Quick Questions

- How much wood would a woodchuck chuck if a woodchuck would chuck wood?
  He’d chuck as much wood as a woodchuck could if a woodchuck could chuck wood.

- Can I get web access to our grades? Read and write permission?

- DecreaseGrade(double amountToDecreaseYourGrade)

- Where can I get good ostrich meat? http://www.ostrichesonline.com/meat/meatindex.html

- When will the dot.com stocks recover? 21 months

- Why wasn’t web invented earlier? Al Gore wasn’t around

- How many nodes in the web graph? 3 billion?

More Questions

- How does a web page request (email) know where to go? How does it know how to find me?

- How do I eliminate pop-up ads?

- How much bandwidth gets used each day on the net serving web pages?

- What is the number of pages on the web that have not had their content updated in the past year?

- What happens when we run out of IP addresses? How do we keep them from colliding?

- Why doesn’t most web routers verify the sender’s IP before forwarding? Does this make them vulnerable?

- Is Microsoft’s HailStorm idea of software as a service realistic?

- How do counters which count the visits (hits) to a page work? What are they used for?

- Why can’t certain ISP’s access some web pages?

- How do I access accounts/web pages/etc. which are restricted?

- How do counters which count the visits (hits) to a page work? What are they used for?

- How does packet routing work?

Search Questions

- How does web searching work?

- How do different search engines differ?

- How do they make money?

- How do they crawl and search such a big web?

More than 600 million OnLine

Search Questions

- How does web searching work?

- How do different search engines differ?

- How do they make money?

- How do they crawl and search such a big web?

What’s the best day to surf?

The chart shows the percentage of web traffic during the week.

- Sunday: 12.0%
- Monday: 14.5%
- Tuesday: 14.0%
- Wednesday: 13.5%
- Thursday: 15.0%
- Friday: 15.5%
- Saturday: 15.0%
What’s the most popular site?

<table>
<thead>
<tr>
<th>Site</th>
<th>AT WORK</th>
<th>AT HOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Yahoo!</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>AOL Time Warner</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Amazon</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>eBay</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Terra Lycos</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>About-Primedia</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>USA Network</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Viacom International</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>CNET Networks</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Excite Network</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Walt Disney Internet Group</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Landmark Communications</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>New York Times Company</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Gannett</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RealNetworks</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Verizon Communications</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>InfoSpace</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>The Gator Corporation</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>TPM Worldwide</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Adobe</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Tribune Interactive</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>CSE326</td>
<td>0.00003</td>
<td></td>
</tr>
</tbody>
</table>

Average Internet Usage Internationally

- Number of sessions per month: 18
- Number of unique domains visited: 48
- Page views per month: 797
- Page view per surfing session: 43
- Time spent per month: 9:49:53
- Time spent during surfing session: 0:32:04
- Duration of page view: 0:00:44

Surfing from 10,000 feet

- Type in web address, e.g., www.amazon.com
- DNS Lookup translates name into 32-bit IP address 207.171.181.16
- Transmission broken up into packets (TCP/IP)
- Packets travel via internet (routers direct packets at each hop) to destination

Where does it go?

```
traceroute www.u-tokyo.ac.jp
```

Tracing route to www.u-tokyo.ac.jp [133.11.128.254] over a maximum of 30 hops:

1  10 ms   10 ms   10 ms  regina-GE3-1.cac washington.edu [128.95.3.100]
2  10 ms   10 ms   10 ms  uwbr2-GE0-1.cac washington.edu [140.142.196.24]
3  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
4  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
5  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
6  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
7  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
8  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
9  10 ms   10 ms   10 ms  TRANSPAC-PWAVE psp.gigapop.net [198.32.170.60]
10 10 ms  10 ms  10 ms  www.u-tokyo.ac.jp [133.11.128.254]

Trace complete.

Web Searching...

What are we looking for?

- Excite
- Yahoo
- AltaVista
- AlltheWeb
- Lycos
- Google
- Metacrawler
- MSN

A case study...
PageRank

- Idea! Use link structure of web to determine a page’s value
- Interpret a link from page A to page B as a vote, by page A, for page B
- Weight A’s vote for B by the value of the voting page A (divided by the number of outgoing links on page A)

PageRank (cont.)

Assume page A has pages T₁, T₂,..., Tₙ which point to it (i.e., are citations). Let C(i) be the number of outgoing links from a page B. Then the PageRank of page A is given by:

\[ \text{PR}(A) = d \left( \frac{\text{PR}(T₁)}{C(T₁)} + \frac{\text{PR}(T₂)}{C(T₂)} + \ldots + \frac{\text{PR}(Tₙ)}{C(Tₙ)} \right) \]

If we view the web as a graph, where each node is a web page and each edge is a link, then the PageRank corresponds to the principal eigenvector of the adjacency matrix and d corresponds to the principal eigenvalue.

PageRank Example

PageRank Stability

Page Info and Anchor Text

<table>
<thead>
<tr>
<th>Font</th>
<th>Capitalization</th>
<th>Count</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Associate the text of a link with the page that the link points to!

Architecture
An alternative model...
Hubs and Authorities

- Use text based search to generate list of candidate pages
- Calculate hub score and authority score for all pages in the candidate set
- Return set of pages with highest authority score