/N_j \)

Question: Will it converge?
Example

Figure 1. Directed graph representing web of six pages
Turn the Problem into a Markov Process

\[
P = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}.
\]

[LM04]
Evenly Split Rank of Dangling Links

\[ \bar{P} = \alpha \bar{P} + (1 - \alpha) ee^T / n = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix} \]
Final Solution

- Eigenvector of $P = \text{steady state rank}$
Fig. 3. The energy of the target community grows at least linearly with the number of pages of the promoting community, regardless of its pattern of connections. This makes it very hard to detect such a spamming method.

[BGS05]
Questions

- Where to start?
  - Find a nondegenerate start vector

- What if there are two pages that point to each other and no one else and there is a page that points to one of them?
  - Role of dumping factor guarantees no rank sink
References


