CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Graph Analysis and Page Rank
Review from previous Lecture

- We have looked the Page Rank in the form of
  - Flow Model Equations
  - Matrix Formulations
- Find out the Eigen vector form the Matrix Formulations for the Page rank
Random Walk Interpretation

\[ r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{\text{out}}(i)} \]
Imagine a random web surfer
- At any time $t$, the surfer is on some page $i$

Random Walk Interpretation

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{out}(i)}$$
Random Walk Interpretation

Imagine a random web surfer

- At any time $t$, the surfer is on some page $i$
- At time $t + 1$, the surfer follows an out-link from $i$ uniformly at random
  - Ends up on some page $j$ linked from $i$

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{out}(i)}$$
Imagine a random web surfer
- At any time $t$, the surfer is on some page $i$
- At time $t + 1$, the surfer follows an out-link from $i$ uniformly at random
  - Ends up on some page $j$ linked from $i$
- Process repeats infinitely

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{out}(i)}$$
Imagine a random web surfer

- At any time $t$, the surfer is on some page $i$
- At time $t + 1$, the surfer follows an out-link from $i$ uniformly at random
  - Ends up on some page $j$ linked from $i$
- Process repeats infinitely

$P(t)$ is the vector whose $i$th coordinate is the probability that the surfer is at page $i$ at time $t$

So $P(t)$ is a probability distribution over pages
Random Walk Interpretation

● Where is the surfer at time $t+1$?
  ○ Follows a link uniformly at random
    $$p(t + 1) = M \cdot p(t)$$

- Suppose the random walk reaches a state
  $$p(t + 1) = M \cdot p(t) = p(t)$$
  then $p(t)$ is **stationary distribution** of a random walk

- Our original rank vector $r$ satisfies $r = M \cdot r$
  ○ So, $r$ is a stationary distribution for the random walk
Page RanK: Google Formulations
Google Formulation

\[ r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \]

or equivalently

\[ r = Mr \]
Google Formulation

\[ r_{j}^{(t+1)} = \sum_{i \rightarrow j} \frac{r_{i}^{(t)}}{d_{i}} \]

or equivalently

\[ r = Mr \]

Does this value converge?
Google Formulation

$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \quad \text{or equivalently} \quad r = Mr$$

Does this value converge?

Does it converge to the results that we want?
Google Formulation

\[ r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \]

or equivalently

\[ r = Mr \]

Does this value converge?

Does it converge to the results that we want?

Are the results reasonable?
Does this converge?

\[ r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \]
Does this converge?

\[ r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \]

**Example:**

\[
\begin{align*}
    r_a &= \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \\
    r_b &= \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}
\end{align*}
\]

Iteration 0, 1, 2, ...
Does this converge to what we want?

\[ r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i} \]
Does this converge to what we want?

Example:

\[
\begin{align*}
\mathbf{r}_a & = 1 \ 0 \ 0 \ 0 \ 0 \\
\mathbf{r}_b & = 0 \ 1 \ 0 \ 0 \ 0 \\
\end{align*}
\]

Iteration 0, 1, 2, …
Page Rank: Two Problems

1. Dead ends:
   - Some pages are **dead ends** (have no out-link)
   - Such pages cause important information to leak

2. Spider traps
   - All out-links are within the group
   - Random walk gets stuck in a trap
   - And eventually spider traps absorbs all importance
Spider Traps

m is a spider trap

\[
\begin{array}{ccc}
   y & a & m \\
   \frac{1}{2} & \frac{1}{2} & 0 \\
   \frac{1}{2} & 0 & 0 \\
   0 & \frac{1}{2} & 1 \\
\end{array}
\]

\[
r_y = r_y/2 + r_a/2 \\
r_a = r_y/2 \\
r_m = r_a/2 + r_m
\]
Spider Traps

- **Power Iteration:**
  - Set $r_j = 1$
  - $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$
  - And iterate

- **Example:**

\[
\begin{pmatrix}
    r_y \\
    r_a \\
    r_m
\end{pmatrix} =
\begin{pmatrix}
    1/3 & 2/6 & 3/12 & 5/24 & 0 \\
    1/3 & 1/6 & 2/12 & 3/24 & \ldots & 0 \\
    1/3 & 3/6 & 7/12 & 16/24 & 1
\end{pmatrix}
\]

*Iteration 0, 1, 2, …*

All the PageRank score gets “trapped” in node m.
Solution: Teleports!

- The Google solution for spider traps: At each time step, the random surfer has two options
  - With prob. $\beta$, follow a link at random
  - With prob. $1-\beta$, jump to some random page
  - Common values for $\beta$ are in the range 0.8 to 0.9
  - This will help the surfer to teleport out of spider trap within a few steps
Dead Ends

\[ r_y = r_y / 2 + r_a / 2 \]
\[ r_a = r_y / 2 \]
\[ r_m = r_a / 2 \]
Dead Ends

Power Iteration:
Set \( r_j = 1 \)

\[
 r_j = \sum_{i \to j} \frac{r_i}{d_i}
\]
And iterate

Here the PageRank “leaks” out since the matrix is not stochastic.
Solution: Teleports

Teleport with probability 1.0 at dead ends
Adjust matrix accordingly
Why Teleports Solve the Problem?

Why are dead-ends and spider traps a problem and why do teleports solve the problem?

- **Spider-traps** are not a problem, but with traps PageRank scores are not what we want
  - **Solution**: Never get stuck in a spider trap by teleporting out of it in a finite number of steps

- **Dead-ends** are a problem
  - The matrix is not column stochastic so our initial assumptions are not met
  - **Solution**: Make matrix column stochastic by always teleporting when there is nowhere else to go
Google Solution: Random Teleports

- **Google’s solution that does it all:**
  At each step, random surfer has two options:
  - With probability $\beta$, follow a link at random
  - With probability $1 - \beta$, jump to some random page

- **PageRank equation** [Brin-Page, 98]

\[
r_j = \sum_{i \rightarrow j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}
\]

This formulation assumes that $M$ has no dead ends. We can either preprocess matrix $M$ to remove all dead ends or explicitly follow random teleport links with probability 1.0 from dead-ends.
The Google Matrix

- **PageRank equation** [Brin-Page, ‘98]
  \[ r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N} \]

- **The Google Matrix** \( A \):
  \[ A = \beta M + (1 - \beta) \left[ \frac{1}{N} \right]_{N \times N} \]

- We have a recursive problem: \( r = A \cdot r \)

- And the Power method still works!

- **What is \( \beta \)?**
  - In practice \( \beta = 0.8, 0.9 \) (make 5 steps on avg., jump)

[1/N]_{N \times N}…N by N matrix where all entries are 1/N
Random Teleports ($\beta = 0.8$)

$y = \frac{1}{3}, \quad 0.33, \quad 0.24, \quad 0.26$

$a = \frac{1}{3}, \quad 0.20, \quad 0.20, \quad 0.18$

$m = \frac{1}{3}, \quad 0.46, \quad 0.52, \quad 0.56$

$A = \begin{bmatrix} 
\frac{7}{15} & \frac{7}{15} & \frac{1}{15} \\
\frac{7}{15} & \frac{1}{15} & \frac{1}{15} \\
\frac{1}{15} & \frac{7}{15} & \frac{13}{15} 
\end{bmatrix}$

$M = \begin{bmatrix} 
\frac{1}{2} & \frac{1}{2} & 0 \\
\frac{1}{2} & 0 & 0 \\
0 & \frac{1}{2} & 1 
\end{bmatrix} + 0.2 \begin{bmatrix} 
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} 
\end{bmatrix}$
References


[2] Chapter 5 Lin and Dyer