

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

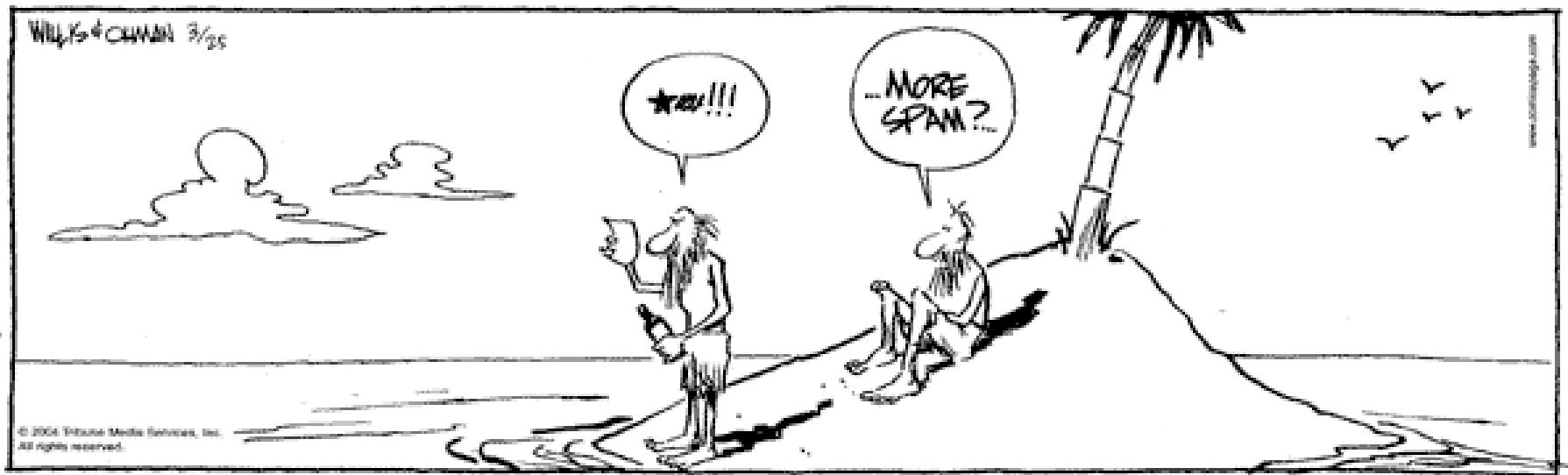
... but it can be unreliable

JUMP START by Robb Armstrong



Anyone can be an author on the web!

Email Spam anyone?



50% of emails received at Wellesley College are spam!

The Web has Spam too!



Web Images Groups New! News Froogle more »

can human growth hormone increase muscle mass?

Search

Advanced Search
Preferences

Web

Results 1 - 10 of about 313,000 for can human growth hormone increase muscle mass? (0.26 seconds)

[Human Growth Hormone](#)

www.HGH-Info-Buyers-Guide.org

Doctor Recommendations & Warnings. Find Info on Effective & Safe HGH!

Sponsored Link

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.hgplus.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.somatroprx.com/hgh.htm - 19k - [Cached](#) - [Similar pages](#)

Sponsored Links

[Human Growth Hormone](#)

US Prescribed HGH and Testosterone Amazing results. Great service.
www.AntiAgingGroup.com

[Human Growth Hormones](#)

Don't buy HGH until you read this Get the latest research and reviews
www.hgreview.net

[Muscle Without Steroids](#)

Step by step method to pack on new **muscle** every workout.
www.PrecisionTraining.com

[HGH - The Truth](#)

What They Don't Want You To Know What It Does To Your Body Affiliate
www.GenF20.com

[Muscle Growth](#)

Want Faster Muscle Growth?
Learn The Tips & Secrets
www.steroid-encyclopaedia.com

[Build muscle fast](#)

System used by 150,000+ worldwide Put on 30+ pounds of pure **muscle**!
www.BestWeightliftingProgram.com

[Muscle Growth Secrets](#)

Any controversial issue will be spammed!

The image shows a Google search results page for the query "is adhd a real disease?". The search bar contains the query. Below it, a note says "The following words are very common and were not included in your search: is a. [details]". The results are categorized under "Web" and show 10 results out of approximately 200,000. Each result includes a snippet, a link to the original page, and links to "Cached" and "Similar pages".

Web Results 1 - 10 of about 200,000 for [is adhd a real disease?](#). (0.18 seconds)

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 ...
MAKE CHADD ANSWER: IS ADHD A REAL DISEASE—YES OR NO? by Fred A. Baughman Jr., MD, May 26, 2001. Is **ADHD** a **real disease**—yes or no? ...
www.adhdfraud.org/commentary/5-27-01-1.htm - 5k - [Cached](#) - [Similar pages](#)

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 tellsInsight, "is a total 100 percent ... to bring an end to the **ADHD** diagnosis ... as an adult and child neurologist, diagnosing **real** diseases. ...
www.becomehealthynow.com/article/dirty/209 - 62k - [Cached](#) - [Similar pages](#)

... you like it or not!



Web Images Groups News Froogle^{New!} more »

miserable failure

Search

Advanced Search
Preferences

Web

Results 1 - 10 of about 255,000 for [miserable failure](#).

[Biography of President George W. Bush](#)

Home > President > Biography President George W. Bush En Español.

George W. Bush is the 43rd President of the United States. He ...

www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)

[Biography of Jimmy Carter](#)

Home > History & Tours > Past Presidents > Jimmy Carter. Jimmy Carter.

Jimmy Carter aspired to make Government "competent and compassionate ...

www.whitehouse.gov/history/presidents/jc39.html - 35k - [Cached](#) - [Similar pages](#)

[Michael Moore.com](#)

Click Here To Continue To MichaelMoore.Com "The War President". The

War President Photo courtesy of American Leftist, (click here ...

www.michaelmoore.com/ - 5k - [Cached](#) - [Similar pages](#)

[Senator Hillary Rodham Clinton: Online Office Welcome Page](#)

Dear Friend,. Thank you for visiting my on-line office! I appreciate

your interest in the issues before the United States Senate. ...

clinton.senate.gov/ - 9k - Apr 17, 2004 - [Cached](#) - [Similar pages](#)

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



Why?

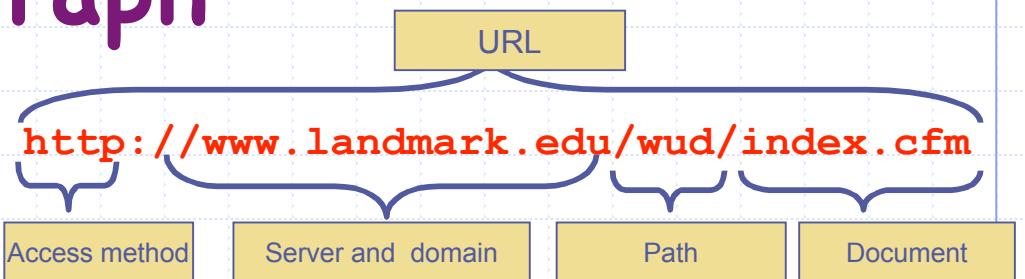


◆ Web Spam:

- Attempt to **modify** the web (its structure and contents), and thus **influence** search engine results in ways **beneficial** to web spammers

How do they do it?

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**

 World Usability Day
November 14 2006

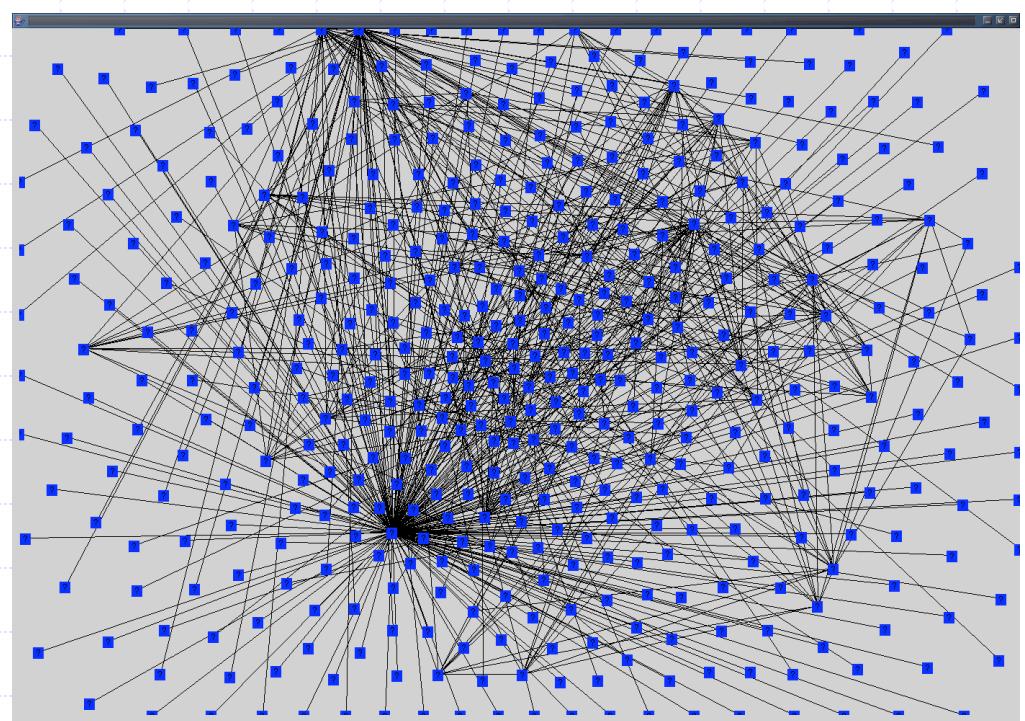
[Home](#)
[Program Details](#)
[Registration](#)
[Logistics](#)
[Accommodations](#)

World Usability Day New England 2006 *Universal Usability for Teaching and Learning*

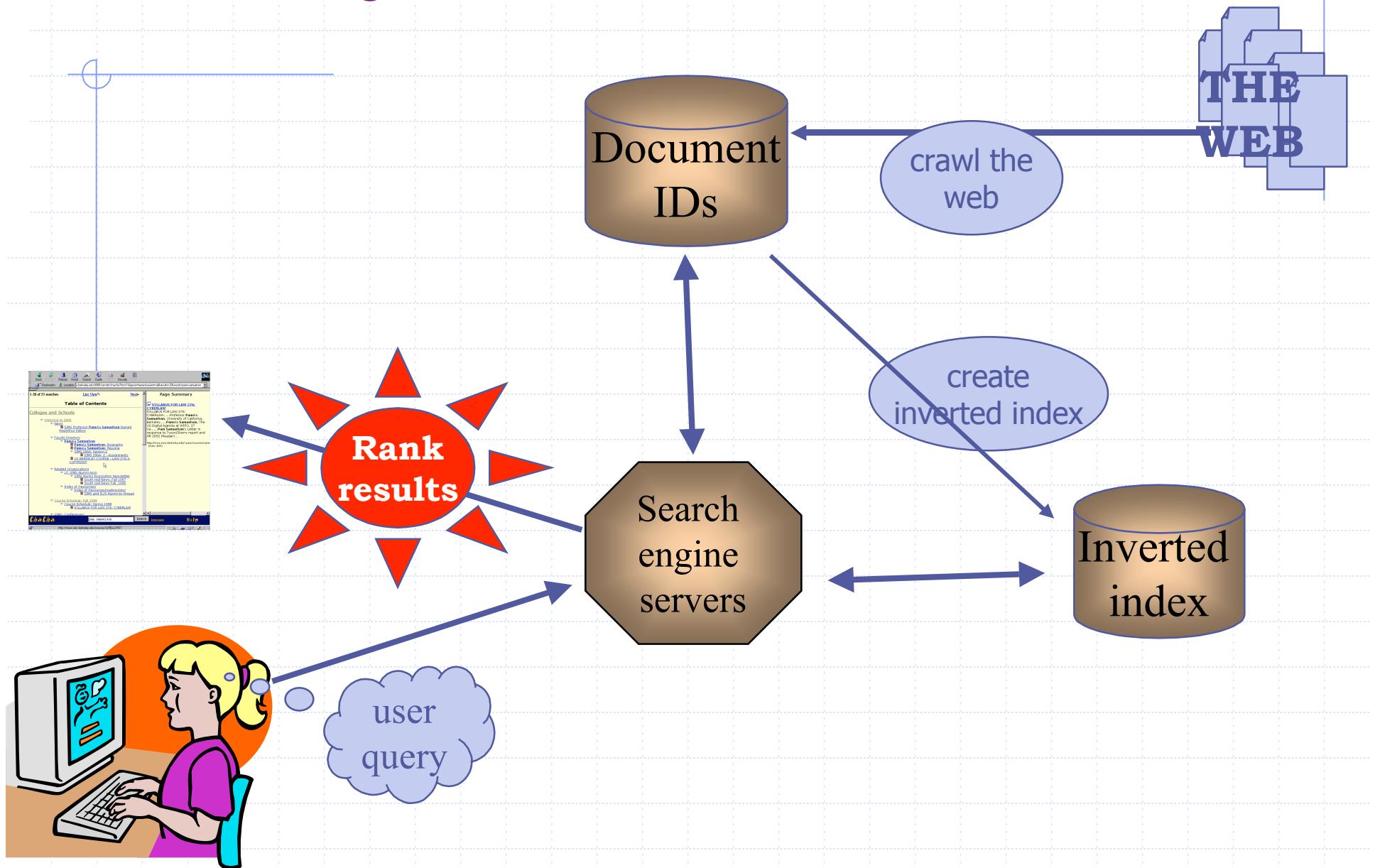
2nd Annual World Usability Day New England

The [Usability Professionals' Association](#) is organizing [World Usability Day](#) events worldwide with the goal of increasing awareness of the importance of usability and user-centered design. Last year's event lasted for 36 hours, with 115 events in 35 countries, on 6 continents all around the world.

[Dartmouth College](#) and [Landmark College](#) join together again to organize this conference with the theme, "universal usability to enhance learning, effectiveness, and understanding for people of all abilities." World Usability Day New England will offer two tracks: **Universal Usability in Teaching and Learning**, and **Usability of Products and Sy**



How Google (and the other search engines) Work



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"

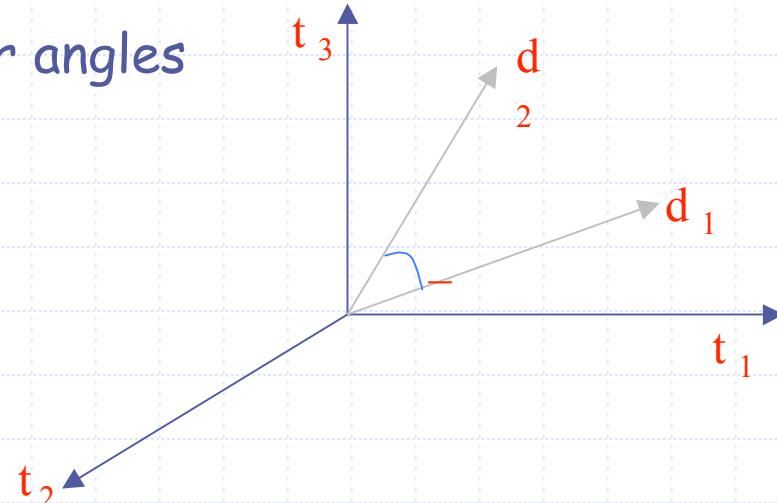
(no effort to "understand" the contents)

- ◆ Similarity is measured by vector angles

- ◆ Query Results are **ranked**

by sorting the angles
between query and documents

- ◆ **How To Spam?**



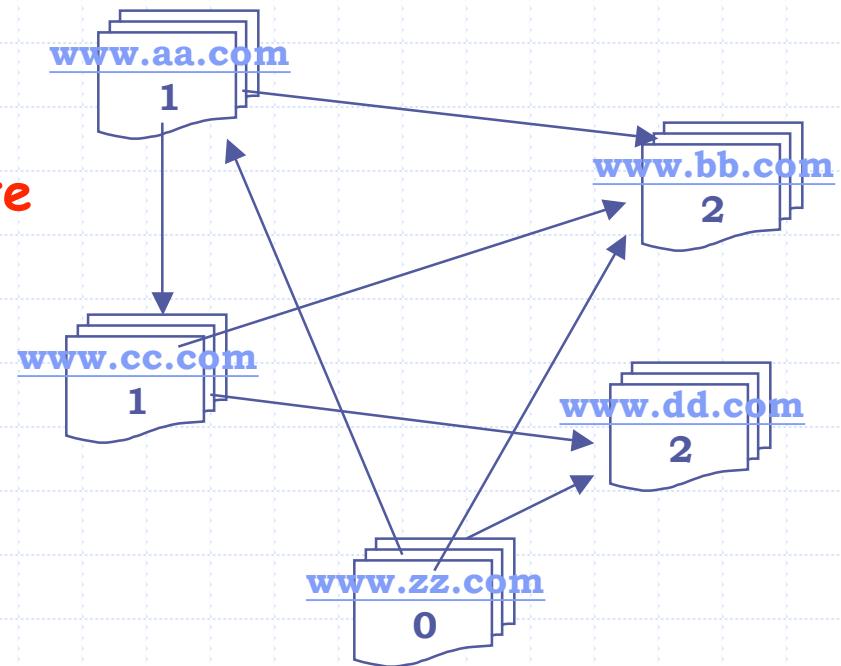
1st Generation: How to Spam

- ◆ "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 CASTA 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 CASTA 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 CASTA
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 CASTA BETTIE PAGE HEIDI KLUM PATRICIA FORD DAISY
FUENTES KELLY BROOK

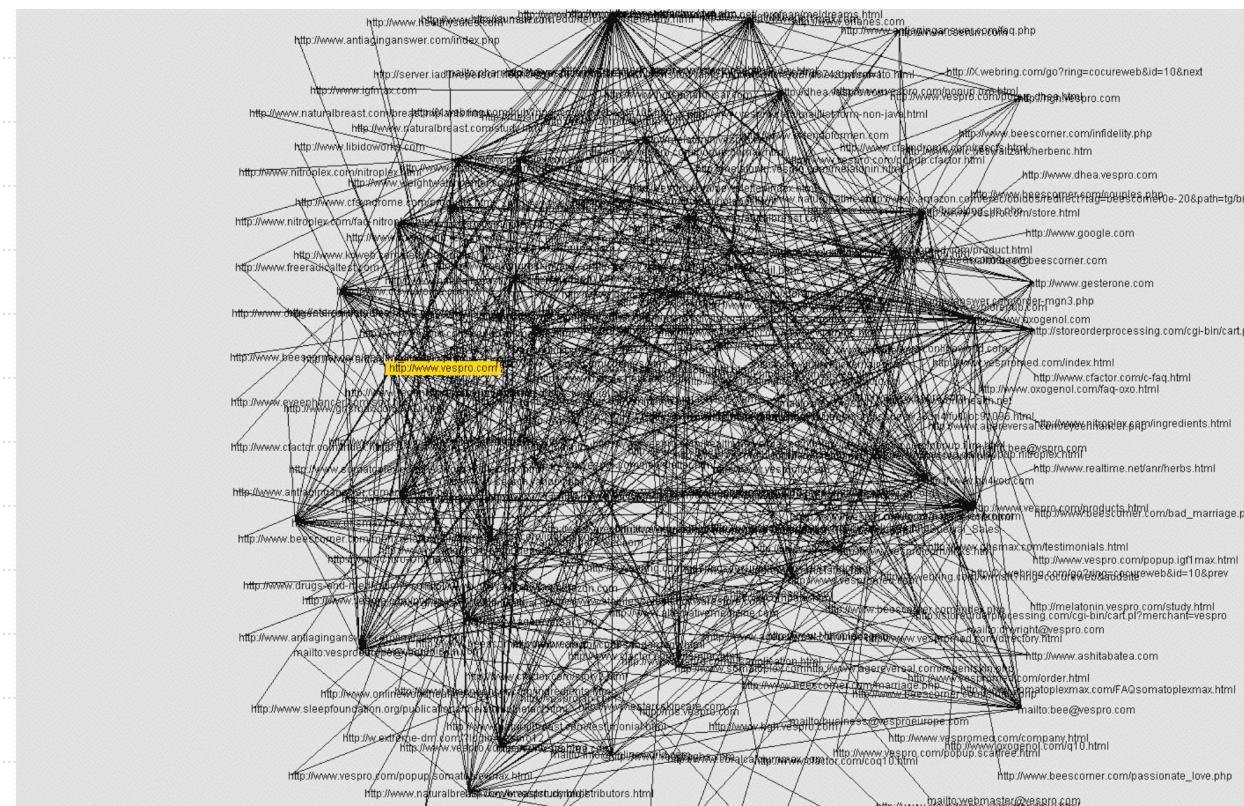
2nd Generation: Add Popularity

- ◆ 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

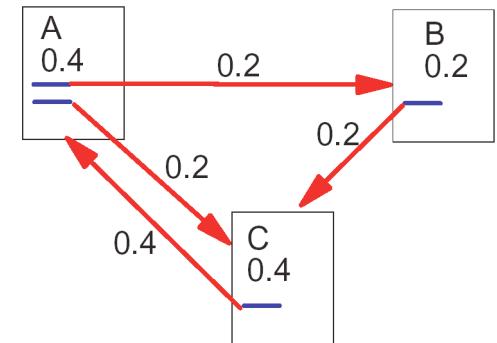
- ◆ The reputation "PageRank" of a page P_i =
the sum
of a fraction of the reputations
of all pages P_j that point to P_i

$$P_i = (1 - d) + d \cdot \sum_{j \rightarrow i} \frac{P_j}{C_j}$$

- ◆ 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?



Simplified PageRank Calculation

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><A HREF="<http://www.1st-Hgh.com>" TARGET="_blank">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.</P>

<!-- End Copying here -->

It should look like this:

[1st Hgh, Homeopathic Human Growth Hormone Spray](http://www.1st-Hgh.com) - 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

High PR Link Club

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!**

Boost your Google Pagerank with Reciprocal Linking & Web site Positioning Strategy - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media Mail PageRank 395 blocked AutoFill Opt

Address http://www.marketingeffects.com/linking_strategies.htm

Google

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 the [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 get on your favorite search engine (I recommend Google since it's

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

◆ In the Works

- Ranking based on
"the 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

◆ ??

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

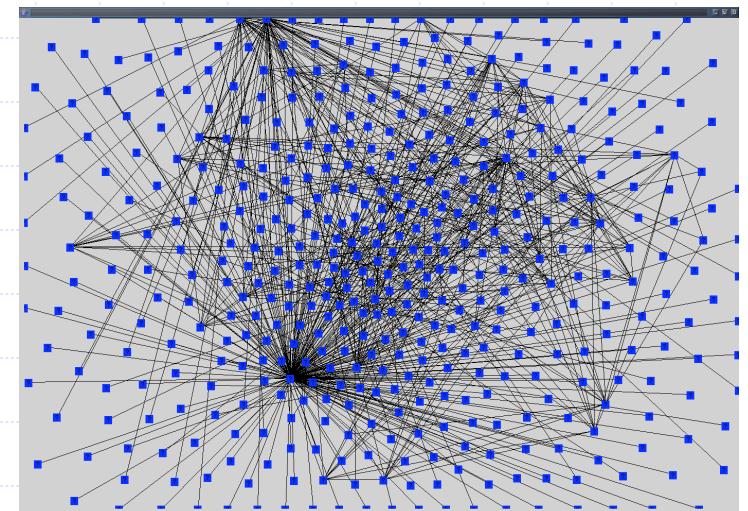
- Developed by the Institute for Propaganda Analysis 1938-42

◆ 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

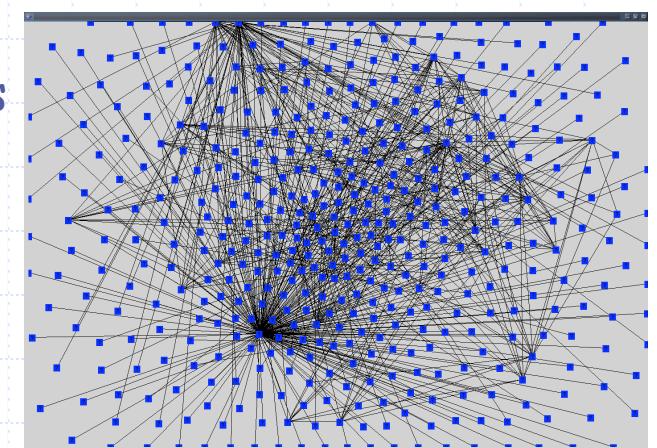
◆ Modify Node content & keep weights

◆ Insert Arcs b/w irrelevant nodes

◆ Modify Arcs

◆ Misdirect Arcs

◆ Modify Arcs & generate nodes



Web Spammers as Propagandists

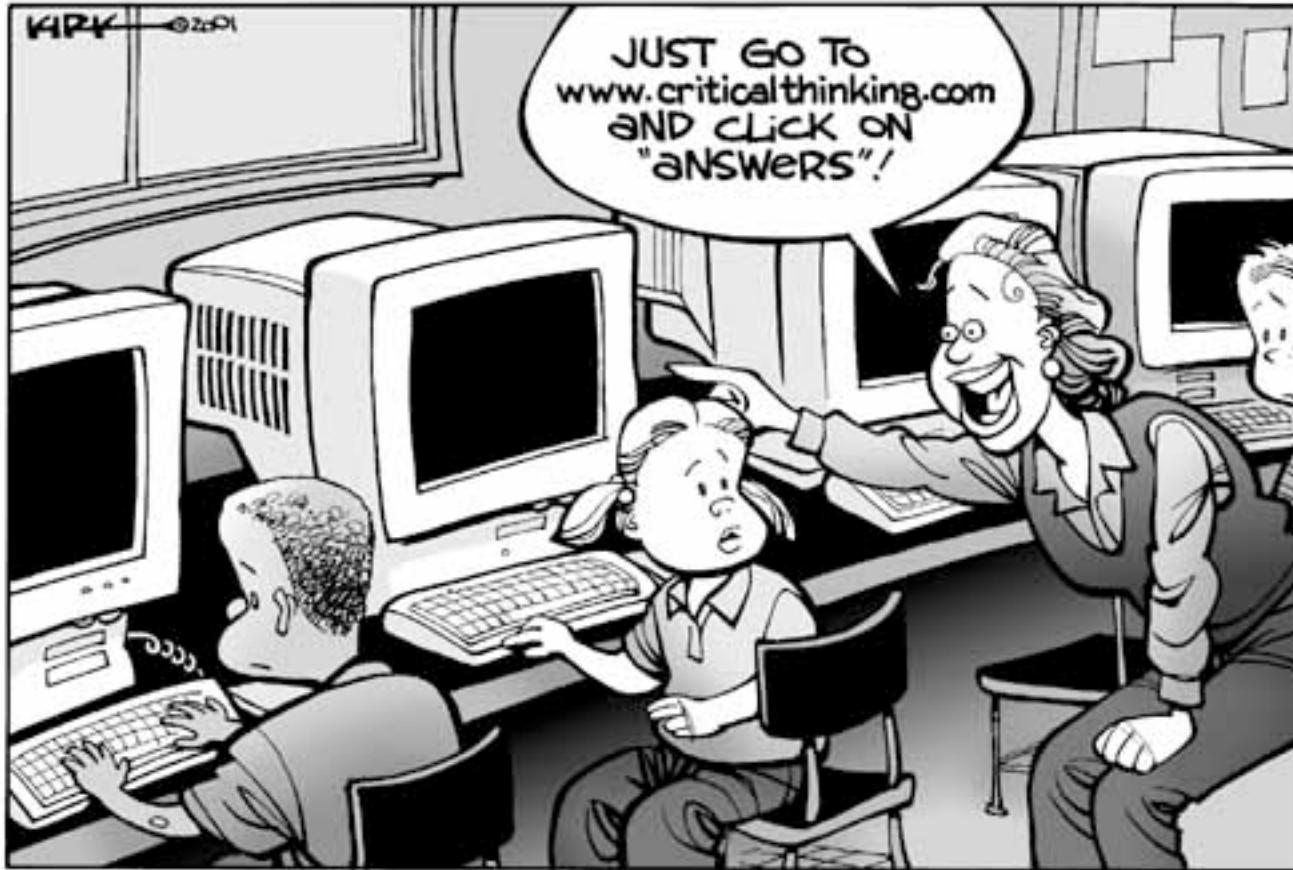
- ◆ Web Spammers can be seen as employing propagandistic techniques in order to modify the Web Graph
- ◆ **There is a pattern on how to spam!**

- | | | |
|-----------|--|-----------------|
| ◆ 1st Gen | ◆ "keyword stuffing" to increase content similarity | ◆ Word Games |
| ◆ 2nd Gen | ◆ Create "link farms" of heavily interconnected sites | ◆ Band wagon |
| ◆ 3rd Gen | ◆ Organize "mutual admiration societies" of irrelevant reputable sites | ◆ Testimonials |
| ◆ ? | ◆ Create Google-bombs | ◆ Card-stacking |

Cognitive Hacking and Cyber Trust

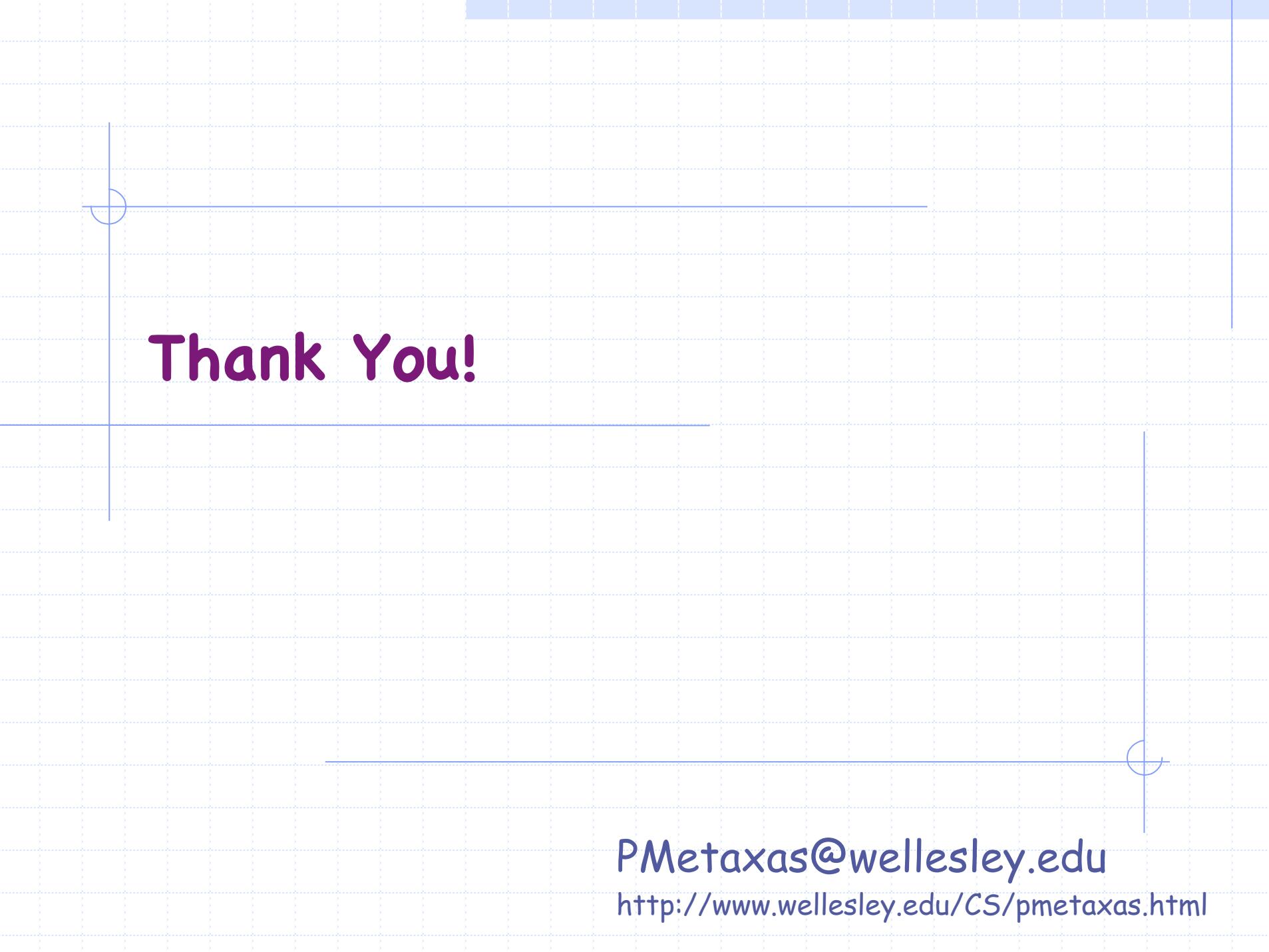
- ◆ Gaining Access or Breaking into a computer system for the purpose of modifying certain behaviors of a human user in a way that violates the integrity of the overall system
- ◆ Does not necessarily aim to fool a search engine
- ◆ Famous examples:
 - ◆ Pump & Dump stock schemes
 - ◆ The Emulex case
 - ◆ Word Games
 - ◆ Transfer

How (not) To Solve The Problem



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>