Agenda

- Brief overview of topics in ad-hoc web search (Emily).
- Discussion of PageRank and related work (Tomer).

Ad-hoc web search

- The web as a collection of documents is very different from the collections we’ve been working on so far.
  - enormous
  - always changing
  - multilingual
  - duplicated content
  - varying quality and reliability
  - unscrupulous content providers
- Manning et al: “Web is...unprecedented in scale, unprecedented in the almost complete lack of coordination in its creation, and unprecedented in the diversity of backgrounds and motives of its participants.”

The World Wide Web

- Anyone can publish anything, as long as it can be read and rendered by a browser.
- Millions of contributors with a wide variety of backgrounds and motivations.
- Massive amount of data of varying content and quality:
  - how to process and store the information?
  - how to quickly access the information the user needs?
  - how to assess the quality (reliability, accuracy, authoritativeness) of the information?

User experience in web search

- Users of web search are different from users of traditional IR.
  - don’t know or care about syntax of query language, crafting perfect queries
  - all they want are relevant, reliable results, returned quickly and presented in an easy-to-process way
- Three types of web queries:
  - informational
  - navigational
  - transactional
- As we’ve discussed, search engines try to determine the nature of the query and return results accordingly.

Identifying duplicates

- Search engine can save storage, time and return better results if it can identify duplicate or near-duplicate pages.
- One technique for this: *shingling.*
  - given $k > 0$ and a document *d,* the *k*-shingles of *d* is the set of all consecutive sequences of *k* terms in *d* (typical *k*=4)
- Example document: *a rose is a rose is a rose*
  - 4-shingles are:
    - a rose is a (count = 2)
    - rose is a rose (count = 2)
    - is a rose is (count = 1)
- Two documents are near-duplicates if they have nearly the same set of shingles. (Efficient computation described in detail in the textbook.)

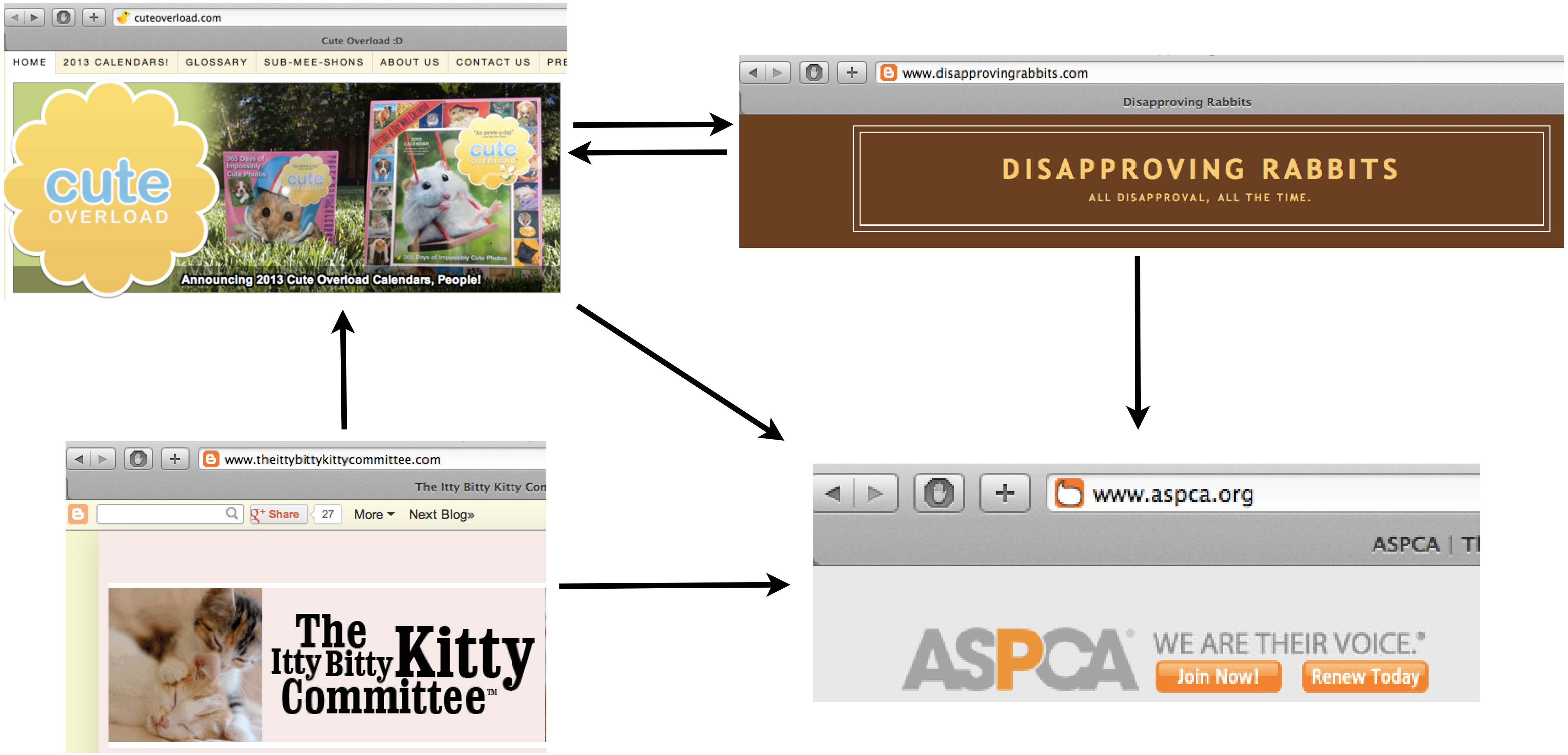
Early approaches to web search

- Hand-crafted taxonomies:
  - User searches through hierarchical tree of categories.
  - Pros: seems intuitive and easy for new users.
  - Cons: tons of human intervention.
- Traditional IR approaches:
  - Use indices and term-weighting-based ranking mechanisms.
  - Pros: proven in traditional IR tasks.
  - Cons: characteristics of the web make these approaches less feasible and appropriate.
- Big problem for both approaches: deciding which of the billions of pages are worthwhile, authoritative, trustworthy.

Web as a directed graph

Web pages are nodes.

Hyperlinks are directed edges.

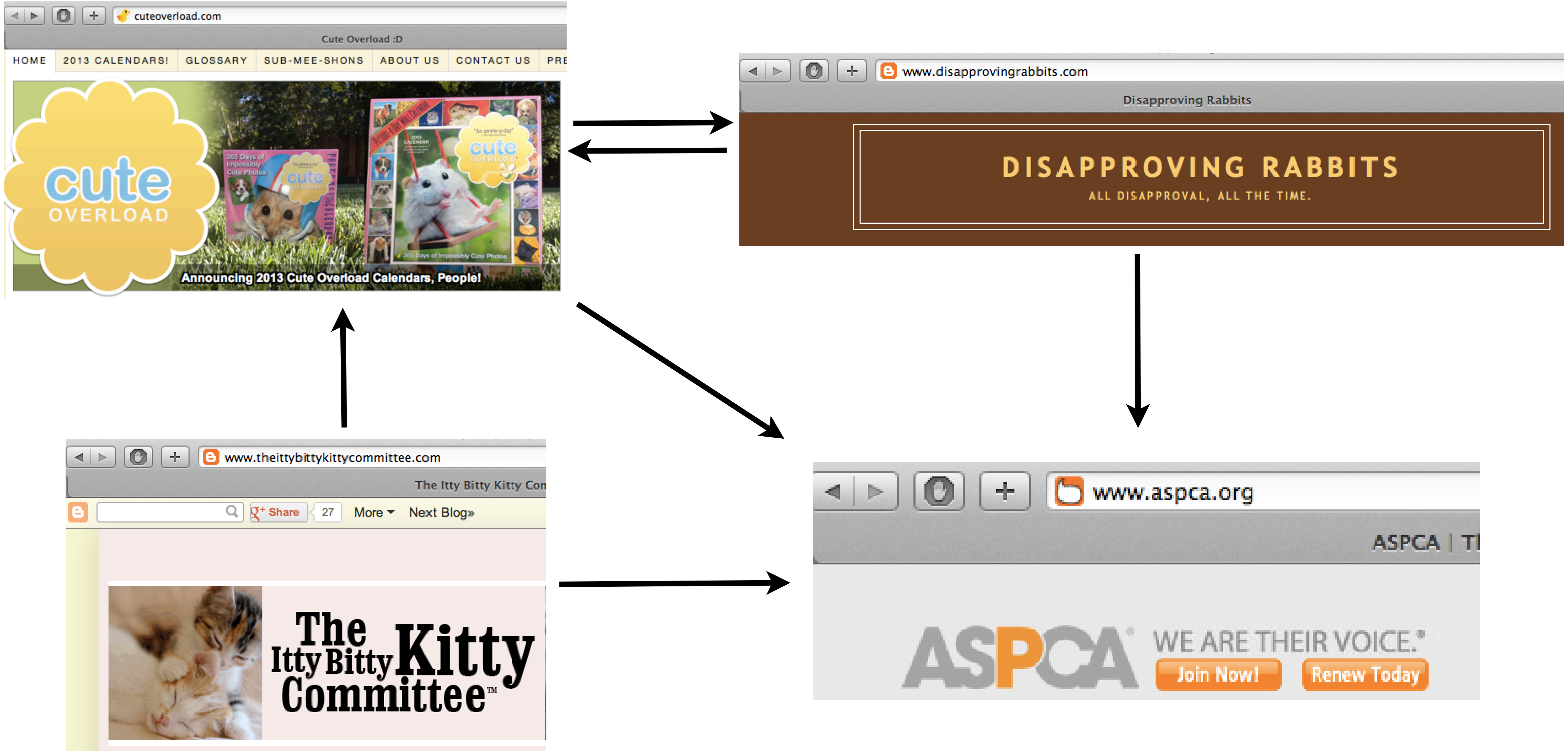


Web is not a strongly connected graph.

Web as a directed graph

in-degree = # incoming links

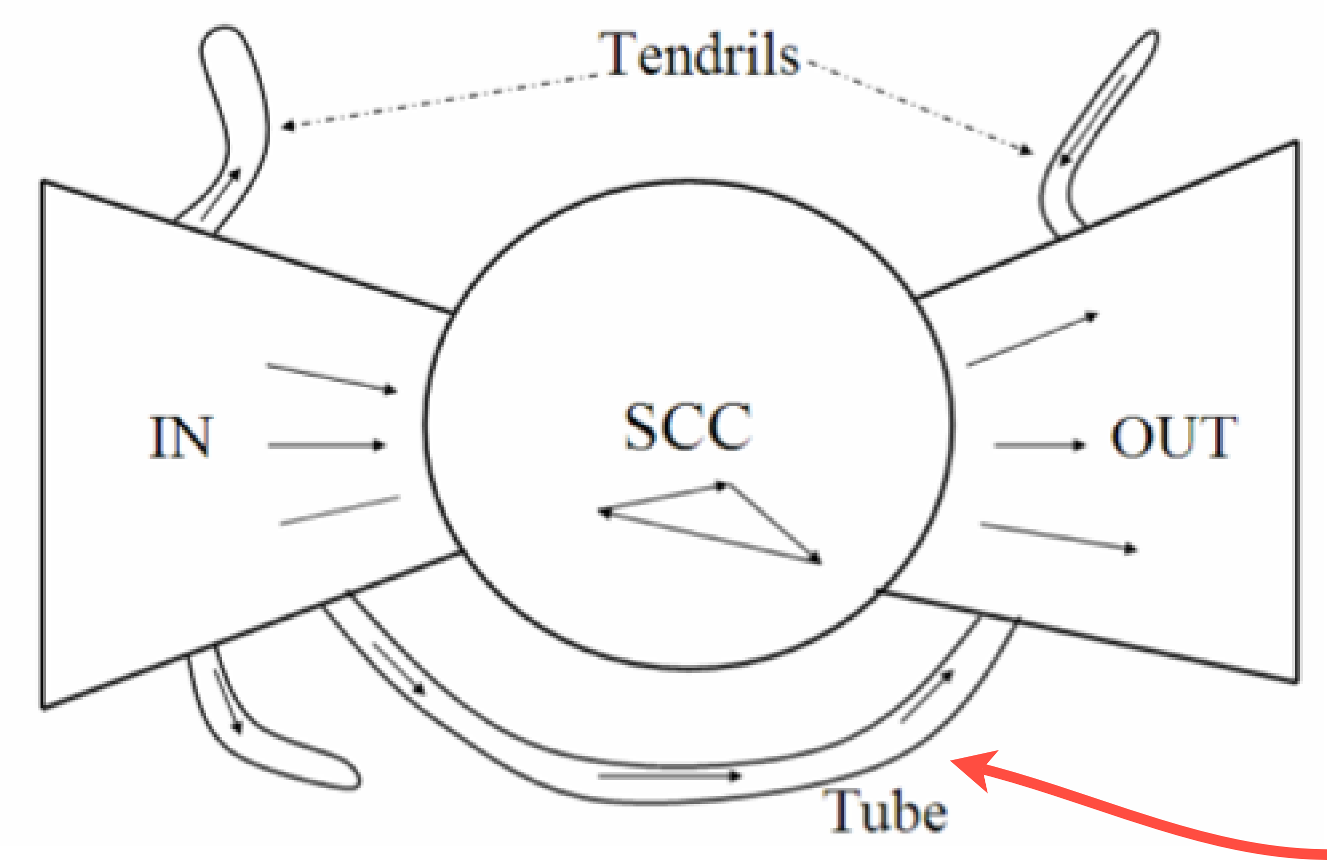
out-degree = # outgoing links

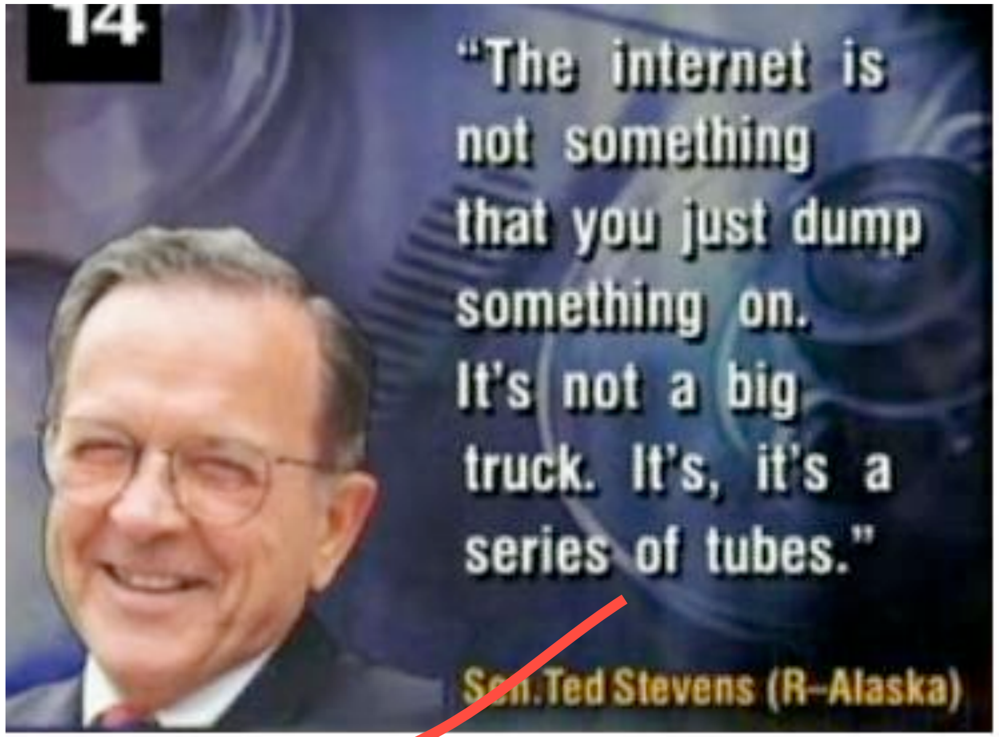


Disapproving Rabbits: in-degree=1, out-degree=2

Features of web graph

- Links are not randomly distributed.
- Distribution reported to be a power law:
  - number of pages with in-degree *i* is proportional to $1/i^{\alpha}$
  - α is reported to be 2.1
- Graph of web has a bow-tie shape:




14
"The internet is not something that you just dump something on. It's not a big truck. It's, it's a series of tubes."
Sen. Ted Stevens (R-Alaska)

Utility of web as directed graph

- Traverse the graph in order to find pages to index (a.k.a. web crawling or spidering).
- Use the information implicit in the graph to help rank retrieved pages not only for relevance but for authoritativeness:
  - Assume that pages with more incoming links are pages that are more trustworthy, authoritative, etc.
  - All the details coming up in Tomer's presentation...
- One challenge that will never entirely go away, however, is how to deal with spamdexing.

Spamdexing

- How do you get more people to come to your website?
  - Make a great website and a great product, or...
  - ...trick a search engine into assigning high rank to your site!
  - Traditional IR (e.g., for medical journals, legal cases) didn't need to worry about this problem.
- In early days of web search, one strategy was *keyword stuffing:*
  - To avoid user detection, include them in *metadata* or *hidden fields*, or have them be *literally invisible* (e.g., very tiny in a white font on a white background).
  - Kind of like AAAAppliance Repair in the Yellow Pages.

Advanced spamdexing

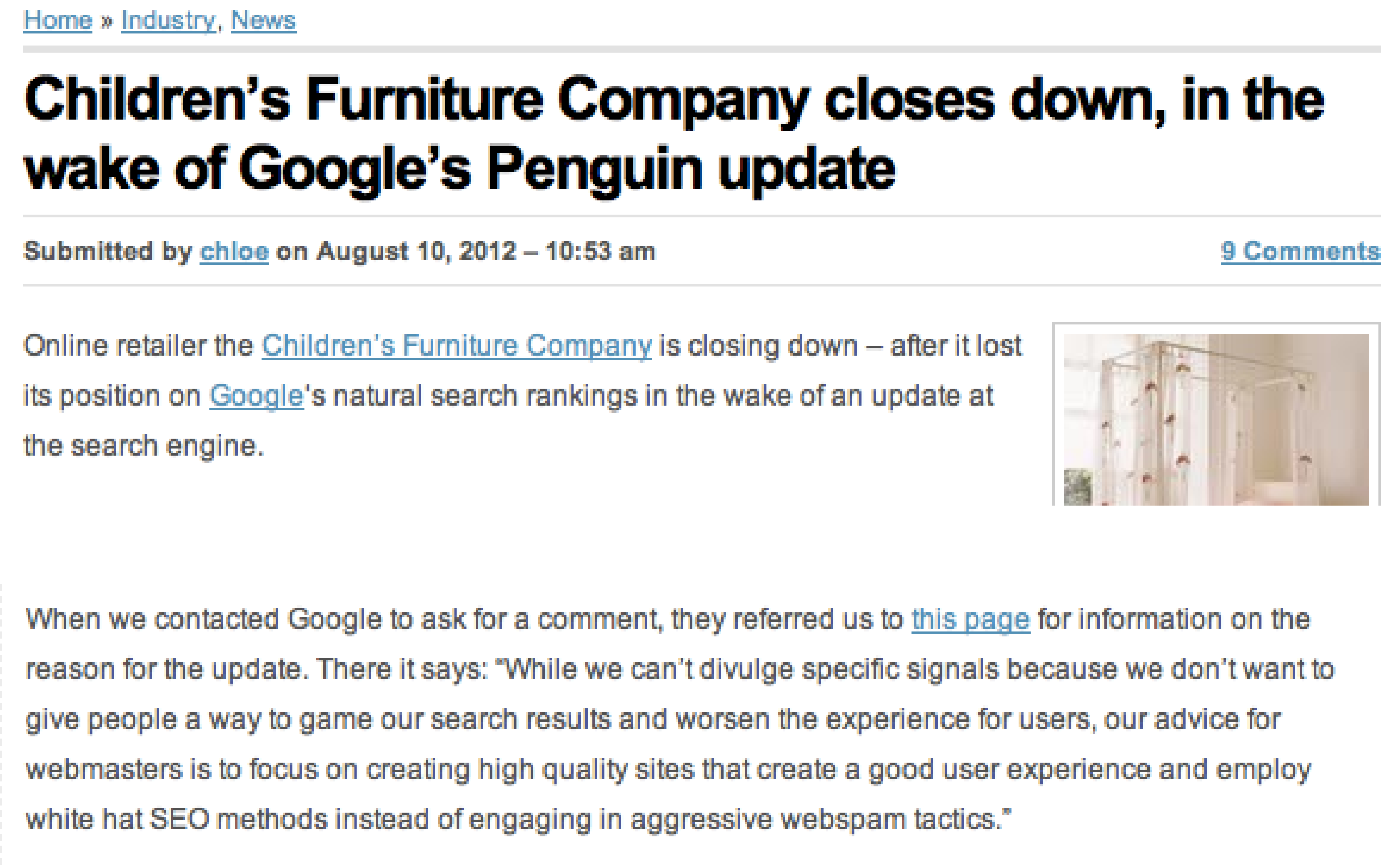
- *Cloaking*: returning one page for users and another for crawlers to avoid detection by crawlers.
- *Doorway pages:* show one page that looks legit to a web crawler, but all links go to commercial pages.
- *Scraping:* steal content from legit sites that will get ranked highly, then link or redirect to your commercial pages.
- *Link buying and exchange:* take advantage of link-based ranking algorithms by paying for links or making agreements with other spammers to link to your site if you link to theirs.
- *Comment spam:* put links to your site in the comments section of legitimate websites, again exploiting link-based ranking.

Search engine optimization (SEO)

- Techniques on previous pages are examples of *black hat SEO*.
  - These can get a site or domain excluded by Google et al.
- SEO can be a legitimate marketing strategy: *white hat SEO*.
  - research keywords and include them naturally in content
  - use valid, correct HTML, CSS
  - create site maps to make sure all pages are linked
- *Grey hat SEO*
  - marginally acceptable techniques
  - might not get banned, but could be penalized
  - today’s grey hat SEO may be tomorrow’s black hat SEO

Google vs. SEO

- A cat-and-mouse game between search engines and websites seeking more visibility.
- Can lead to legitimate sites getting banned or ranked down.



Home » Industry, News

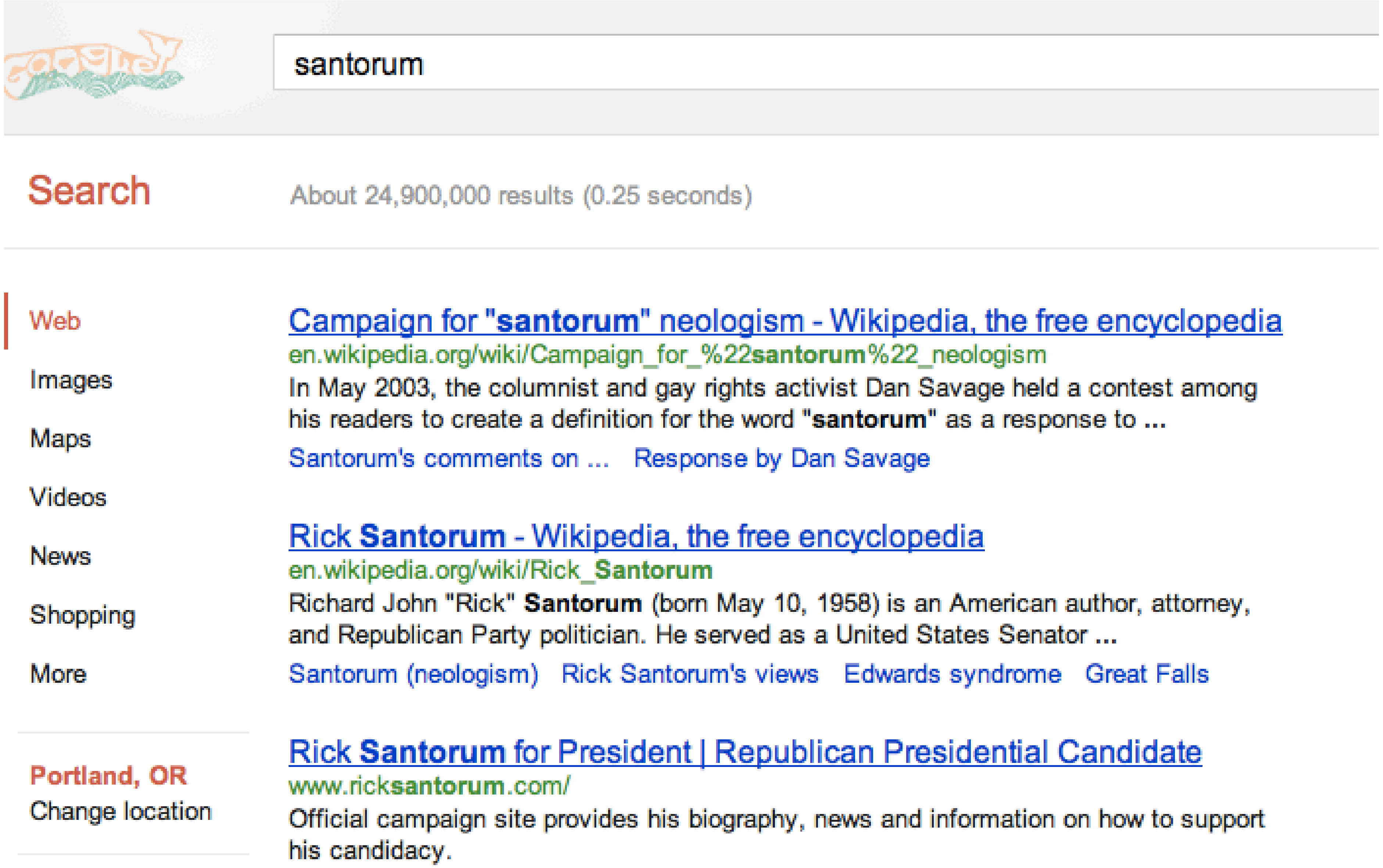
Children's Furniture Company closes down, in the wake of Google's Penguin update

Submitted by chloe on August 10, 2012 – 10:53 am **9 Comments**

Online retailer the Children's Furniture Company is closing down – after it lost its position on Google's natural search rankings in the wake of an update at the search engine.

When we contacted Google to ask for a comment, they referred us to this page for information on the reason for the update. There it says: "While we can't divulge specific signals because we don't want to give people a way to game our search results and worsen the experience for users, our advice for webmasters is to focus on creating high quality sites that create a good user experience and employ white hat SEO methods instead of engaging in aggressive webspam tactics."

SEO for fun: Google bombing



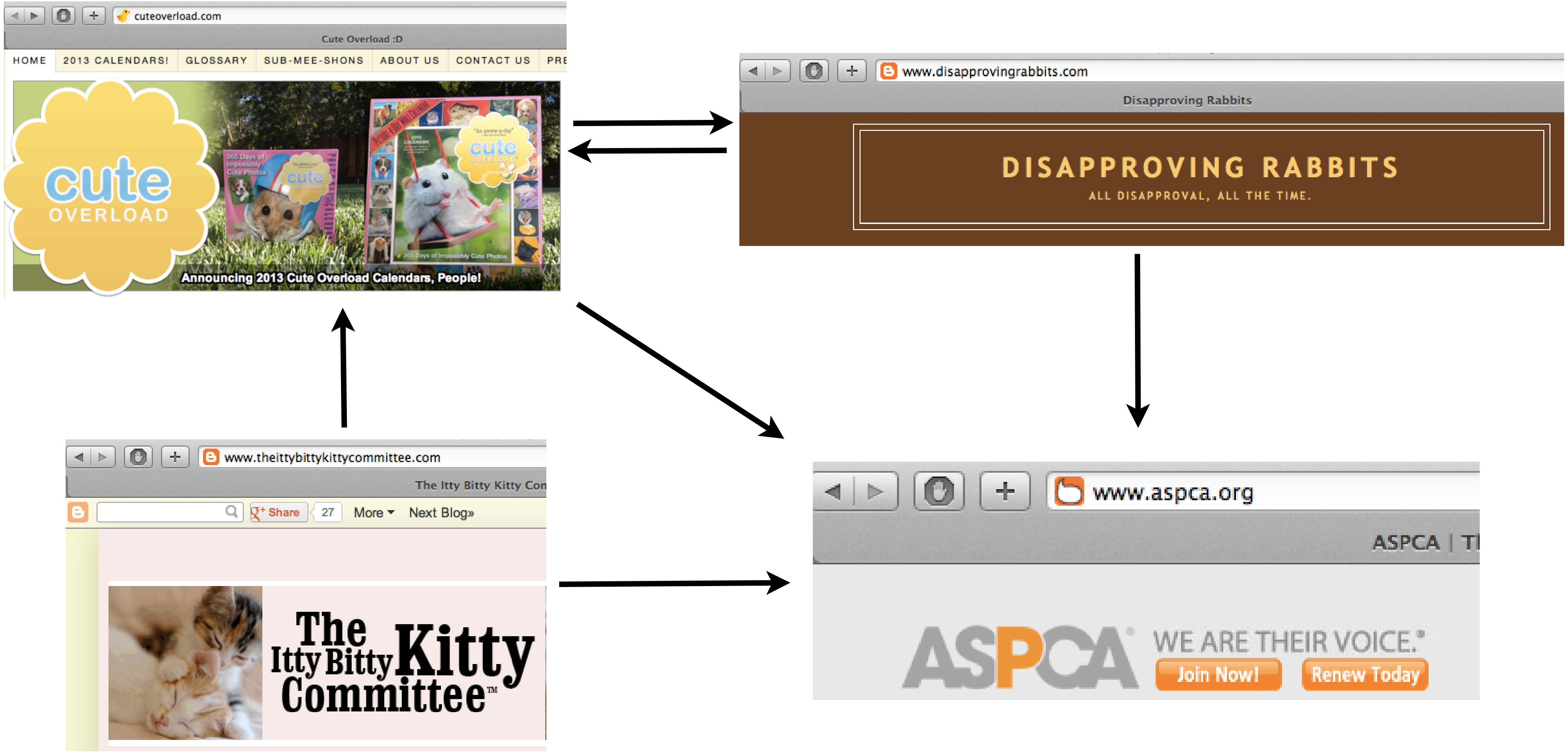
About 24,900,000 results (0.25 seconds)

Campaign for "**santorum**" neologism - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Campaign_for_%22**santorum**%22_neologism
In May 2003, the columnist and gay rights activist Dan Savage held a contest among his readers to create a definition for the word "**santorum**" as a response to ...
Santorum's comments on ... Response by Dan Savage

Rick **Santorum** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rick_**Santorum**
Richard John "Rick" **Santorum** (born May 10, 1958) is an American author, attorney, and Republican Party politician. He served as a United States Senator ...
Santorum (neologism) Rick Santorum's views Edwards syndrome Great Falls

Rick **Santorum** for President | Republican Presidential Candidate
www.rick**santorum**.com/
Official campaign site provides his biography, news and information on how to support his candidacy.

Web as a directed graph

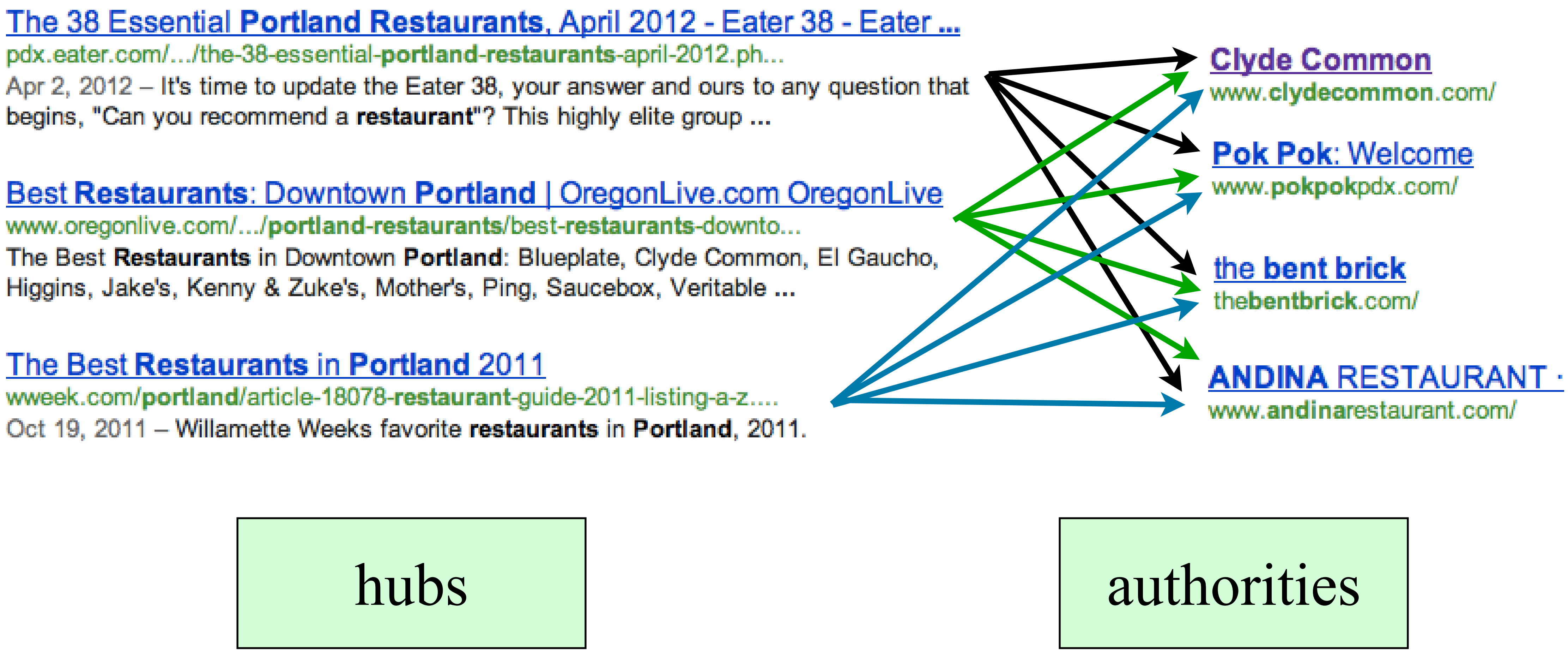


Link Analysis

- WWW is a graph, with pages for nodes and links for edges.
- A link between Page A and Page B is an implicit endorsement of Page B by the authors of Page A.
- Tomer will be talking about PageRank.
- A precursor to PageRank is Hyperlink-induced Topic Search (HITS).

Hyperlink-induced Topic Search (HITS)

- Pages with lots of outgoing links (to authorities) are *hubs*.
- Pages with lots of incoming links (from hubs) are *authorities*.
- Circular definition.



The 38 Essential Portland Restaurants, April 2012 - Eater 38 - Eater ...
pdx.eater.com/.../the-38-essential-portland-restaurants-april-2012.ph...
Apr 2, 2012 – It's time to update the Eater 38, your answer and ours to any question that begins, "Can you recommend a **restaurant**"? This highly elite group ...

Best Restaurants: Downtown Portland | OregonLive.com OregonLive
www.oregonlive.com/.../portland-restaurants/best-restaurants-downto...
The Best **Restaurants** in Downtown **Portland**: Blueplate, Clyde Common, El Gaucho, Higgins, Jake's, Kenny & Zuke's, Mother's, Ping, Saucebox, Veritable ...

The Best Restaurants in Portland 2011
wweek.com/portland/article-18078-restaurant-guide-2011-listing-a-z....
Oct 19, 2011 – Willamette Weeks favorite **restaurants** in **Portland**, 2011.

Clyde Common
www.clydecommon.com/

Pok Pok: Welcome
www.pokpokpdx.com/

the bent brick
thebentbrick.com/

ANDINA RESTAURANT ·
www.andinarestaurant.com/

Hyperlink-induced Topic Search (HITS)

- Start with a subset of the web: pages retrieved via standard IR text-based search from a query, plus the pages those pages link to and from.
- Calculate for every web page *v* in the subset **two** scores:
  - hub score: *h(v)*
  - authority score: *a(v)*

$$
\begin{aligned}
h(v) &\leftarrow \sum_{v \mapsto y} a(y) \\
a(v) &\leftarrow \sum_{y \mapsto v} h(y).
\end{aligned}
$$

Matrix math for HITS

- Let $\vec{h}$ and $\vec{a}$ be the vectors of hub and authority scores.
- Let A be the adjacency matrix, where $A_{ij} = 1$ if there is a link between page i to page j, and $A_{ij} = 0$ otherwise.

set h and a vectors uniformly

$$\begin{aligned} \vec{h} &\leftarrow A\vec{a} \\ \vec{a} &\leftarrow A^T\vec{h}, \end{aligned}$$

substitution

$$\begin{aligned} \vec{h} &\leftarrow AA^T\vec{h} \\ \vec{a} &\leftarrow A^TA\vec{a}. \end{aligned}$$

becomes an eigenvector problem

$$\begin{aligned} \vec{h} &= (1/\lambda_h)AA^T\vec{h} \\ \vec{a} &= (1/\lambda_a)A^TA\vec{a}. \end{aligned}$$

HITS Algorithm

```
$HITS(A, k)$
$a_0 \leftarrow [1 \quad 1 \quad \cdots \quad 1]^T$
$h_0 \leftarrow [1 \quad 1 \quad \cdots \quad 1]^T$
$for\ i = 1,2, \ldots, k$
    $a_i \leftarrow A^T h_{i-1}$
    $h_i \leftarrow A a_i$
    $a_i \leftarrow a_i / \|a_i\|$
    $h_i \leftarrow h_i / \|h_i\|$
$end$
$output\ a_k, h_k$
```

set h and a vectors uniformly

update a using last h

use a to update h

normalize

Ad-hoc web search

- Many other interesting topics related to web search, some of which will be covered later on.
- Moving on now to PageRank.