FAQs
- Midterm
  - 10/8/2018 Monday
  - 4:00PM ~ 5:15PM
  - Closed book
  - No computer/No cheat sheets
  - 5 problems
    - Problem 1: 40 points total, 20 T/F style questions
    - Problem 2: 15 points total, 3-6 T/F or 4 multiple choice questions
    - Problem 3: 15 points total, 5-6 T/F or 4 multiple choice questions
    - Problem 4: 15 points total, 4-5 multiple choice questions
    - Problem 5: 15 points total, 1-2 multiple choice questions and 1 short answer
  - Total: 100 points and 20% of your final grade

Topics
- Large-scale Analytics 1. Web-Scale Link and Social Network Analysis
  - Link spam

Architecture of a Spam Farm
- Spam Farm
  - A collection of pages whose purpose is to increase the PageRank of a certain page or pages
  - From the point of view of the spammer, the Web is divided into two parts
    - Inaccessible pages
      - The pages that the spammer cannot affect
      - Most of the Web
    - Accessible pages
      - Those pages that, while they are not controlled by the spammer, can be affected by the spammer

The Web from the point of view of the link spammer
Understanding Spam Farm (1/2)

- Setting the links to the target page
  - Without link from outside, the spam farm is not useful
  - e.g. Blogs or newspapers
    - Comments like "I agree. Please see my article at www.mySpamFarm.com"

Understanding Spam Farm (2/2)

- There is one page, the target page
  - Spammer attempts to place as much PageRank as possible

- There are a large number of supporting pages
  - Accumulate the portion of the PageRank that is distributed equally to all pages
  - The fraction $1-\beta$ of the PageRank that represents surfers going to a random page
  - Prevent the PageRank of $t$ from being lost
    - Note that all of the supporting pages link only to $t$

Analysis of a Spam Farm (1/6)

- A taxation parameter $\beta$
  - The fraction of a page's PageRank that gets distributed to its successors at the next round

- Let there be,
  - $n$ pages on the Web in total
  - A target page $t$
  - $m$ supporting pages

Analysis of a Spam Farm (2/6)

- Let $x$ be the amount of PageRank contributed by the accessible pages
  - $x$ is the sum over all accessible page $p$ with a link to $t$, of the PageRank of $p$ times $\beta$
divided by the number of successors of $p$

- Finally, let $y$ be the unknown PageRank of $t$

Analysis of a Spam Farm (3/6)

- The PageRank of each supporting page
  - $\beta m y / (m+1-\beta)/n$

- First term represents the contribution from $t$
  - $\beta y$ is distributed to $t$'s successors

- Second term is the supporting page's share of the fraction $1-\beta$ of the PageRank that is divided equally among all pages on the Web

Analysis of a Spam Farm (4/6)

- PageRank of $y$ of target page $t$ is (1)+(2)+(3)
  1. Contribution $x$ from outside

  2. $\beta$ times the PageRank of every supporting page
     $\beta m y / m+1-\beta/n$

  3. $(1-\beta)/n$, the share of the fraction $1-\beta$ of the PageRank that belongs to $t$
     This amount is negligible
Analysis of a Spam Farm (5/6)

- From (1) and (2),
  \[ y = x + \beta \left( \frac{\beta y}{m} \right) \left( 1 - \beta \right) = x + \beta y (1 - \beta) \frac{m}{n} \]
  \[ y = x + \beta + c \frac{m}{n} \]

  Where:
  \[ c = \beta (1 - \beta) / (1 - \beta^2) = \beta / (1 + \beta) \]

Analysis of a Spam Farm (6/6)

- If we choose \( \beta = 0.85 \), then \( 1/(1 - \beta^2) = 3.6 \)
  \[ c = \beta (1 + \beta) = 0.46 \]

- The structure has amplified the external PageRank contribution by 360% for each iteration
  
- Also, it obtained an amount of PageRank that is 46% of the fraction of the Web, \( m/n \), that is in the spam farm

Combatting Link Spam

- Detecting and eliminating link spam have been critical for search engines
  - Just as it was critical to eliminate term spam in the previous decade

- Detecting particular structures
  - Spam farm
    - One page links to a very large number of pages
    - Each of which links back to it

TrustRank

- TrustRank is a topic-sensitive PageRank
  - "topic" is a set of pages believed to be trustworthy (not spam)

- Develop a suitable teleport set of trustworthy pages
  - Let humans examine a set of pages and decide which of them are trustworthy

- Pick a domain whose membership is controlled
  - University pages
    - .mil, .gov
Calculating TrustRank (1/2)

- Then the topic-sensitive PageRank for $S$ is the limit of the iteration,

$$v' = \beta M v + (1 - \beta) e_S / |S|$$

- $M$ is the transition matrix of the Web, and $|S|$ is the size of set $S$

Calculating TrustRank (2/2)

- Suppose we use $\beta = 0.8$, and our trust rank is represented by the teleport set (trustworthy pages) $S = \{B, D\}$

<table>
<thead>
<tr>
<th>Web Page</th>
<th>PageRank</th>
<th>TrustRank</th>
<th>SpamMass</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3/9</td>
<td>54/210</td>
<td>0.229</td>
</tr>
<tr>
<td>B</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
<tr>
<td>C</td>
<td>2/9</td>
<td>38/210</td>
<td>0.186</td>
</tr>
<tr>
<td>D</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
</tbody>
</table>

Spam Mass

- Measures the fraction of its PageRank that comes from spam for each page

- For an arbitrary page $p$,
  - PageRank $r$
    - Computing ordinary PageRank
  - TrustRank $t$
    - Computing the TrustRank based on some teleport set of trustworthy pages
  - The spam mass $< r - t > r$

- A negative or small positive spam mass
  - $p$ is probably not a spam page

Example

- Suppose that both the PageRank and TrustRank were computed
- Teleport set was page $B$ and $D$
  - Which nodes are not the link spams?
  - Is there any link spam?

<table>
<thead>
<tr>
<th>Web Page</th>
<th>PageRank</th>
<th>TrustRank</th>
<th>SpamMass</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3/9</td>
<td>54/210</td>
<td>0.229</td>
</tr>
<tr>
<td>B</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
<tr>
<td>C</td>
<td>2/9</td>
<td>38/210</td>
<td>0.186</td>
</tr>
<tr>
<td>D</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
</tbody>
</table>

- Suppose that both the PageRank and TrustRank were computed
- Teleport set was page $B$ and $D$
  - Which nodes are not the link spams?
    - $B$ and $D$
  - $C$ has lower chance to be the link spam compared to $A$