Propaganda, Spam and the Evolution of Search Engines

Panagiotis Takis Metaxas
@takis_metaxas
Professor, Computer Science, Wellesley College
Affiliate, CRCS, Harvard University

AACAC, October 26, 2012
Who do you consult for answers?

- Myself and my experiences
- A friend or family member
- A religious authority
- A professor
- Someone else
- A book or librarian
- A newspaper
- An ad that I like
- The Internet

How did you decide who to ask:

- Are they knowledgeable on the issue?
- Can I trust them?
- Is it convenient to ask them?
Your Trust Network

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

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

How is your trust network challenged?
Challenges 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 (Glittering Generalities & Name Calling)** - associate good/bad concept with social entity
- **Plain Folk** - people like us think this way
- **Bandwagon** - everybody’s doing it, jump on the wagon
- **Testimonial** - famous non-experts’ claims
- **Card Stacking** - use of bad logic
- **Transfer** - use special privileges (e.g., office) to breach trust
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

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

- Network of Nodes and Arcs (directed edges)
  - Nodes = web pages
  - Arcs = hyperlinks from a page to another
- A network can be explored
- A network can be indexed

![Diagram of a network with nodes and arcs]

**URL**

http://cs.brynmawr.edu/content/wginfo

**Access method** **Server and domain** **Path of Document**

---

**AALAC/Mellon 23 Working Group on Information**

**Fall 2012 Meeting**

**AALAC/Mellon 23 Working Group on Information**

Co-Sponsored by the Center for Science of Information

October 26-27, 2012

Bryn Mawr College

Bryn Mawr, PA 19010

Maps & Directions

Parking: Between Buildings 30 & 40 on Campus Map

Keynote speaker: Prof. Sanjeev Kulkarni, Princeton University & Center for Science of Information, Title: Science of Information Meets the Liberal Arts

**Workshop Leaders**

Desak Kunar, Computer Science – Bryn Mawr College (Project Leader)

Diana Boley, Computer Science – Williams College

Tzu-Yi Chen, Computer Science – Pomona College

Paragupta T. Melawe, Computer Science – Wesley College

Richard Wicentowski, Computer Science – Swarthmore College

**Agenda:**

All sessions will be held in Room 300, Dalton Hall (Building 30 on Campus Map)
Use Web, Search Engines, Twitter?

- 74% of the Americans had internet access in May, 2008
- More than half of connected Americans use broadband
- 54% of all US adults went online to get news or information about the 2010 midterm elections
- 89% of all internet users use a search engine regularly
- 6% of US adults received candidate or campaign information on social networking sites or Twitter

What does this tell you about the importance of Search Engines and Social Networks?
Your Trust Network 2.0

- Your Trust Network is greatly increased
- You access it through Search Engines
- A few of us, also through Social Networks

The Web

Prof. X

NYTimes

Your Boss

Mom

Rev. Y

YOU

Democracy

Famous Actress

Joe (a plumber)
Reliability of Information on the Web

Anyone can be an author on the web!
In the past, the greatest problem facing researchers was finding information; now, with the advent of the Internet, the greatest problem is evaluating the vast wealth of information available.

Students [...] placed greater emphasis on the process of finding an answer than on analyzing the actual information.
The Web has misinformation!

Search results steroid drug HGH (human growth hormone)
Any controversial issue will be targeted.

Search results for mental disease ADHD (attention-deficit/hyperactivity disorder)

Tip: Have a question? Ask the researchers at Google Answers.

Texans for Safe Education Resolution
Texans for Safe Education Is ADHD A Real Disease? Dr. Fred Baughman is a neurologist who has discovered real diseases. By directly ...
www.wildestcolts.com/safeEducation/real.html - 8k - Cached - Similar pages

Truths About ADHD and Stimulant Drugs
... Informative Websites. www.adhdfraud.org - Dr. Fred Baughman's excellent website, containing the best of his essays revealing that ADHD is not a real disease ...
www.wildestcolts.com/mentalhealth/stimulants.html - 14k - Cached - Similar pages
[ More results from www.wildestcolts.com ]

To all website visitors: I urge one and all concerned with the fraud of Attention Deficit Hyperactivity Disorder (ADHD)--Compiled by ... Making "disease" (real diseases--epilepsy, brain tumor, multiple sclerosis, etc.) or ...
www.adhdfraud.org/ - 21k - Jan 29, 2005 - Cached - Similar pages
[ More results from www.adhdfraud.org ]

ADD and ADHD Fraud. Find out the truth about ADD and ADHD. ...
... the fraud of Attention Deficit Hyperactivity Disorder (ADHD)--Compiled by ... Making "disease" (real diseases--epilepsy, brain tumor, multiple sclerosis, etc.) or ...
www.adhdfraud.org/ - 21k - Jan 29, 2005 - Cached - Similar pages
[ More results from www.adhdfraud.org ]

ADHD a Real Disorder
... Some people say, ADHD is not a real disorder because ... He says: "ADHD is a disorder that cannot be identified in the same way as polio, heart disease or other ...
web4health.info/en/answers/adhd-real-disorder.htm - 20k - Cached - Similar pages

Baughman Dispels The Myth of ADHD
... The "disease," Baughman tells Insight, "is a total 100 percent ... to bring an end to the ADHD diagnosis ... as an adult and child neurologist, diagnosing real diseases. ...
www.becomeshealthynow.com/article/dirty/209 - 62k - Cached - Similar pages
Why?

- If you do not **rank high** in search results, you may well not exist!

- **Web spam:**
  - Attempt to **modify** the web (its structure and contents), and thus **influence** search engine results in ways **beneficial** to web spammers
Ranking is The Important Decision

- Crawl the web
- Create inverted index
- Document IDs
- Search engine servers
- Inverted index

Rank results

User query
Evolution of Search Engines

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

4th Generation? (In the Works)
- Ranking based on “the need behind the query”
1st Generation: Content Similarity

Content Similarity Ranking:
The more *rare* words two documents share, the more similar they are

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>157</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brutus</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>232</td>
<td>227</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>worse</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Documents are treated as “**bags of words**”
(no effort to “understand” the contents)

Term frequencies are computed
(Above: a few rare words in W.S. books)

Very common words ("a", "of", "the", "in", ...) are ignored
1st Generation: Content Similarity

**Content Similarity Ranking:**
*The more rare words two documents share, the more similar they are*

<table>
<thead>
<tr>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>worse</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Above: **Normalized** term frequencies.
- Results are **ranked** by sorting the sum of normalized frequencies of the query terms

**Query:** “In the mercy of Caesar”  
**Rank Results:**

**How To Spam?**
"Keyword stuffing": Add keywords, text, to increase content similarity.
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.

- For example, if these sites had the same content similarity, the highest ranking result would be:

  ![Diagram of web pages and links]

  - www.aa.com: 1
  - www.bb.com: 3
  - www.cc.com: 1
  - www.dd.com: 2
  - www.zz.com: 0

**How To Spam?**
2nd Generation: How to spam

Create “Link Farms”:
Heavily interconnected sites spam popularity

Interconnected sites owned by vespro.com promote main site
3rd Generation: Add Reputation...

- The reputation "PageRank" of a page $P_i$ = the sum of a fraction of the reputations of all pages $P_j$ that point to $P_i$

- Idea similar to academic **co-citations**

- Beautiful Math behind it

- The PageRank of a page
  - Can be calculated by a sequence of matrix multiplications
  - It is equivalent to the probability of randomly surfing to a page

**How To Spam?**
3rd Generation: How to spam

Organize Mutual Admiration Societies:
“link farms” of irrelevant reputable sites

Resource Partner Additions

Currently we are only adding websites to this Resource page with PR5 or higher AND WHOSE LINKING PAGE IS PR5 OR HIGHER. The first thing we check when we receive a link request is the PageRank. If the page you are planning to put our link on does not have a PR5 or higher we delete the request without a response. Please honor this request. Those websites that meet the above requirements are added within a few days.

To add your website to our Directory simply do 3 things:

1) Copy this code to your website:

<!-- Start Copying here -->

<A HREF="http://www.1st-Hgh.com" TARGET="_blank">1st Hgh, Homeopathic Human Growth Hormone Spray</A>

1st Hgh, Homeopathic Human Growth Hormone Spray - The primary hormone in the body, information and sales. Hgh is the ultimate youth reviving hormone. For everyone over 30 years old. This product is safe, inexpensive, and comes with a full money back guarantee.

<!-- End Copying here -->

2) E-Mail us the url where we can view our link on your website (it must be PR4 or higher!);

3) include in this e-mail your website url, title, and description.

Thank You, Partner!
An Industry is Born

“Search Engine Optimization”

Advertisement Consultants

Conferences

Marketing Effects
Reciprocal Link Tools and Website Positioning Strategies

Drive Thousands of Targeted Visitors to your website and boost your Google Pagerank

Why Having Links to Your Website Are Important

Having links from other websites is very important, yet, it's one of the most overlooked strategies for increasing traffic and for increasing your position in the search engines.

As soon as you launch a website, you should spend some time working on a strategic web site linking strategy.

By working on getting sites to link to you, you'll not only see an increase in traffic, but you'll also see your website achieve better listings in the search engines, especially the most important search engine, Google.

How to find good sites to link to you

It's all about getting the right sites to link to you. Just remember quality before quantity. That means to avoid links farms and FFAs like the plague. Many search engines will penalize you for having a link from a link farm. Ask yourself, "Would I want people to associate me with this website?" If the answer is no, then move on to the next site. With over 40,000,000 websites, you should be able to find some high quality websites willing to link to you.

Before you get started, I recommend these two tools, as they'll make finding the quality sites you want to be linked to much easier.

The first is Google Toolbar. With the Google Toolbar, you'll be able to see how important Google thinks the page is.

The second is Alexa Toolbar. With Alexa, you can see how popular the site is. Alexa ranks every site based on how often other toolbar users visit the site. It's an excellent tool, and the best part about both of these tools are that they're absolutely free.

The next step is to set up your favorite search engine (I recommend Google since it's...
About me

I was born July 23, 1977 in Albany, NY. I received a BA from Reed College in biochemistry and molecular biology in 1999, an MS from Cornell University in biophysics in 2002, and an MS from Stanford University in computational and mathematical engineering in 2009. I'm currently working at Genia Technologies Inc., analyzing single-molecule measurements of DNA threaded through protein nanopores embedded in phospholipid bilayers. I also have a part-time position in the Stanford University Radiology Department with the Informatics in Imaging Sciences (ISIS) group, developing the BIMM biomedical image metadata manager and ATM Aim Template Manager applications (pre-release version only).
Link exchanges
Creative ideas for valentine's day gifts and christening gift idea including holiday office party ideas

Posted by Holiday office party ideas and Creative Ideas for Valentine's Day gifts on August 29, 2005 at 01:26:26:

Spam Detection through Statistics

Web page in-degrees follow a power law distribution.
Word count variance

Variance of word counts on given host (0 values offset by 1E-7 to cope with log scale)
3rd Generation: Reputation & Anchor Text

- Anchor text tells you what the reputation is about.
- It adds synonyms for a term.

Page A

Anchor

Page B

www.wellesley.edu

Founded in 1875, Wellesley College is known...

US News Best Liberal Arts Colleges:
- Williams College
- Wellesley College
- Pomona College

The contributions of the nation’s women’s colleges, in the education...

How To Spam?
“Google-bombs” spam Anchor Text...

“miserable failure hits Obama in January 2009

Political Google-bombs
... for fun and profit!

- Business weapons
  - “more evil than satan”

- Political weapon in pre-election season
  - “miserable failure”
  - “waffles”
  - “Clay Shaw” (+ 50 Republicans)

- Misinformation
  - Promote steroids
  - Discredit AD/HD 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…”
There is a pattern on How to Spam!

Web spammers are employing propagandistic techniques in order to modify the Trust Network of Search Engines

<table>
<thead>
<tr>
<th>S.E.’s</th>
<th>Ranking</th>
<th>Spamming</th>
<th>Propaganda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt; Gen</td>
<td>Doc Similarity</td>
<td>Keyword Stuffing</td>
<td>Word Games</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Gen</td>
<td>+ Site Popularity</td>
<td>+ Link Farms</td>
<td>+ Bandwagon</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Gen</td>
<td>+ Page Reputation</td>
<td>+ Mutual Admiration Societies</td>
<td>+ Testimonials</td>
</tr>
<tr>
<td></td>
<td>+ Anchor Text</td>
<td>+ Link Bombs</td>
<td>+ Card Stacking</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt; Gen</td>
<td>+ “user need behind search”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Google to pay US $500m in drug ad inquiry
Admits aiding Canadian sales

August 25, 2011 | By Hiawatha Bray, Globe Staff

In one of the biggest settlements of its kind in US history, Google Inc. has agreed to pay $500 million to the Justice Department and admitted that the company's advertising service helped online pharmacies sell prescription drugs illegally.

US attorney for Rhode Island Peter F. Neronha revealed the agreement with the Internet search giant at a press conference in Providence yesterday. Neronha’s office led an investigation of Google’s role in the online pharmacy ads that began in 2009 and involved the Food and Drug Administration and the Internal Revenue Service, as well as Rhode Island authorities.

Ads by Google

**Free Prescription Savings**
Save up to 75% on your prescription Just print our 56,000 discounts

www.yourrxdiscounts.com
So, how do you decide what is reliable?
CHALLENGES to your Trust Network

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

You

Mom

Partner

Famous Actress

Prof. X

Rev. Y

US Pres.

NYTimes

Your Boss

Democracy

Famous Actress

Partner

Joe
(a plumber)
How do you know what you know?

- **Extrinsic reasons**
  - Trust in the entity supporting the information
  - The majority of people use it extensively
  - Technology can help here

- **Intrinsic reasons**
  - Own experience
  - Own ability to **think critically**, which means:
    - Understanding the Scientific Method and apply it habitually on important matters

- **But this is tough and requires Education**
We also “know”...

- What we learned as children
- What we think we remember correctly
  - There is no database in our brain, but we recreate memories every time we remember them
- What we misunderstood
  - Our brain is a pattern matching machine, we find similarities even where there is none
- What we think under the influence of substances, of voices in sleep, lack of sleep
- What we decide under fear, anger, personal interest, using illogical processes

__________ (Add your own examples here)
Are we so stupid?

- Our brain is impressively complicated, but it is not perfect.
- It is influenced by construction limitations and errors, our feelings, our senses, our environment.
- It was created through an ongoing evolutionary process. Some parts are old and are activated immediately, others are newer and demand great energy to get activated.
- We need to feel that we are in control of our environment. We do not easily accepting random phenomena, we want to discover a couple reasons explaining them.
- Critical thinking uses neocortex, large in size and requiring lots of energy to operate. We try to avoid it by creating heuristics, stereotypes, personal ways of “thinking”.
The take-home message

- **Web misinformation** (like propaganda and advertisement) is here to stay.

- **Finding** information is easy, **evaluating** it is more difficult than it used to.
  - Our trusted sources are being challenged, fooled.

- **Think critically** in terms of your trust network.
  - Understand what you believe and why.

- In the 21st century the basic survival skills:
  - Reading & Writing
  - Thinking critically
  - Understand technology.
Thank you!

2012

Knowledge-Based Education – We oppose the teaching of Higher Order Thinking Skills (HOTS) (values clarification), critical thinking skills and similar programs that are simply a relabeling of Outcome-Based Education (OBE) (mastery learning) which focus on behavior modification and have the purpose of challenging the student’s fixed beliefs and undermining parental authority.

Report of

Platform Committee