How Google* Works
(& why you should care)

*and Yahoo, and MSN, and ...

Panagiotis Takis Metaxas
Computer Science Department
Wellesley College

World Usability Day 2006
Have you used the Web...

- to get informed?
- to help you make decisions?
  - Financial
  - Medical
  - Political
  - Religious
  - Other?
- on your computer?
  - Your cell phone?
  - Your PDA?
  - Your thermostat?
  - Your toaster?

The Web is huge
- >10 billion static pages publicly available,
- ...growing every day
- Three times this size, if you count the “deep web”
- Infinite, if you count dynamically created pages

The Web is omnipresent
... but it can be unreliable

Anyone can be an author on the web!
50% of emails received at Wellesley College are spam!
The Web has Spam too!

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

Human Growth Hormone Natural HGH Supplement Decrease Fat and Gain ...
... Increase stamina. Improve immune functioning. Now, you can benefit from the premiere HGH ... information you need on hGH, human growth hormone hGH releaser, youth ...
www.hghplus.net/ - 23k - Cached - Similar pages

Human Growth Hormone - "Rejuvenation" Medicine?
... the list of the age-depleted hormones that can be replaced ... is natural; it is plentiful in the human body during ... weight gain of 30-50 pounds occur when HGH wanes ...
www.usdoctor.com/gh.htm - 17k - Cached - Similar pages

Secretagogue-One can increase - hgh, hGH - Human Growth Hormone ...
Maximum Human Performance (MHP ... almost embarrassed to carry Secretagogue-One, growth hormone and IGF ... Enhances all other hormone activities; Increases hGH and IGF1; ...
www.powersupplements.com/hgh.htm - 12k - Cached - Similar pages

Best HGH for Bodybuilding - HGH Supplement
... Yes, You Can Now Get Actual Human Growth Hormone ... Human Growth Hormone (HGH) is one of the most powerfully effective anabolic hormones available. ...
www.trulyhuge.com/hgh.htm - 14k - Jan 29, 2005 - Cached - Similar pages

Human Growth Hormone HGH - Pro HGH Supplement | Human Growth ...
... HGH) over the counter, however, you can improve your ... The same study found that administering human growth hormone (hgh) resulted in a 8.8% increase in lean ...
www.bodybuildingforyou.com/ pro-hormones/human-growth-hormone-hgh.htm - 33k - Cached - Similar pages

Buy HGH human growth hormone supplement, oral HGH hormone
Somatrop Rx is a legal HGH hormone supplement with the strongest potency allowed by law. ... Now you can get actual human growth hormones legally without ...
www.pheromone-secrets.com/hgh.html - 13k - Cached - Similar pages
Any controversial issue will be spammed!
... you like it or not!

But Google is usually so good in finding info... Why does it do that?
Why?

Web Spam:
- Attempt to modify the web (its structure and contents), and thus influence search engine results in ways beneficial to web spammers.
The Web is a Graph

- Directed Graph of Nodes and Arcs (directed edges)
  - Nodes = web pages
  - Arcs = hyperlinks from a page to another

- A graph can be explored
- A graph can be indexed
How Google (and the other search engines) Work

1. The web
2. Crawl the web
3. Create inverted index
4. Document IDs
5. Search engine servers
6. Rank results
7. User query
A Brief History 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

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.

- 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?**
1st Generation: How to Spam

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

Searching for Jennifer Aniston?

SEX SEXY MONICA LEWINSKY JENNIFER LOPEZ CLAUDIA SCHIFFER CINDY CRAWFORD JENNIFER ANNISTON GILLIAN ANDERSON MADONNA NIKI TAYLOR ELLE MACPHERSON KATE MOSS CAROL ALT TYRA BANKS FREDERIQUE KATHY IRELAND PAM ANDERSON KAREN MULDER VALERIA MAZZA SHALOM HARLOW AMBER VALLETTA LAETITA BETTIE PAGE HEIDI KLUM PATRICIA FORD DAISY FUENTES KELLY BROOK SEX SEXY MONICA LEWINSKY JENNIFER LOPEZ CLAUDIA SCHIFFER CINDY CRAWFORD JENNIFER ANNISTON GILLIAN ANDERSON MADONNA NIKI TAYLOR ELLE MACPHERSON KATE MOSS CAROL ALT TYRA BANKS FREDERIQUE KATHY IRELAND PAM ANDERSON KAREN MULDER VALERIA MAZZA SHALOM HARLOW AMBER VALLETTA LAETITA BETTIE PAGE HEIDI KLUM PATRICIA FORD DAISY FUENTES KELLY BROOK
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 sites spam popularity
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
  - $PR = \text{principal eigenvector of the web's link matrix}$
  - $PR$ 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

Currently we are only adding websites to this Resource page with PR6 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 -->

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

   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.</P>

   <!-- End Copying here -->

   It should look like this:

   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.

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” Companies
◆ Advertisement Consultants
◆ Conferences

Besides, don’t you have better things to do with your time than chase down links?
Like run your business and make money?

Yet not only is this club a huge time saver… many of these link pages are
undiscovered little gems… meaning you can gain a whole bunch of PageRank
from them as you won’t be sharing the PR with hundreds of other sites.

And get this…

There are even some PR6 pages I’ve found that hardly anyone knows about
and they’re more than willing to swap links with any decent website!

Are you starting to see the big picture here?

As if all that weren’t enough… what if I also made it so simple for you to conduct your
link campaign that it’s as easy as
copy, paste and forget it…

But wait…here’s the best part…

… what if I showed you how to do the whole thing in
… about 1 hour a month!

Yes, you read that right… about one hour a month is all it will take you to boost
your pagerank, increase your search engine rankings and easily grab more
traffic!

Now if you really do ‘get’ what I’m offering, you’re practically ready to knock me over
to get into the private, Members Only section of this Website, right?

(Just think…this entire, awesome program is that close at hand! Now I saw that Quit
trying to peak around the corner…you need a password to get in!)

Get your lifetime membership to the High PR Link Club for ONLY $97!
Unanswered Spam Attacks

- **Business weapons**
  - "more evil than satan"

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

- **Misinformation**
  - Promote hGH
  - Discredit AD/HD research

- **Activism / online protest**
  - "Egypt"
  - "Jew"

- **Other “Google bombs” we do not know?**
  - "...views expressed by the sites in your results are not in any way endorsed by Google..."
Is there a pattern on how to spam?

**Search Engine's Action**

- 1st Generation: *Similarity*
  - Content
- 2nd Generation: + *Popularity*
  - Content + Structure
- 3rd Generation: + *Reputation*
  - Content + Structure + Value

**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

**In the Works**

- Ranking based on “the need behind the query”

Can you guess what they will do?
And Now For Something Completely(?) Different

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
Societal Trust is (also) a Graph

- **Weighted Directed Graph of Nodes and Weighted Arcs**
  - Nodes = Societal Entities (People, Ideas, ...)
  - Arcs = Trust recommendation from an entity to another
  - Arc weight = Degree of entrustment

Then what is Propaganda?
- Attempt to modify the Societal Trust Graph in ways beneficial to propagandist

How to modify the Trust Graph?
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**
- **Misdirect 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

There is a pattern on how to spam!

<table>
<thead>
<tr>
<th>Gen</th>
<th>Technique Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gen</td>
<td>“keyword stuffing” to increase content similarity</td>
</tr>
<tr>
<td>2nd Gen</td>
<td>Create “link farms” of heavily interconnected sites</td>
</tr>
<tr>
<td>3rd Gen</td>
<td>Organize “mutual admiration societies” of irrelevant reputable sites</td>
</tr>
<tr>
<td></td>
<td>Create Google-bombs</td>
</tr>
</tbody>
</table>

- Word Games
- Band wagon
- Testimonials
- Card-stacking
Cognitive Hacking and Cyber Trust

- Gaining Access or Breaking into a computer system for the purpose of modifying certain behaviors of a human user in a way that violates the integrity of the overall system.

- Does not necessarily aim to fool a search engine.

- Famous examples:
  - Pump & Dump stock schemes
  - The Emulex case
  - Word Games
  - Transfer
How (not) To Solve The Problem

JUST GO TO www.criticalthinking.com AND CLICK ON "ANSWERS"!
Living in Cyberspace

- Critical Thinking, Education
  - Realize how do we know what we know
  - “Of course it’s true; I saw it on the Internet!”

- Cyber-social Structures that mimic Societal ones
  - Know why to trust or distrust
  - Who do you trust on a particular subject?

- A Search Engine per Browser
  - Easier to fool one search engine than to fool millions of readers
  - Enable readers to keep track of their trust network
  - Personalized tools of cyber trust
Thank You!

PMetaxas@wellesley.edu
http://www.wellesley.edu/CS/pmetaxas.html