Have you ever used the Web…

- to get informed?
- to help you make decisions?
- Financial
- Medical
- Political
- Religious…

- Have you ever found something that is not correct in the top-10?
- How do you know that what you find is correct?
- Who are responsible for highly visible, incorrect information?

We depend on search engines to find information.

The effects of Web Spam on The Evolution of Search Engines

Any controversial issue will be spammed

The Web has Spam (not talking about email spam). . .
Political issues will be spammed

... you like it or not!

Why is there Web Spam?

What do Web Spammers do
Understanding S.E. History through Web Spam

1st Generation (ca 1994):
- AltaVista, Excite, Infoseek...
- Ranking based on Content
  + Pure Information Retrieval

2nd Generation (ca 1996):
- Lycos
- Ranking based on Content + Structure
  + Site Popularity

3rd Generation (ca 1998):
- Google, Teoma, Yahoo
- Ranking based on Content + Structure + Value
  + Page Reputation

In the Works
- Ranking based on "the user's need behind the query"

1st Generation: How to Spam

“Keyword stuffing”: Add keywords, text, to increase content similarity

- Page stuffed with casino-related keywords
- Keywords
- Well-formed sentences stitched together
- Links to keep crawlers going

1st Generation: Content Similarity

- Content Similarity Ranking:
  - The more rare words two documents share, the more similar they are
  
  - Documents are treated as "bags of words"
    (no effort to "understand" the contents)
  
  - Similarity is measured by vector angles
  
  - Query Results are ranked by sorting the angles between query and documents

How To Spam?

Really good synthetic content

- "Nigritude Ultramarine": An SEO competition

- Links to keep crawlers going
- Grammatically well-formed but meaningless sentences
  
  - Page stuffed with casino-related keywords
  - Keywords
  - Well-formed sentences stitched together
  - Links to keep crawlers going

- "Nigritude Ultramarine": An SEO competition
  
  - Links to keep crawlers going
  - Grammatically well-formed but meaningless sentences
  
  - Page stuffed with casino-related keywords
  - Keywords
  - Well-formed sentences stitched together
  - Links to keep crawlers going
2nd Generation: Add Popularity

- A hyperlink from a page in site A to some page in site B is considered a popularity vote from site A to site B.
- Rank similar documents according to popularity.

How To Spam?

2nd Generation: How to Spam

- Create "Link Farms": Heavily interconnected owned sites spam popularity.

Interconnected sites owned by vespro.com promote main site.

3rd Generation: Add Reputation...

- The reputation "PageRank" of a page $P_i$ is given by:
  $$ P_i = (1 - d) + d \cdot \sum_{j \in C_i} \frac{P_j}{C_j} $$
- Idea similar to academic co-citations.
- Beautiful Math behind it:
  - $PR$ = principal eigenvector of the web's link matrix.
  - $PR$ equivalent to the chance of randomly surfing to the page.

How To Spam?

3rd Generation: How to Spam

- Organize Mutual Admiration Societies: "link farms" of irrelevant reputable sites.

Resource Partner Additions:

1. Visit and add site to your browser’s "TABS".
2. Go to: [http://www.vespro.com](http://www.vespro.com)
3. Create "Link Farms": Heavily interconnected owned sites spam popularity.
4. Visit and add site:
   - [http://www.vespro.com](http://www.vespro.com)
   - [http://www.vespro.com](http://www.vespro.com)
   - [http://www.vespro.com](http://www.vespro.com)

3rd Generation: How to Spam

- Organize Mutual Admiration Societies: "link farms" of irrelevant reputable sites.

How To Spam?

1. Copy this code to your website:
   ```
   \begin{align*}
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \text{PR}_i = \frac{1}{n} \sum_j \text{PR}_j \cdot \text{C}_j \\
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \frac{1}{n} \sum_i \text{PR}_i = \text{PR}_i \\
   \end{align*}
   
   \text{PR} = \text{PR}_i
   \end{align*}
   ```
2. Copy this code to your website:
   ```
   \begin{align*}
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \text{PR}_i = \frac{1}{n} \sum_j \text{PR}_j \cdot \text{C}_j \\
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \frac{1}{n} \sum_i \text{PR}_i = \text{PR}_i \\
   \end{align*}
   
   \text{PR} = \text{PR}_i
   \end{align*}
   ```
3. Copy this code to your website:
   ```
   \begin{align*}
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \text{PR}_i = \frac{1}{n} \sum_j \text{PR}_j \cdot \text{C}_j \\
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \frac{1}{n} \sum_i \text{PR}_i = \text{PR}_i \\
   \end{align*}
   
   \text{PR} = \text{PR}_i
   \end{align*}
   ```
4. Copy this code to your website:
   ```
   \begin{align*}
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \text{PR}_i = \frac{1}{n} \sum_j \text{PR}_j \cdot \text{C}_j \\
   \text{PR} = \frac{1}{n} \sum_i \text{PR}_i \\
   \frac{1}{n} \sum_i \text{PR}_i = \text{PR}_i \\
   \end{align*}
   
   \text{PR} = \text{PR}_i
   \end{align*}
   ```

Thank You.
An Industry is Born

- "Search Engine Optimization" Companies
- Advertisement Consultants
- Conferences

Marketing Blasts
Free Thousands of Targeted Visitors to your website and boost your Google PageRank

3rd Generation: Reputation & Anchor Text

- Anchor text tells you what the reputation is about

How To Spam?

- www.ibm.com

Joe’s computer hardware link
- IBM
- HP

Big Blue today announced record profits for the quarter

Google-bombs” spam Anchor Text...

- Business weapons
  - “more evil than satan”
- Political weapon in pre-election season
  - “miserable failure”
  - “waffles”
  - “Clay Shaw” (+ 50 Republicans)
- Misinformation
  - Promote steroids
  - Discredit ADHD research
- Activism / online protest
  - “Egypt”
  - “Jew”
- Other uses we do not know?
  - “views expressed by the sites in your results are not in any way endorsed by Google…”

... mostly for political purposes

Activists openly collaborating to Google-bomb search results of political opponents in 2006...
Spammers are kept busy

Term Spamming
- Keyword stuffing
- Synthetic page creation
- Re-purposed content
- Blog content creation for spam

Link Spamming
- Link farms
- Mutual admiration societies (link exchanges)
- Expired high-ranked domains
- Post links to high-quality blogs

Hiding Spam:
- Content hiding (making it invisible to humans)
- Cloaking (making it invisible to SEs) by serving different pages for the same URL
- Redirecting through meta-refresh

Search Engines vs Web Spam

Search Engine’s Action
1st Generation: Similarity
- Content
2nd Generation: + Popularity
- Content + Structure
3rd Generation: + Reputation
- Anchor Text
- Content + Structure + Value
4th Generation (in the Works)
- Ranking based on the user’s “need behind the query”

Web Spammers Reaction
- Add keywords so as to increase content similarity
- Create “link farms” of heavily interconnected sites
- Organize “mutual admiration societies” of irrelevant reputable sites
- Googleboms

Is there a pattern on how to spam?

The World According to YOU

friends and family
trusted advisors (religious, political)
trusted sources (news, books)
adverts and infomercials

The Web

We interrupt our program to discuss something completely different...

The World According to YOU

friends and family
trusted advisors (religious, political)
trusted sources (news, books)
adverts and infomercials

The Web

Your TRUST network

friends and family
trusted advisors (religious, political)
trusted sources (news, books)
adverts and infomercials
Your Trust Network

- Network of Nodes and Arcs (directed edges)
  - Nodes = social entities (people, entities, sources, ideas)
  - Arcs = trust relationships from an entity to another
    - Length of arc = strength of trust
  - We can explore it (mentally)
  - We change/verify/augment it all the time

Your network:
- YOU
- Mom
- Partner
- Famous Actress
- Prof. X
- NYTimes
- Rev. Y
- Joe (a plumber)
- US Pres.
- Democracy
- Your Boss

Societal Trust is (also) a Graph

<table>
<thead>
<tr>
<th>Graph Theory</th>
<th>Web Graph</th>
<th>Trust Social Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node weight</td>
<td>rank (accord. to a search engine)</td>
<td>social entity reputation (accord. to a person)</td>
</tr>
<tr>
<td>weight computation</td>
<td>ranking formula (e.g., pagerank)</td>
<td>dissonant (e.g., 2 recommenders)</td>
</tr>
<tr>
<td>Arc semantics</td>
<td>“vote of confidence”</td>
<td>computed on demand</td>
</tr>
<tr>
<td>weight</td>
<td>degree of confidence</td>
<td></td>
</tr>
<tr>
<td>weight range</td>
<td>[0, 1]</td>
<td></td>
</tr>
</tbody>
</table>

CHALLENGES to your Trust Network

- By your friends and family
- By teachers and colleagues
- By trusted advisors
- By trusted sources
- By others
- By ads

YOU
- Devastation
- Famous Actress
- Joe (a plumber)

Challenges of your Trust Network through Propaganda

- Propaganda:
  - Attempt to modify human behavior; and thus influence people’s actions in ways beneficial to propagandists
- Theory of Propaganda
  - Developed by the Institute for Propaganda Analysis 1938-42
- Propagandistic Techniques (and ways of detecting propaganda)
  - Word games: associate good/bad concept with social entity
  - Glittering Generalities: Name Calling
  - Transfer: use special privileges (e.g., office) to breach trust
  - Testimonial: famous non-experts' claims
  - Plain Folk: people like us think this way
  - Bandwagon: everybody’s doing it, jump on the wagon
  - Card Stacking: use of bad logic
The Bandwagon Technique

- with it the propagandist attempts to convince us that all members of a group to which we belong are accepting his program and that we must therefore follow our crowd and "jump on the band wagon."

The Testimonial Technique

- having some respected person say that a given idea or program or product or person is good or bad

Propaganda in Graph Terms

- Word Games
  - Name Calling
  - Glittering Generalities
- Transfer
- Testimonial
- Plain Folk
- Card stacking
- Bandwagon
- Modify Node weights
- Decrease node weight
- Increase node weight
- Modify Node content - keep weights
- Insert Arcs b/w irrelevant nodes
- Modify Arcs
- Mislabel Arcs
- Modify Arcs & generate nodes

Web Spammers as Propagandists

- Web Spammers can be seen as employing propagandistic techniques in order to modify the Web Graph

<table>
<thead>
<tr>
<th>S.E.'s</th>
<th>Ranking</th>
<th>Spamming</th>
<th>Propaganda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gen</td>
<td>Doc</td>
<td>keyword stuffing</td>
<td>glittering generalities</td>
</tr>
<tr>
<td></td>
<td>Similarity</td>
<td>+ link farms</td>
<td></td>
</tr>
<tr>
<td>2nd Gen</td>
<td>+ Site popularity</td>
<td>+ mutual admiration societies</td>
<td>+ bandwagon</td>
</tr>
<tr>
<td>3rd Gen</td>
<td>+ Page reputation</td>
<td>+ link bombs</td>
<td>+ testimonials</td>
</tr>
<tr>
<td></td>
<td>+ anchor text</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>