Analysis of Large Graphs: TrustRank and WebSpam
Can we measure popularity within a topic?

**Goal:** Evaluate Web pages by how close they are to a particular topic, e.g. “sports” or “history”

**Allows search queries to be answered based on interests of the user**

**Example:** Query “Jaguars” wants different pages depending on whether you are interested in sports, animals, cars, or Apple computers
Random walker has a small probability of teleporting at any step

**Teleport can go to:**

- **Standard PageRank:** Any page with equal probability
  - To avoid dead-end and cycle problems
- **Topic Specific PageRank:** A topic-specific set of “relevant” pages (teleport set)

**Idea: Bias the random walk**

- When walker teleports, she pick a page from a set $S$
- $S$ contains only pages that are relevant to the topic
  - E.g., Open Directory (DMOZ) pages for a given topic/query
- For each teleport set $S$, we get a different vector $r_S$
Matrix Formulation

- To make this work all we need is to update the teleportation part of the PageRank formulation:

\[ A_{ij} = \begin{cases} 
\beta M_{ij} + (1 - \beta)/|S| & \text{if } i \in S \\
\beta M_{ij} + 0 & \text{otherwise}
\end{cases} \]

- \( A \) is stochastic!
- We weighted all pages in the teleport set \( S \) equally
  - Could also assign different weights to pages!
- **Compute as for regular PageRank:**
  - Multiply by \( M \), then add a vector
  - Maintains sparseness
Suppose $S = \{1\}$, $\beta = 0.8$

<table>
<thead>
<tr>
<th>Node</th>
<th>Iteration</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>stable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.25</td>
<td>0.4</td>
<td>0.28</td>
<td>0.294</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.25</td>
<td>0.1</td>
<td>0.16</td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.25</td>
<td>0.3</td>
<td>0.32</td>
<td>0.327</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.25</td>
<td>0.2</td>
<td>0.24</td>
<td>0.261</td>
<td></td>
</tr>
</tbody>
</table>

$S = \{1\}$, $\beta = 0.90$: $r = [0.17, 0.07, 0.40, 0.36]$

$S = \{1\}$, $\beta = 0.8$: $r = [0.29, 0.11, 0.32, 0.26]$

$S = \{1\}$, $\beta = 0.70$: $r = [0.39, 0.14, 0.27, 0.19]$

$S = \{1, 2, 3, 4\}$, $\beta = 0.8$: $r = [0.13, 0.10, 0.39, 0.36]$

$S = \{1, 2, 3\}$, $\beta = 0.8$: $r = [0.17, 0.13, 0.38, 0.30]$

$S = \{1, 2\}$, $\beta = 0.8$: $r = [0.26, 0.20, 0.29, 0.23]$

$S = \{1\}$, $\beta = 0.8$: $r = [0.29, 0.11, 0.32, 0.26]$
Create different PageRanks for different topics

- The 16 DMOZ top-level categories:
  - arts, business, sports,…

Which topic ranking to use?

- User can pick from a menu
- Classify query into a topic
- Can use the context of the query
  - E.g., query is launched from a web page talking about a known topic
  - History of queries e.g., “basketball” followed by “Jordan”
- User context, e.g., user’s bookmarks, …
TrustRank: Combating the Web Spam
What is Web Spam?

- **Spamming:**
  - Any deliberate action to boost a web page’s position in search engine results, incommensurate with page’s real value

- **Spam:**
  - Web pages that are the result of spamming
  - This is a very broad definition
    - SEO industry might disagree!
    - SEO = search engine optimization

- Approximately **10-15%** of web pages are spam

How do you make your page appear to be about movies?

(1) Add the word movie 1,000 times to your page

Set text color to the background color, so only search engines would see it

(2) Or, run the query “movie” on your target search engine

See what page came first in the listings

Copy it into your page, make it “invisible”

These techniques are term spam
Google’s Solution to Term Spam

- Believe what people say about you, rather than what you say about yourself
  - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text

- PageRank as a tool to measure the “importance” of Web pages
Our hypothetical shirt-seller loose

Saying he is about movies doesn’t help, because others don’t say he is about movies.

His page isn’t very important, so it won’t be ranked high for shirts or movies.

Example:

Shirt-seller creates 1,000 pages, each links to his with “movie” in the anchor text.

These pages have no links in, so they get little PageRank.

So the shirt-seller can’t beat truly important movie pages, like IMDB.
Why it does not work?

Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »

Welcome to MichaelMoore.com!
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - Sep 1, 2005 - Cached - Similar pages

BBC NEWS | Americas | 'Miserable failure' links to Bush
Web users manipulate a popular search engine so an unflattering description leads to the president’s page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages

Google's (and Inktomi's) Miserable Failure
A search for miserable failure on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a ...
searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - Cached - Similar pages
SPAM FARMING
Once Google became the dominant search engine, spammers began to work out ways to fool Google

- **Spam farms** were developed to concentrate PageRank on a single page

- **Link spam:**
  - Creating link structures that boost PageRank of a particular page
Three kinds of web pages from a spammer’s point of view

- **Inaccessible pages**
- **Accessible pages**
  - e.g., blog comments pages
  - spammer can post links to his pages
- **Owned pages**
  - Completely controlled by spammer
  - May span multiple domain names
Link Farms

- **Spammer’s goal:**
  - Maximize the PageRank of target page \( t \)

- **Technique:**
  - Get as many links from accessible pages as possible to target page \( t \)
  - Construct “link farm” to get PageRank multiplier effect
One of the most common and effective organizations for a link farm

Analysis

- **x**: PageRank contributed by accessible pages
- **y**: PageRank of target page \( t \)
- Rank of each “farm” page is given by:
  \[
  \text{Rank} = \frac{\beta y}{M} + \frac{1-\beta}{N}
  \]
- \( y = x + \beta M \left[ \frac{\beta y}{M} + \frac{1-\beta}{N} \right] + \frac{1-\beta}{N} \)
- \( y = x + \beta^2 y + \frac{\beta(1-\beta)M}{N} + \frac{1-\beta}{N} \)
- \( y = \frac{x}{1-\beta^2} + c \frac{M}{N} \)
  where \( c = \frac{\beta}{1+\beta} \)

Very small; ignore

Now we solve for \( y \)
Analysis

\[ y = \frac{x}{1-\beta^2} + c \frac{M}{N} \quad \text{where} \quad c = \frac{\beta}{1+\beta} \]

For \( \beta = 0.85 \), \( 1/(1-\beta^2) = 3.6 \)

- Multiplier effect for acquired PageRank
- By making \( M \) large, we can make \( y \) as large as we want
Combating Spam

- **Combating term spam**
  - Analyze text using statistical methods
  - Similar to email spam filtering
  - Also useful: Detecting approximate duplicate pages

- **Combating link spam**
  - Detection and blacklisting of structures that look like spam farms
    - Leads to another war – hiding and detecting spam farms
  - **TrustRank** = topic-specific PageRank with a teleport set of trusted pages
    - Example: .edu domains, similar domains for non-US schools
**TrustRank: Idea**

- **Basic principle:** *Approximate isolation*
  - It is rare for a “good” page to point to a “bad” (spam) page

- Sample a set of *seed pages* from the web

- Have an *oracle* (*human*) to identify the good pages and the spam pages in the seed set
  - **Expensive task**, so we must make seed set as small as possible
Trust Propagation

- Call the subset of seed pages that are identified as **good** the **trusted pages**

- Perform a topic-sensitive PageRank with teleport set = trusted pages
  - Propagate trust through links:
    - Each page gets a trust value between 0 and 1

- **Solution 1:** Use a threshold value and mark all pages below the trust threshold as spam
Set trust of each trusted page to 1

Suppose trust of page $p$ is $t_p$

- Page $p$ has a set of out-links $o_p$
- For each $q \in o_p$, $p$ confers the trust to $q$
  - $\beta t_p / |o_p|$ for $0 < \beta < 1$

Trust is additive

- Trust of $p$ is the sum of the trust conferred on $p$ by all its in-linked pages

Similarity to Topic-Specific PageRank

- Within a scaling factor, $\text{TrustRank} = \text{PageRank}$ with trusted pages as teleport set
Why is it a good idea?

- **Trust attenuation:**
  - The degree of trust conferred by a trusted page decreases with the distance in the graph.

- **Trust splitting:**
  - The larger the number of out-links from a page, the less scrutiny the page author gives each out-link.
  - Trust is **split** across out-links.
Two conflicting considerations:

- Human has to inspect each seed page, so seed set must be as small as possible.

- Must ensure every good page gets adequate trust rank, so need make all good pages reachable from seed set by short paths.
Suppose we want to pick a seed set of $k$ pages

**How to do that?**

**(1) PageRank:**
- Pick the top $k$ pages by PageRank
- Theory is that you can’t get a bad page’s rank really high

**(2) Use trusted domains** whose membership is controlled, like .edu, .mil, .gov
Spam Mass

- In the TrustRank model, we start with good pages and propagate trust.

- Complementary view:
  What fraction of a page’s PageRank comes from spam pages?

- In practice, we don’t know all the spam pages, so we need to estimate.
Spam Mass Estimation

**Solution 2:**

- $r_p$ = PageRank of page $p$
- $r_p^+$ = PageRank of $p$ with teleport into trusted pages only

- **Then:** What fraction of a page’s PageRank comes from spam pages?
  \[ r_p^- = r_p - r_p^+ \]

- Spam mass of $p = \frac{r_p^-}{r_p}$
  - Pages with high spam mass are spam.