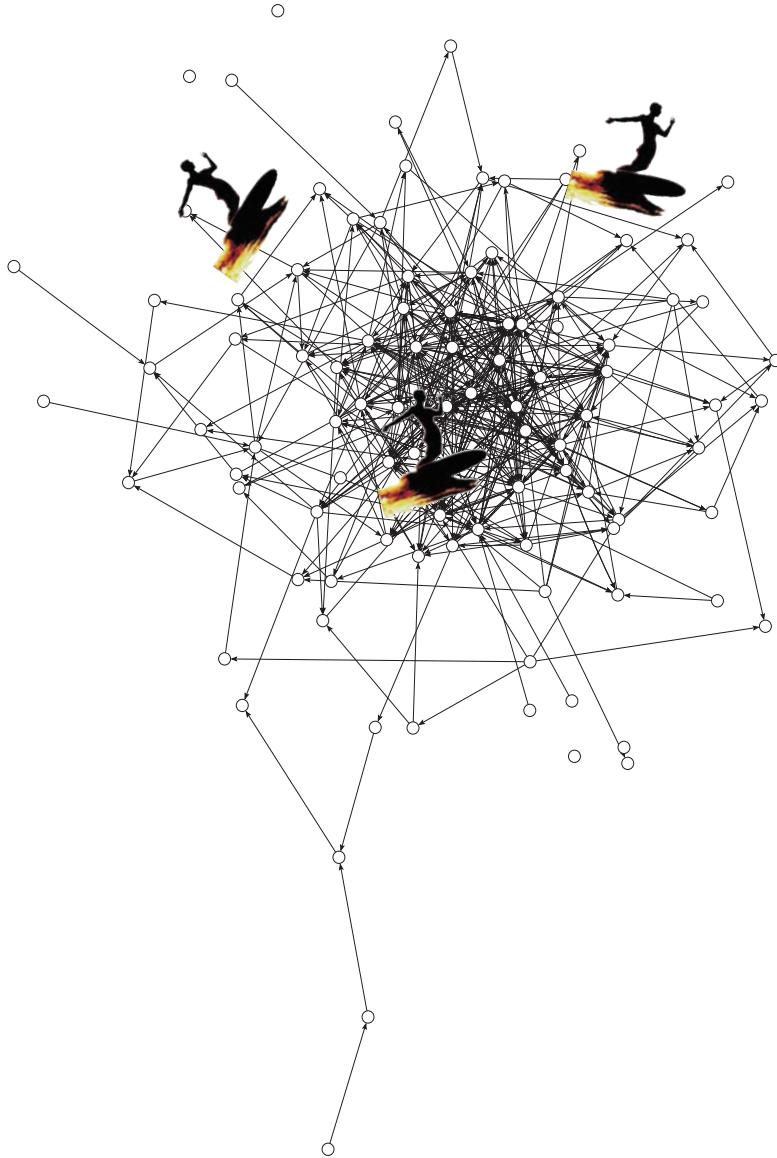


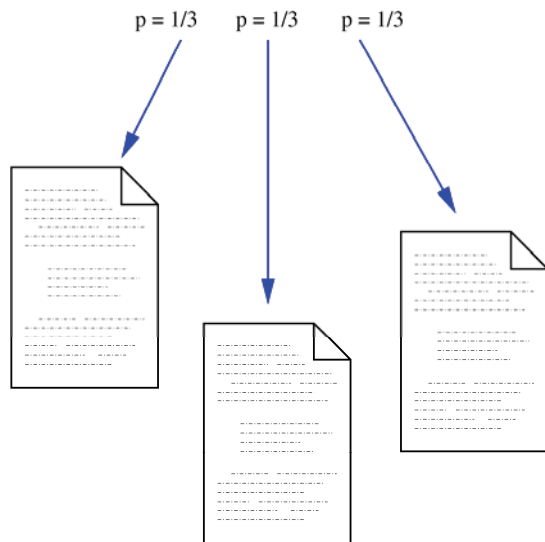
(I) Problem Description



PageRank assumes a ***RANDOM SURFER*** model of Web traffic.

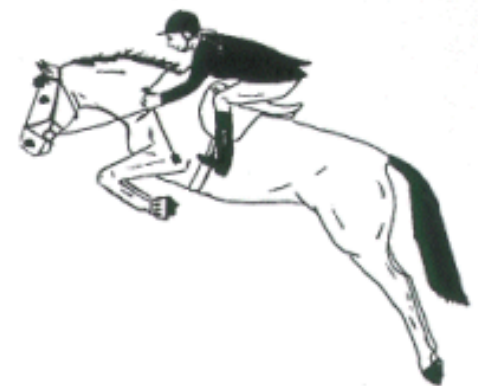
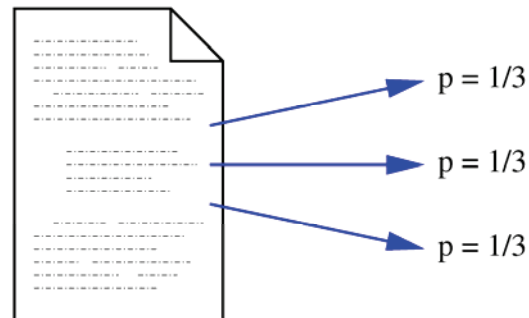
The PageRank of a page is the ***PROBABILITY*** of finding a random surfer at that page.

Standard PageRank is based on *UNIFORM DISTRIBUTIONS*:



Every page is an equally likely starting point.

Every outgoing link is equally likely to be followed.



Every page is equally likely to be jumped from.



Do people *REALLY* behave that way?

How often do you click on the link above?

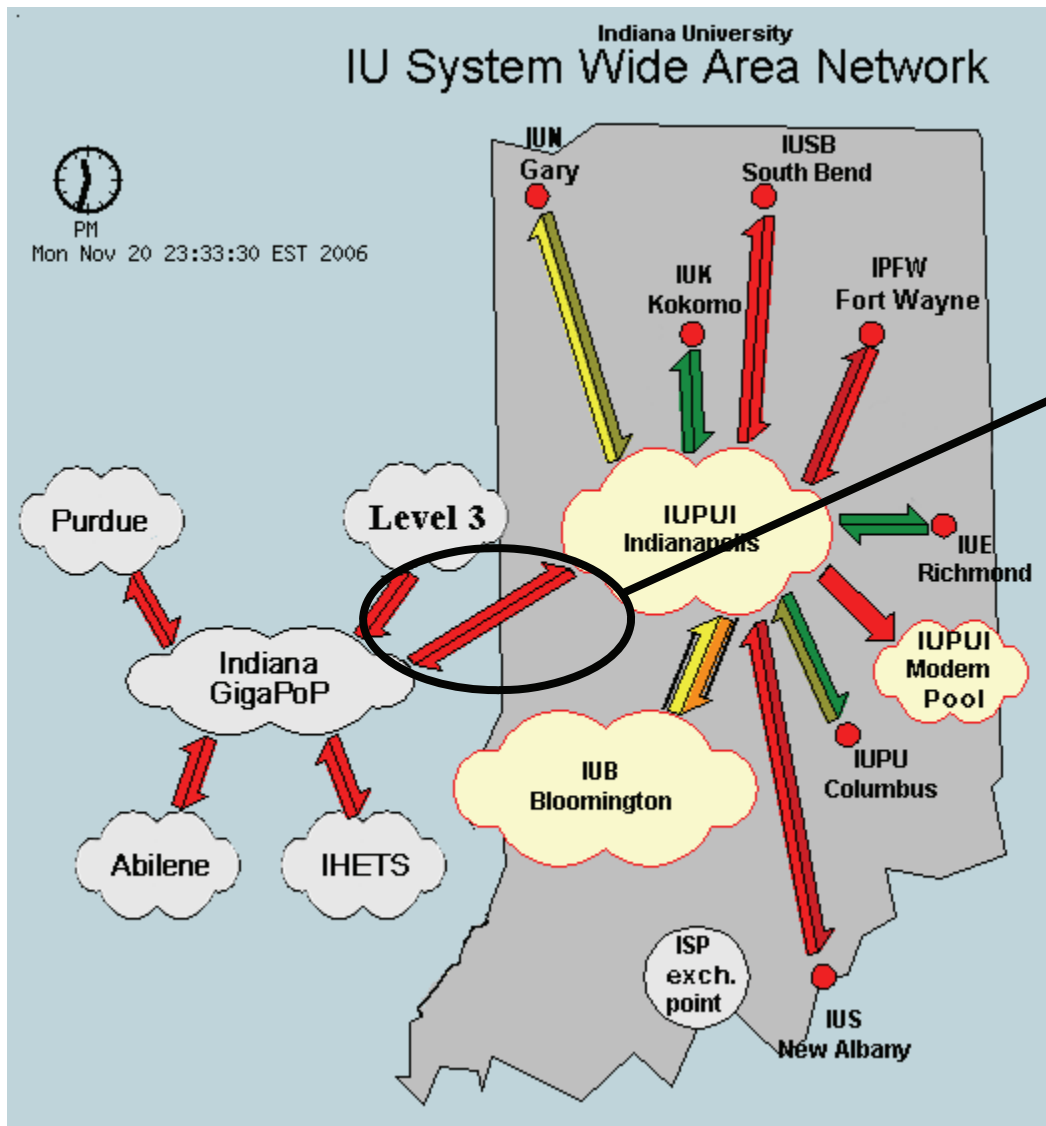
Not all links are created equal!



We can improve the model by collecting Web usage data from *REAL PEOPLE*.



(II) Data Source



Over 100,000 people at Indiana University surf the Web through the ***INDIANA GIGAPOP***.

Through arrangement with the IU Security Office, we can study Web traffic if we ***ANONYMIZE**** the data.

** No information of any kind about client identity is retained.*

We try to examine every *HTTP GET* packet going to TCP port 80.

IP source: `156.56.103.1` ← *Is the client inside the IU network?*
IP destination: `192.168.55.10`

Source port: `9421`
Destination port: `80`

GET `/index.html` HTTP/1.1 ← *What URL is being requested?*

Agent: `fredbot` ← *Is this a browser or a bot?*

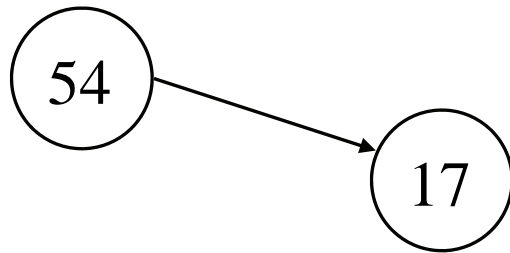
Referer: `http://www.grumpy-puppy.com/links.html`

Host: `www.happy-kitty.com` ← *What page contained this link?*

← *Which virtual host is being queried?*

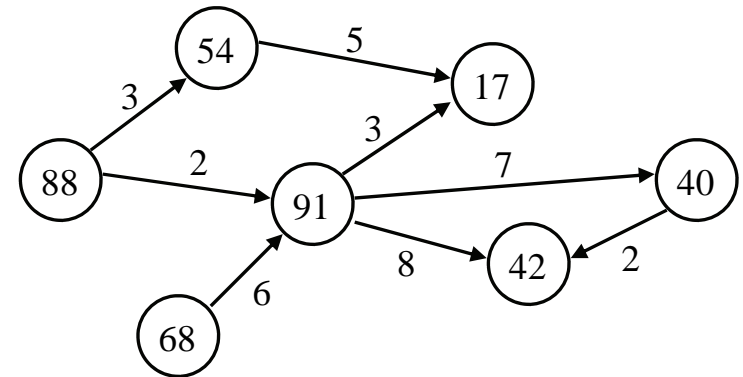
Each request represents the use of an actual link.

We build an *EDGE LIST* and *HOST INDEX* from the requests...



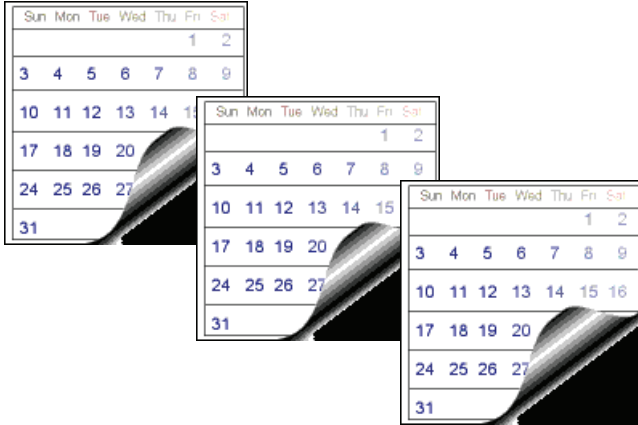
17: fruit-salad.food.com
54: www.google.com

...and then *AGGREGATE* all the edges over a period of time...



...yielding a *WEIGHTED SUBSET* of the actual host graph.

(III) Preliminary Results



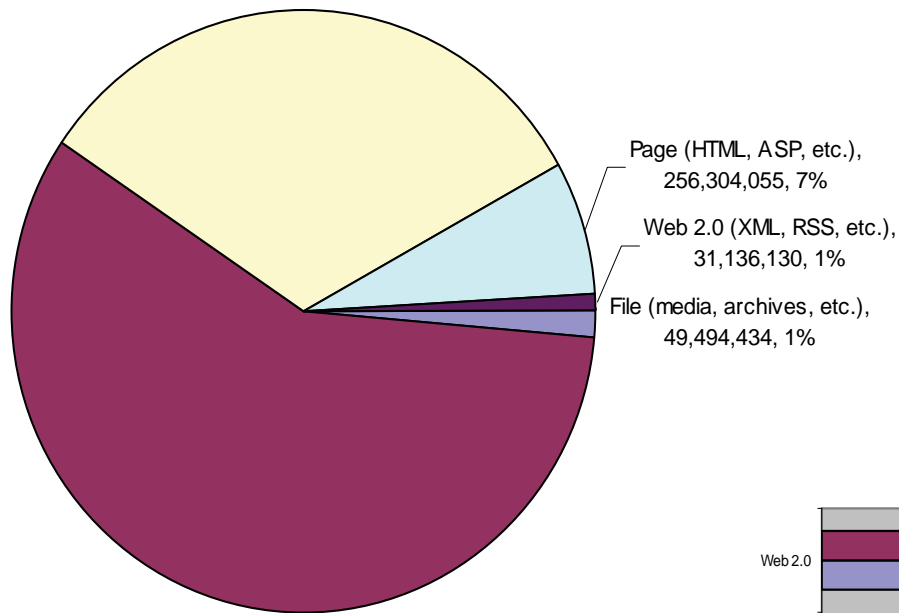
We began capturing click data with full target URLs on ***SEPTEMBER 26, 2006.***

Between then and November 17, the data has grown to include:

- . ***3,520,000,000 clicks***
- . ***3,510,000 distinct servers***
- . ***over 350 GB of disk storage***

Requests by type of URL

Other / Unknown,
1,138,443,208, 32%

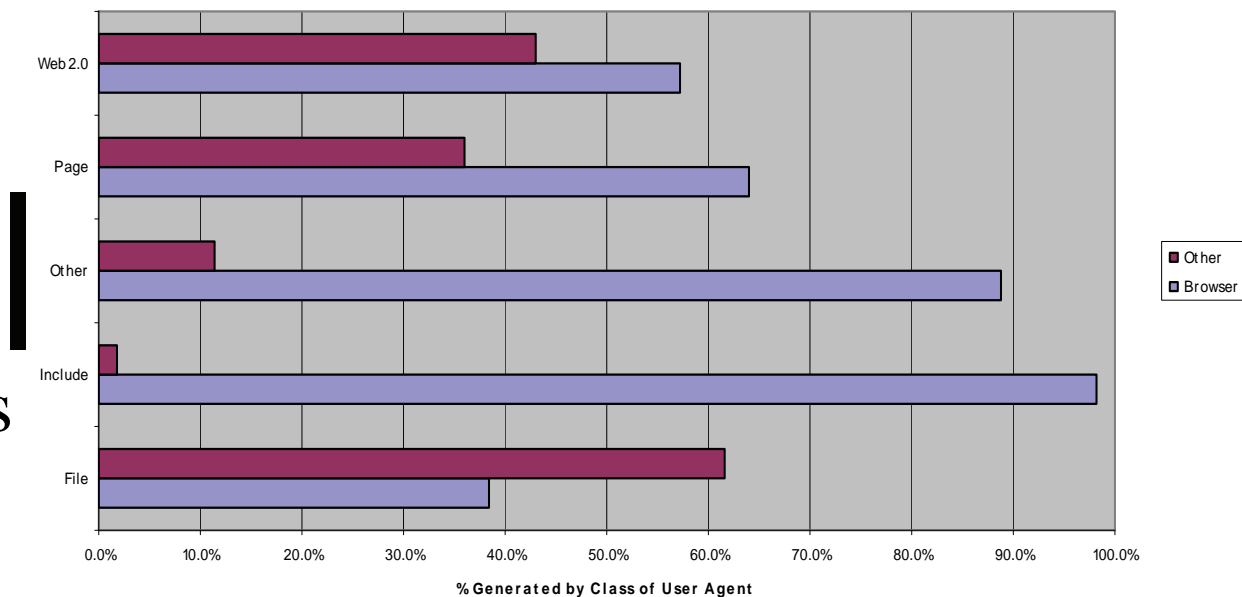


Include (images, CSS, etc.),
2,048,305,421, 59%

Verified page requests are only **7%** of the total.

Over **10** additional requests are made for every page retrieved.

Breakdown by User Agent



Browsers and other agents (i.e., crawlers) have very different profiles.

Many data properties vary strongly by *TIME OF DAY*:



Highest during the day

- Proportion of interior nodes
- Proportion of self-loops
- Proportion of “page” and “include” URLs
- Browsers more focused on files
- Browser traffic

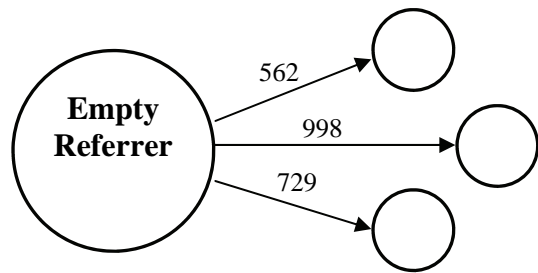


Highest late at night

- Mean node degree
- Mean node strength
- Mean edge weight
- Proportion of “file” and “other” URLs
- Browsers more focused on pages
- Crawler traffic

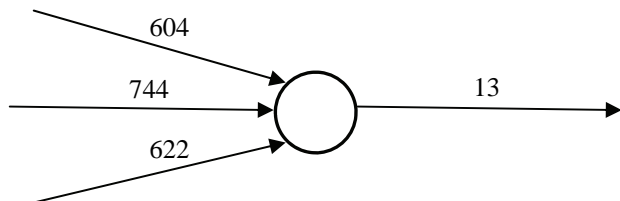
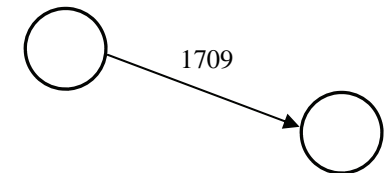
(IV) Modeling

Using our data, we can derive *NEW DISTRIBUTIONS* for PageRank:



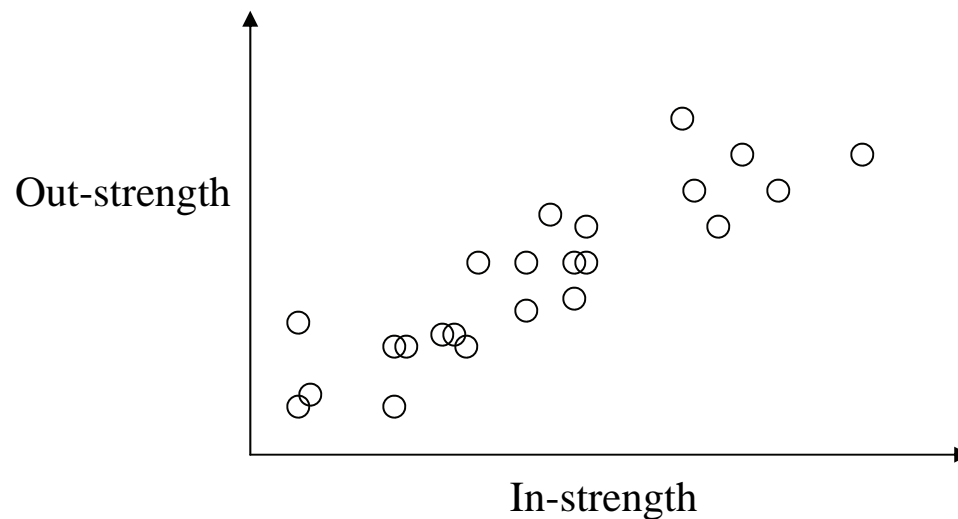
Pages with a high in-strength from the *EMPTY REFERRER* are more likely to be *STARTING PAGES*.

Edges with a large weight represent *POPULAR LINKS* that can be followed preferentially.



Pages with *HIGH IN-STRENGTH* and *LOW OUT-STRENGTH* are more likely to be *JUMP POINTS*.

We can also evaluate the effectiveness of HITS by examining the *ASSOCIATIVITY* of the traffic graph:



Does HITS identify pages that *BEHAVE* as hubs and authorities in actual Web traffic?

We can also conduct the first attempt at modeling a major subset of Web traffic:



Do “adult” sites behave like other sites, or must they be modeled in a different way?

Can the locations of adult sites in the link graph be used in developing more effective content filters?

M. Meiss,
F. Menczer,
S. Fortunato,
A. Vespignani,
A. Flammini

Improving the Random-Surfer Model with Anonymized Traffic Data