Web traffic: analysis of navigation data and modeling at single user level.

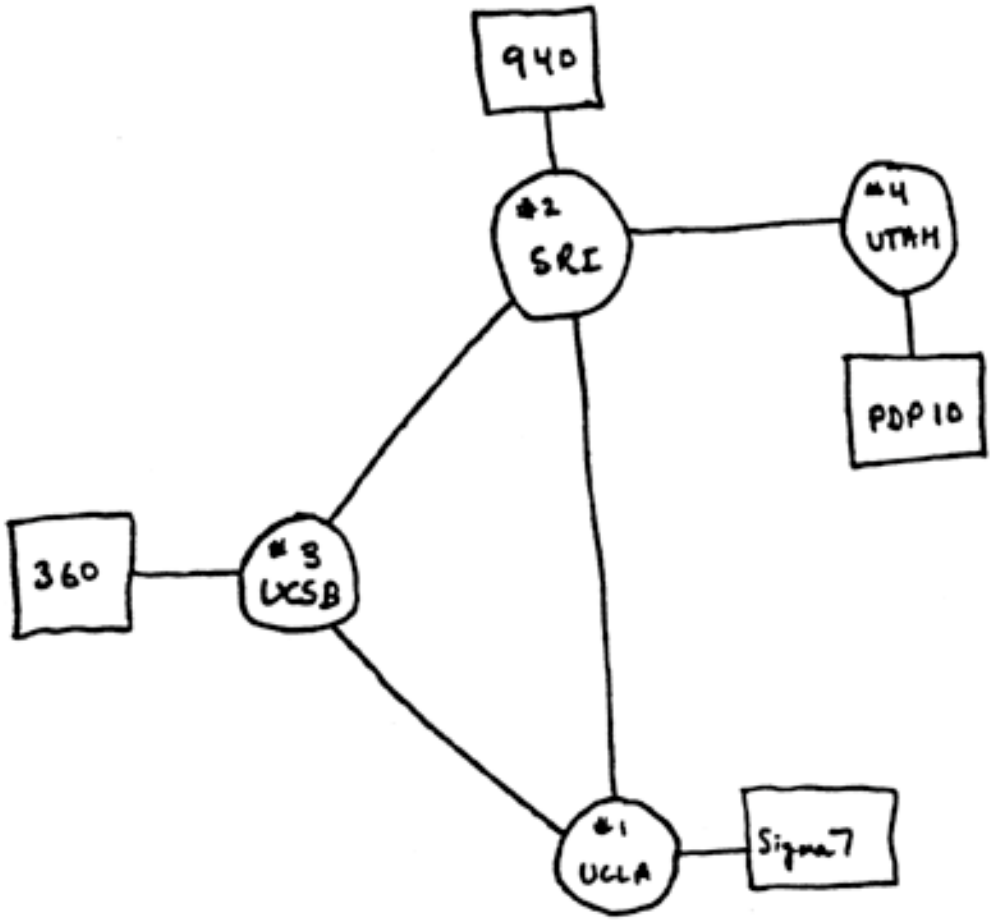
José Javier Ramasco

Outline

- Internet and the Web
- Navigation traces
- Data analysis at an aggregate level
- Individual-level data: navigation trees
- Models of Web navigation

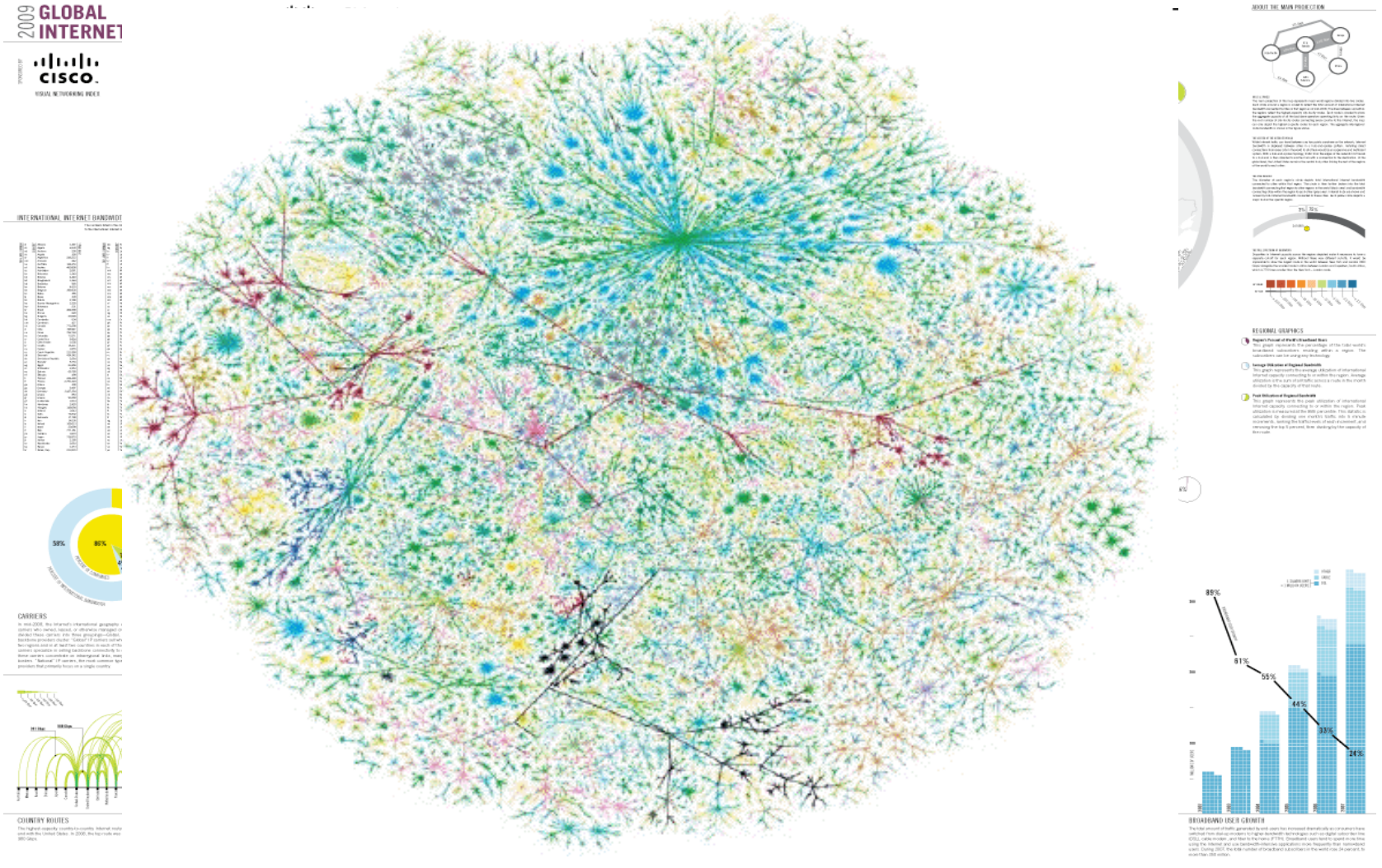
Internet and the WWW (Web)

The Internet in 1969 (ARPA)



Internet and the WWW (Web)

The Internet today



Hulu - Most Popular Episodes / Episodes / All Channels - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.hulu.com/popular/episodes/today

witcom d450

Most Visited | Share on Tumblr | iGoogle | JDK 6 Documentati... | Google Reader | Recently saved pag... | tp://squunkin.hobby-... | Personal - Old Natio...

Stumble! | I like it! | Send to | Channels: | All | Favorites | Friends | Tools

Scribd documents – 745... | Welcome to Google Docs | Welcome to Google Docs | http://kaylee...or/login.php | http://kaylee...or/login.php | Google Reader (1000+) | Google Image Result for ... | Hulu - Most Popular _

hulu™ TV Movies

Login | Forgot Password? | Sign Up

Channels ▾ | Popular Episodes | Popular Clips | Popular Shows | Browse | search

Most Popular Episodes

All Time | Today | This Week | This Month

1 of 589

Filter results by:

Browse

- Most Popular
- Recently Added
- Highest Rated
- Release Date

Programming Type

- All
- All TV
- All Movies
- TV Clips
- TV Full Episodes
- Games
- Movie Clips
- Movie Trailers
- Feature Films

Channel

- Action and Adventure
- Animation and Cartoons
- Comedy
- Drama
- Family
- Food and Leisure
- Home and Garden
- Horror and Suspense

CC Family Guy: Stew-Roids

Season 7 : Ep. 13 (21:54)

+ queue

More: Family Guy
Channel: Comedy

CC The Office: Casual Friday

Season 5 : Ep. 24 (21:47)

+ queue

More: The Office
Channel: Comedy

CC Dollhouse: Briar Rose

Season 1 : Ep. 11 (49:20)

+ queue

More: Dollhouse
Channel: Science Fiction

CC Family Guy: 420

Season 7 : Ep. 12 (21:53)

+ queue

More: Family Guy
Channel: Comedy

CC 30 Rock: The Natural Order

Season 3 : Ep. 20 (21:25)

+ queue

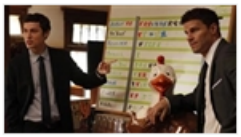
More: 30 Rock
Channel: Comedy

CC The Simpsons: Father Knows Worst

Season 20 : Ep. 18 (21:40)

+ queue

More: The Simpsons
Channel: Comedy



Bones: The Beaver In The Otter

Season 4 : Ep. 22 (43:36)

+ queue

More: Bones
Channel: Drama

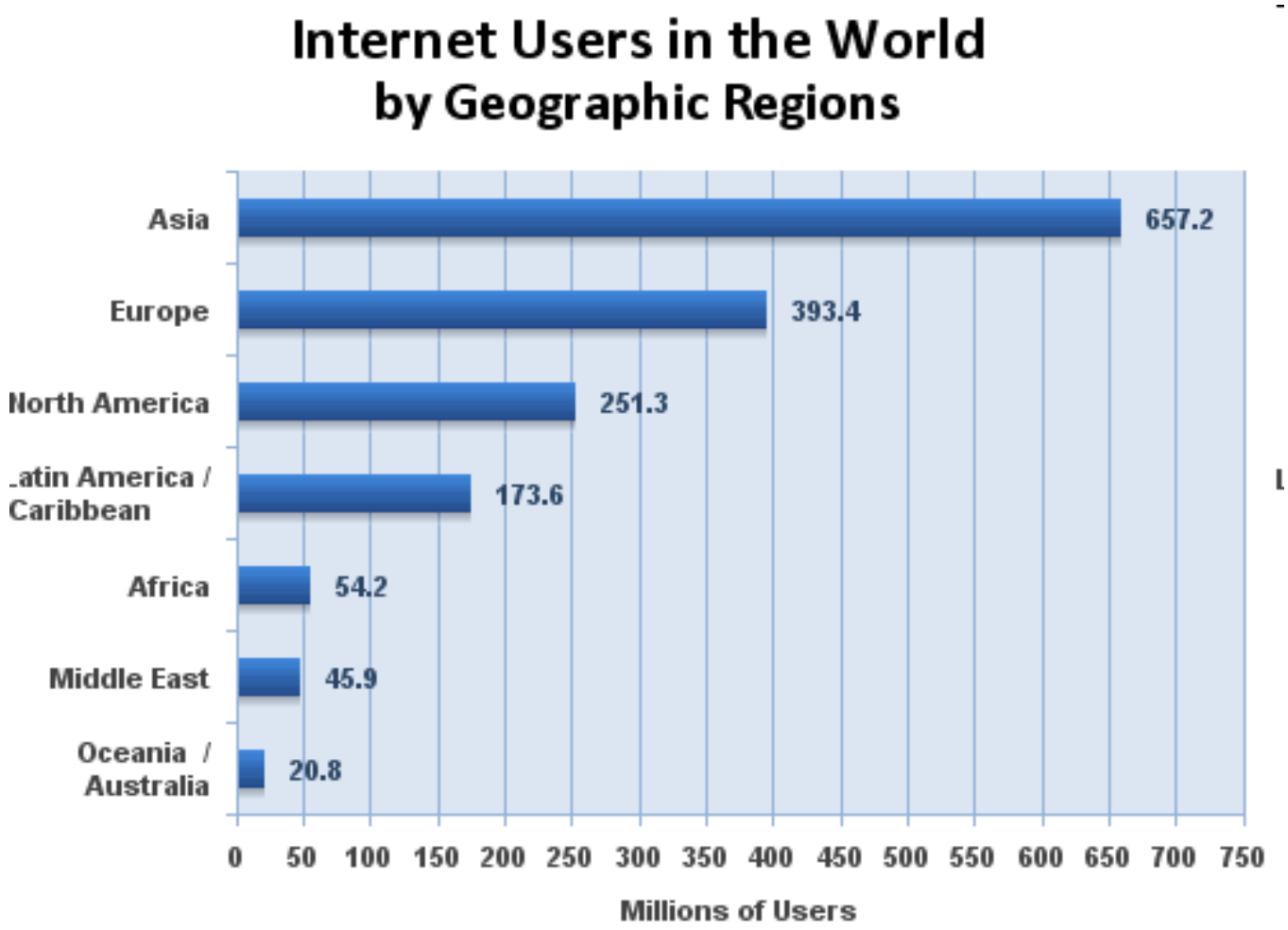
CC The Daily Show with Jon Stewart: Thu, Apr 30, 2009

Season 14 : Ep. 59 (21:36)

+ queue

More: The Daily Show with Jon Stewart
Channel: Comedy

Internet and the Web



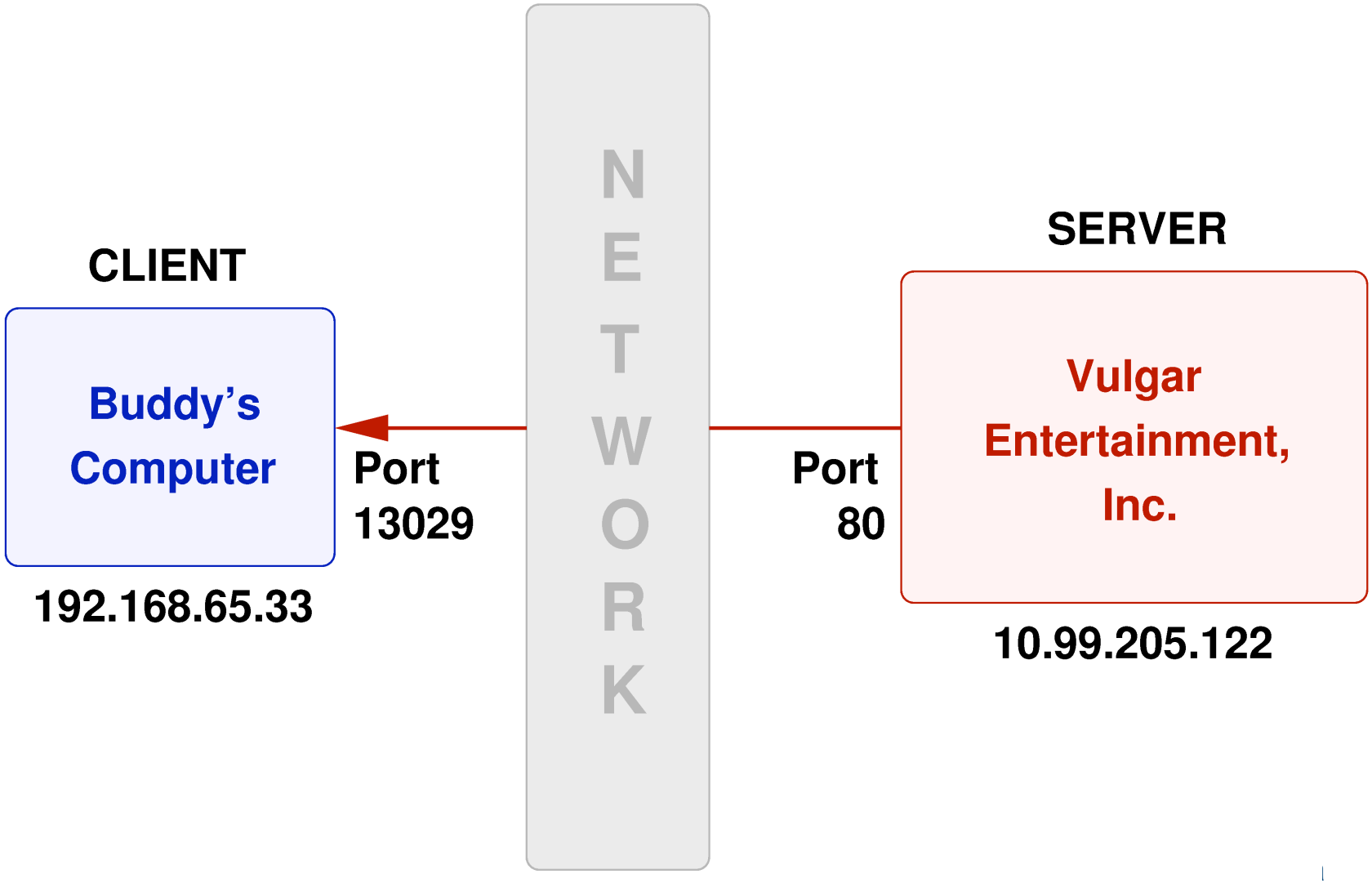
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Estimated Internet users are 1,596,270,108 for March 31, 2009

Web navigation & navigation traces

http://www.a.edu

http://www.b.edu

Navigation traces



Navigation traces (Web requests)

- *Source MAC: 03:5a:66:17:90:5e*
- *Dest. MAC: 10:99:19:3f:51:2f*

- *Source IP: 192.168.39.190*
- *Dest. IP: 127.100.251.3*

- *Source Port: 9421*
- *Dest. Port: 80*

- *GET /index.html HTTP/1.1*
- *Agent: SuperCrawler-2009/beta*
- *Referer: http://www.grumpy-puppy.com/*
- *Host: www.happy-kitty.com*

Why to study navigation traces?

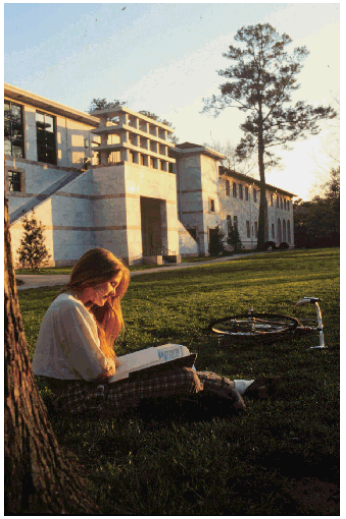
- Privacy concerns
- High potential benefits

Why to study navigation traces?

- We are interested in navigation traces from the point of view of the study of human activity and human interaction with an information-based system.
- Our final aim would be to be able to model this navigation processes in a realistic way.
- All our data is properly anonymized, and we comply with the laws or rules for privacy protection

Databases

Emory University



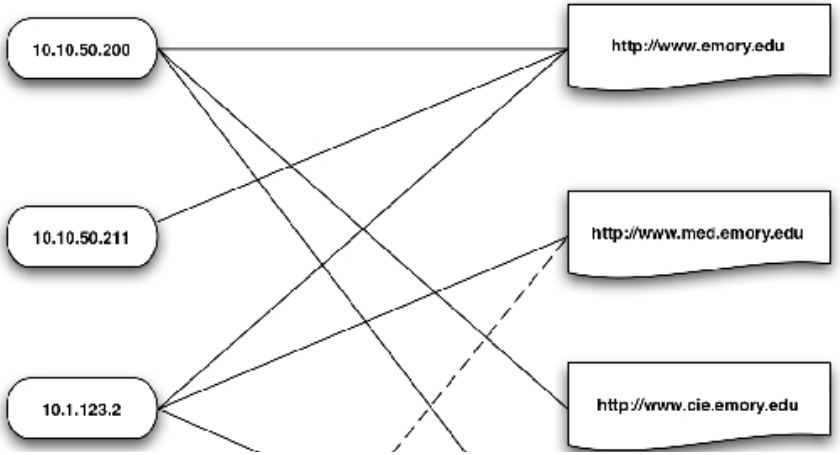
- Students: 12,300
- Faculty: ~ 3,200

Indiana University, Bloomington

- Students: 42,000
- Faculty: ~ 5,000

Databases (Emory University)

- The database is formed by the weblogs of Emory University from Apr. 1st 2005 to Jan. 17th 2006 (41 weeks).
- Each click in a web of the university is registered at the time resolution of 1 second.

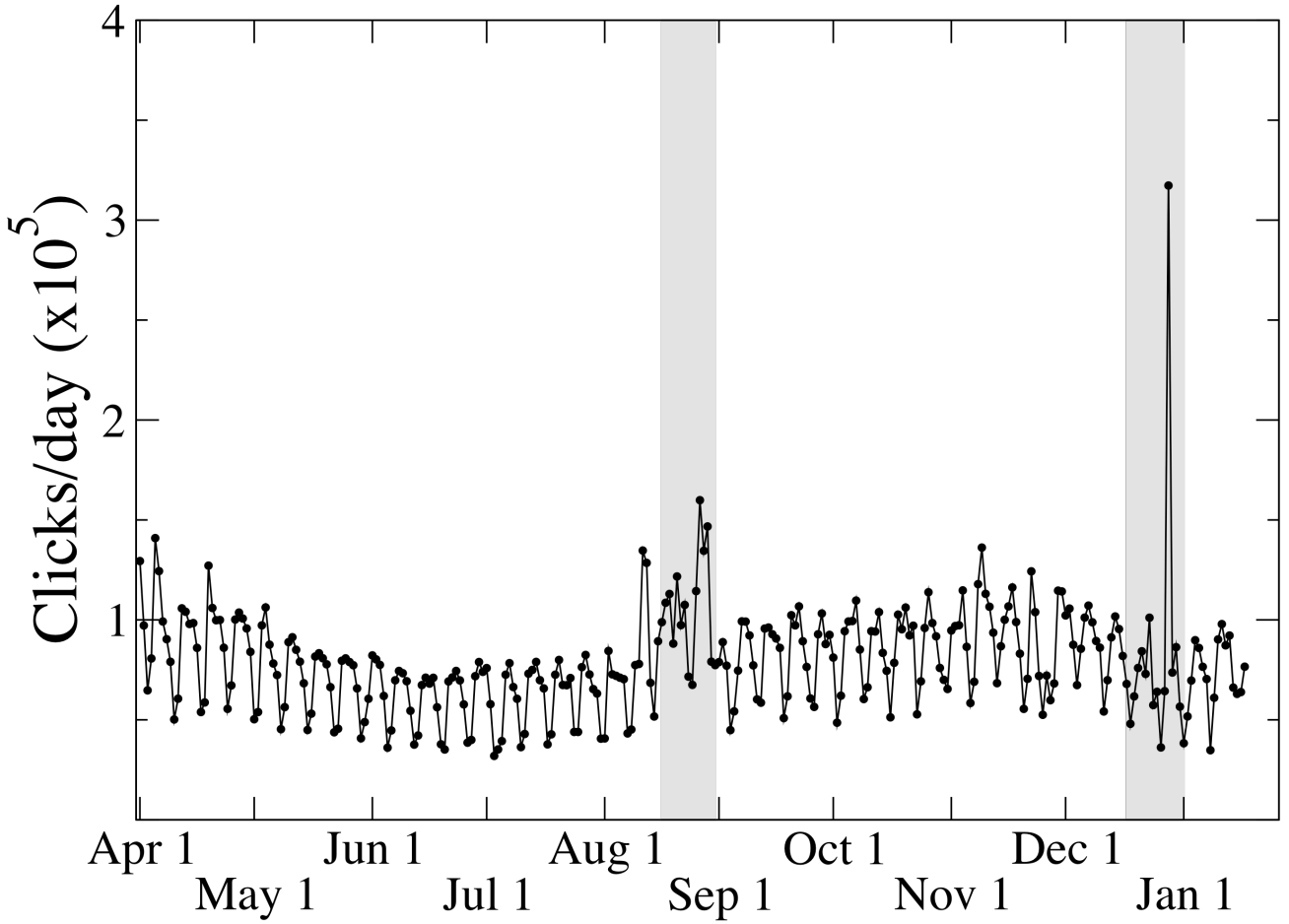


Number of IPs	N_{IP}	$3,179,671$
Number of URLs	N_{URL}	$2,562,398$
Total Number of page requests (weight)	Ω	$53,582,121$
Average number of IPs introduced per day	n_{IP}	$10,742$
Average number of URLs introduced per day	n_{URL}	$8,396$
Average number of edges introduced per day	e	$77,569$
Average wieght increment per day	$\Omega^{\dagger}$	$186,350$

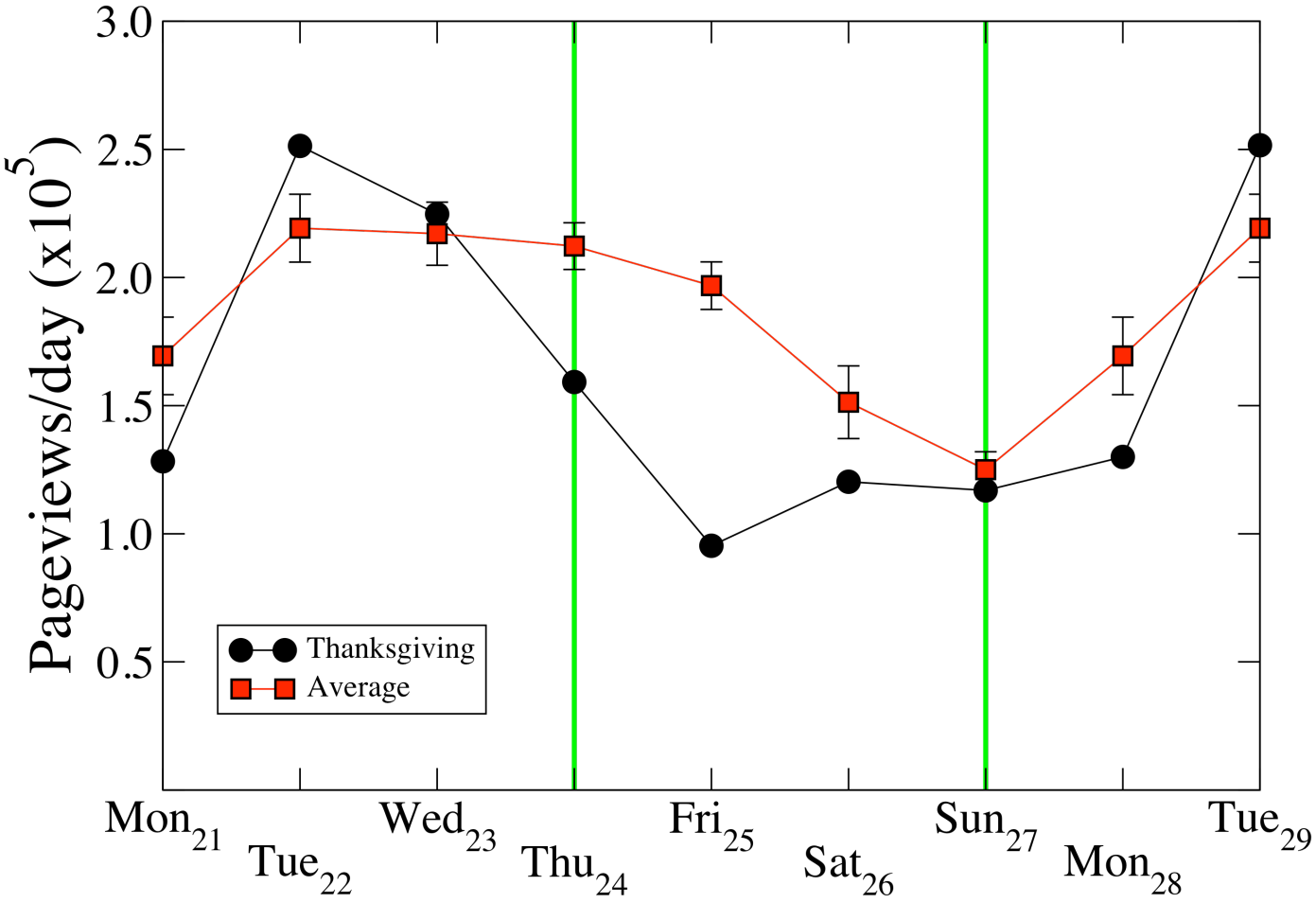
Databases (Indiana University)

- The database is formed by the Web requests from a dorm of the University.
- Data collected from March 5, 2008 through May 3, 2008
- 408 million HTTP requests
- 1083 unique MAC addresses (Computers).
- 29.8 million page requests
- 967 unique users
- 630,000 Web servers
- 110,000 referring hosts

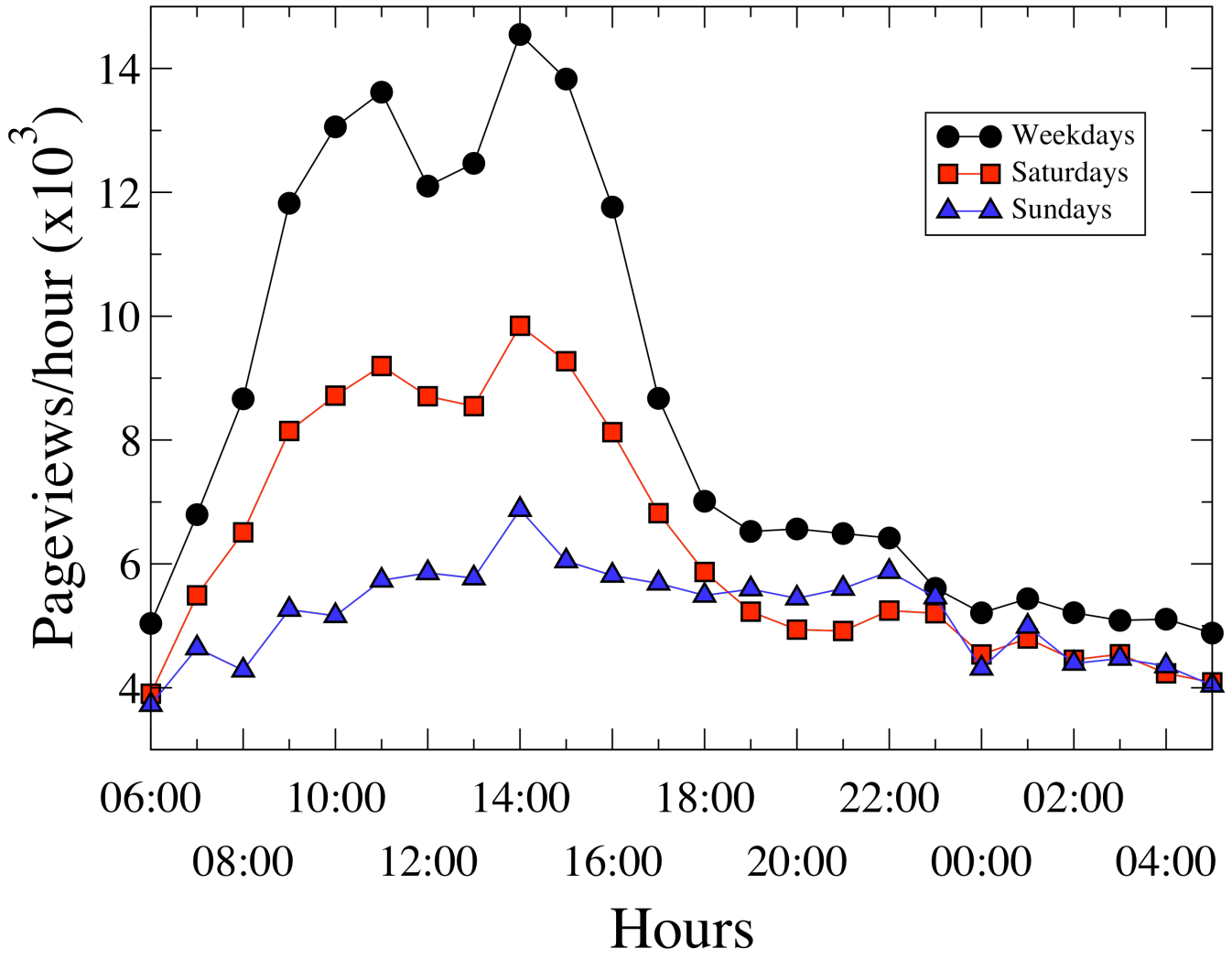
Aggregate results



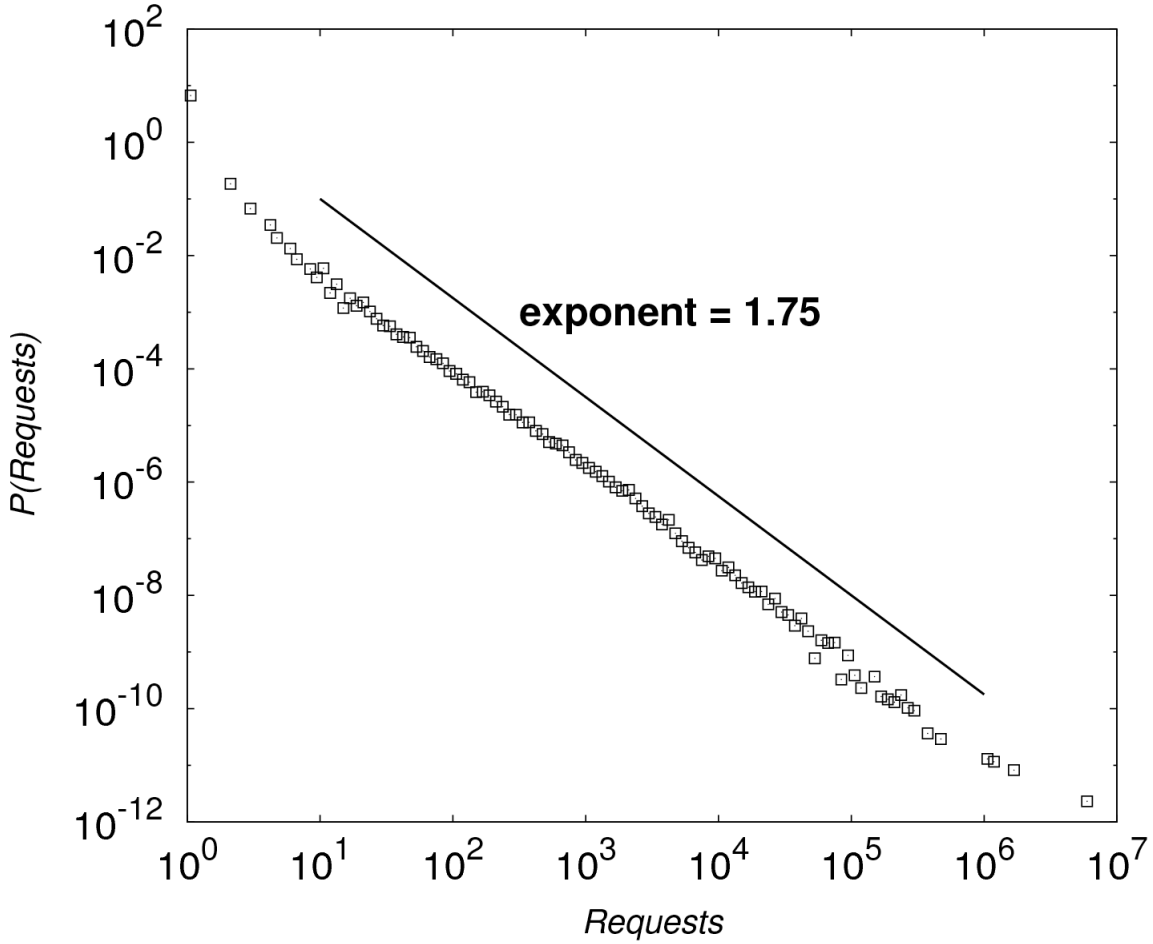
Aggregate results



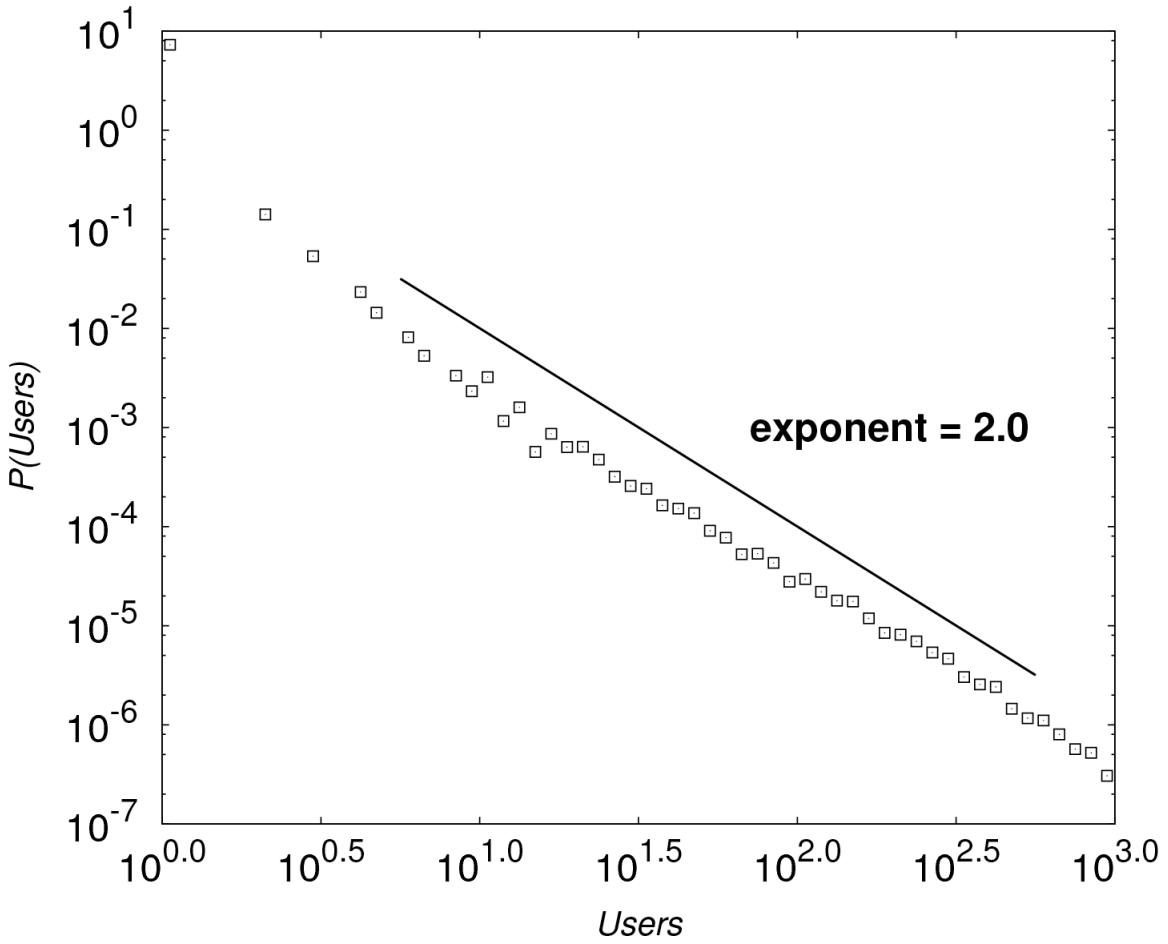
Aggregate results



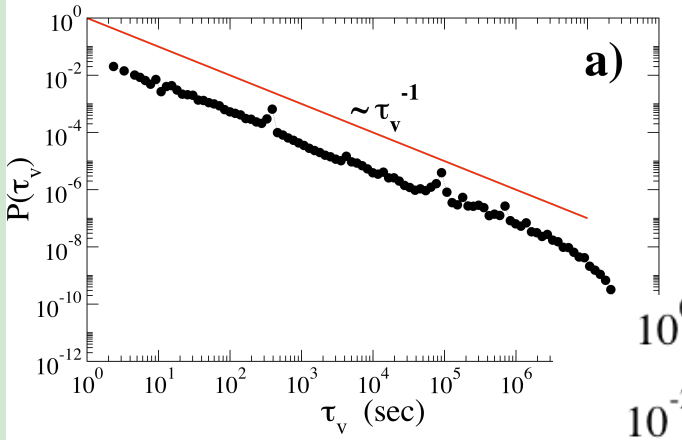
Aggregate results

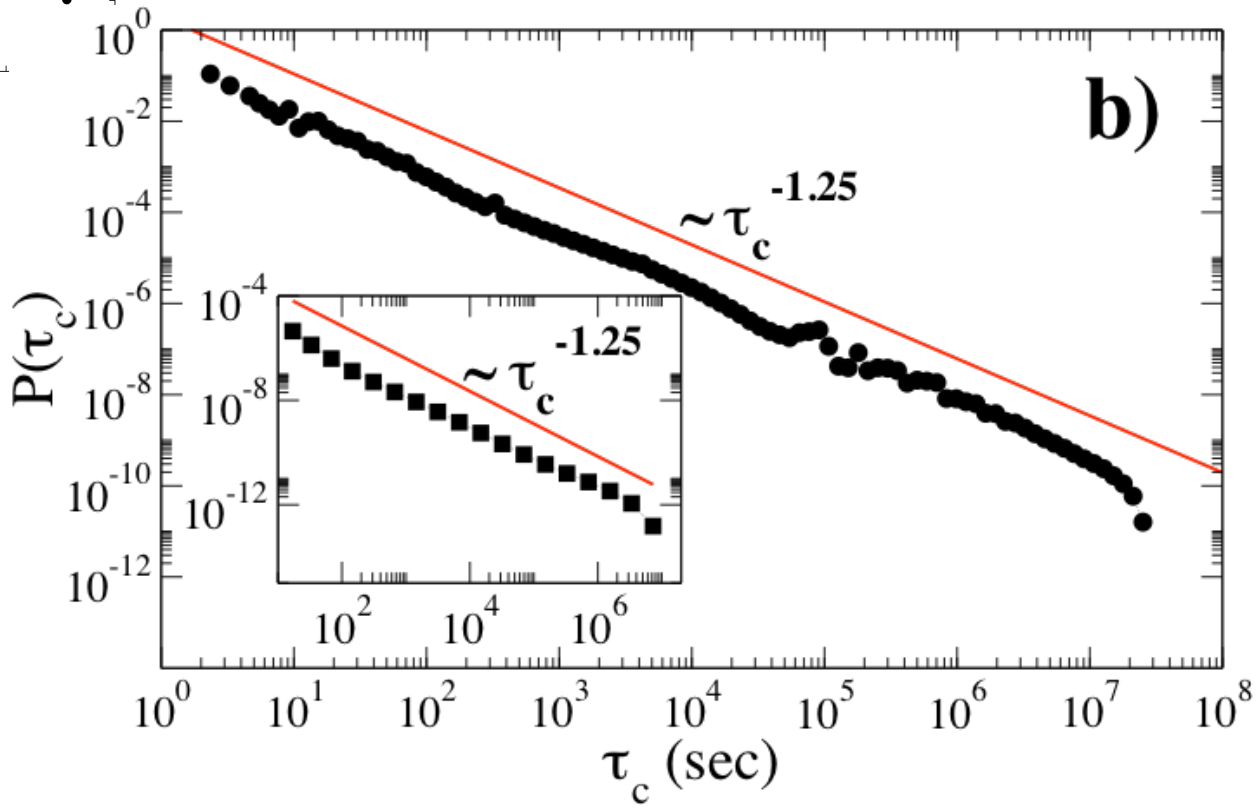


Aggregate results



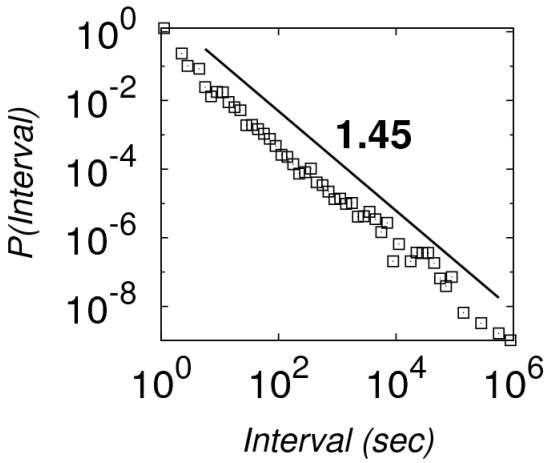
Aggregate results

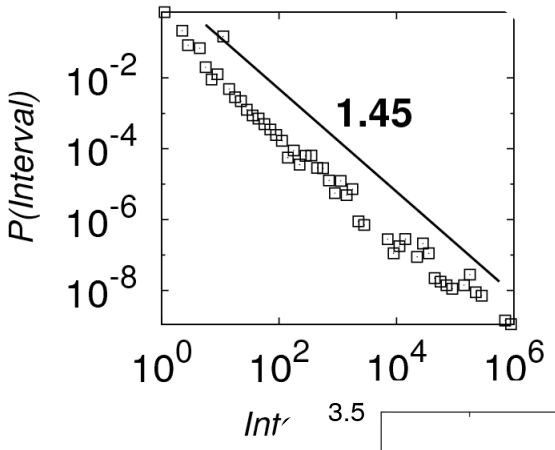


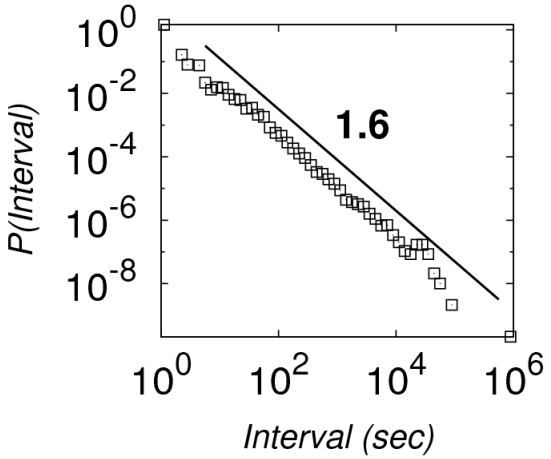


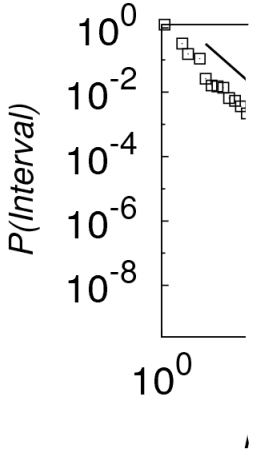
IP ➔ www.x.emory.edu/*

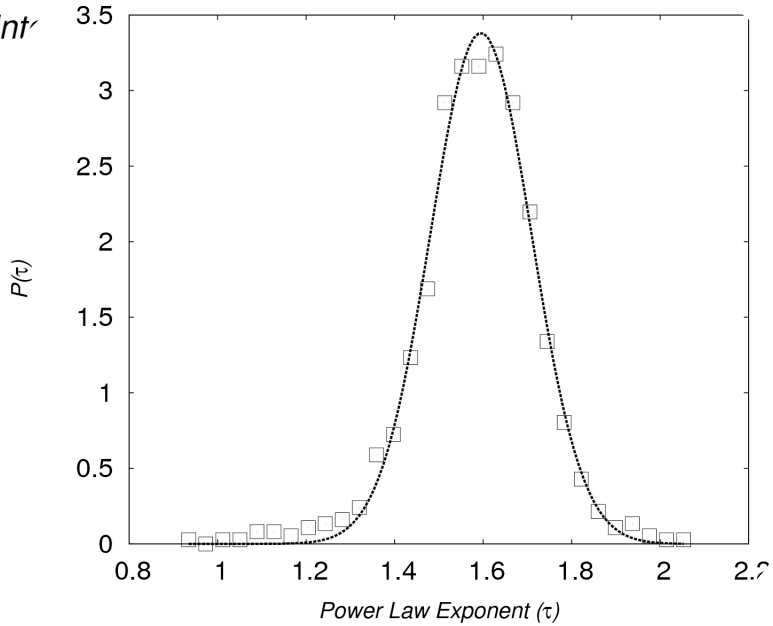
Individual users results



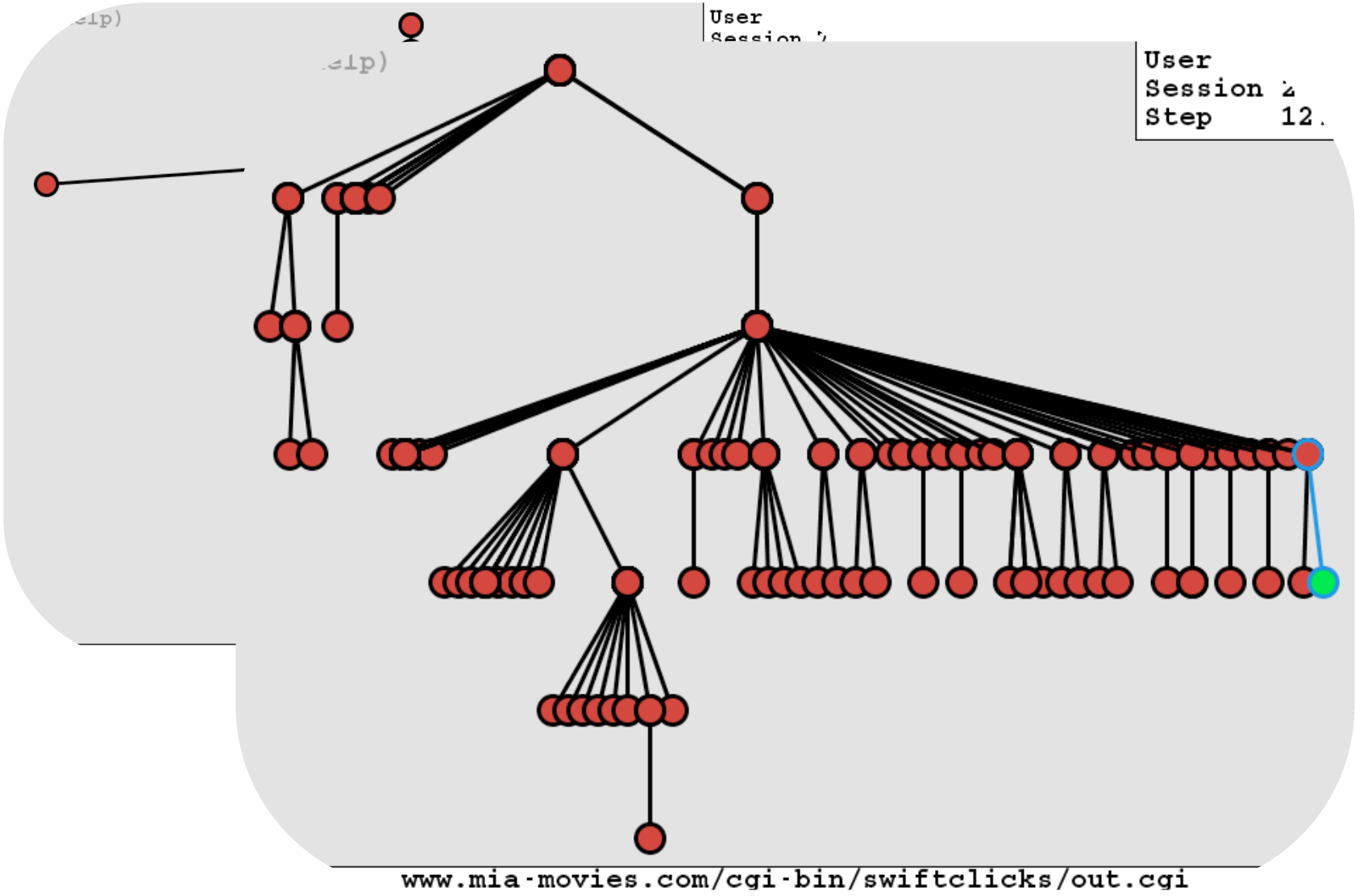








Individual users results (Sessions)



Models: PageRank

PageRank is the simplest navigation model. The basic rules are:

- The users perform a random walk in the Web.
- With a certain probability p, they "teletransport".
- Each of these "teletransportation" events mark the beginning of a new session.

Models: BookRank

We added a further detail in order to mimic the user Web surfing activity:

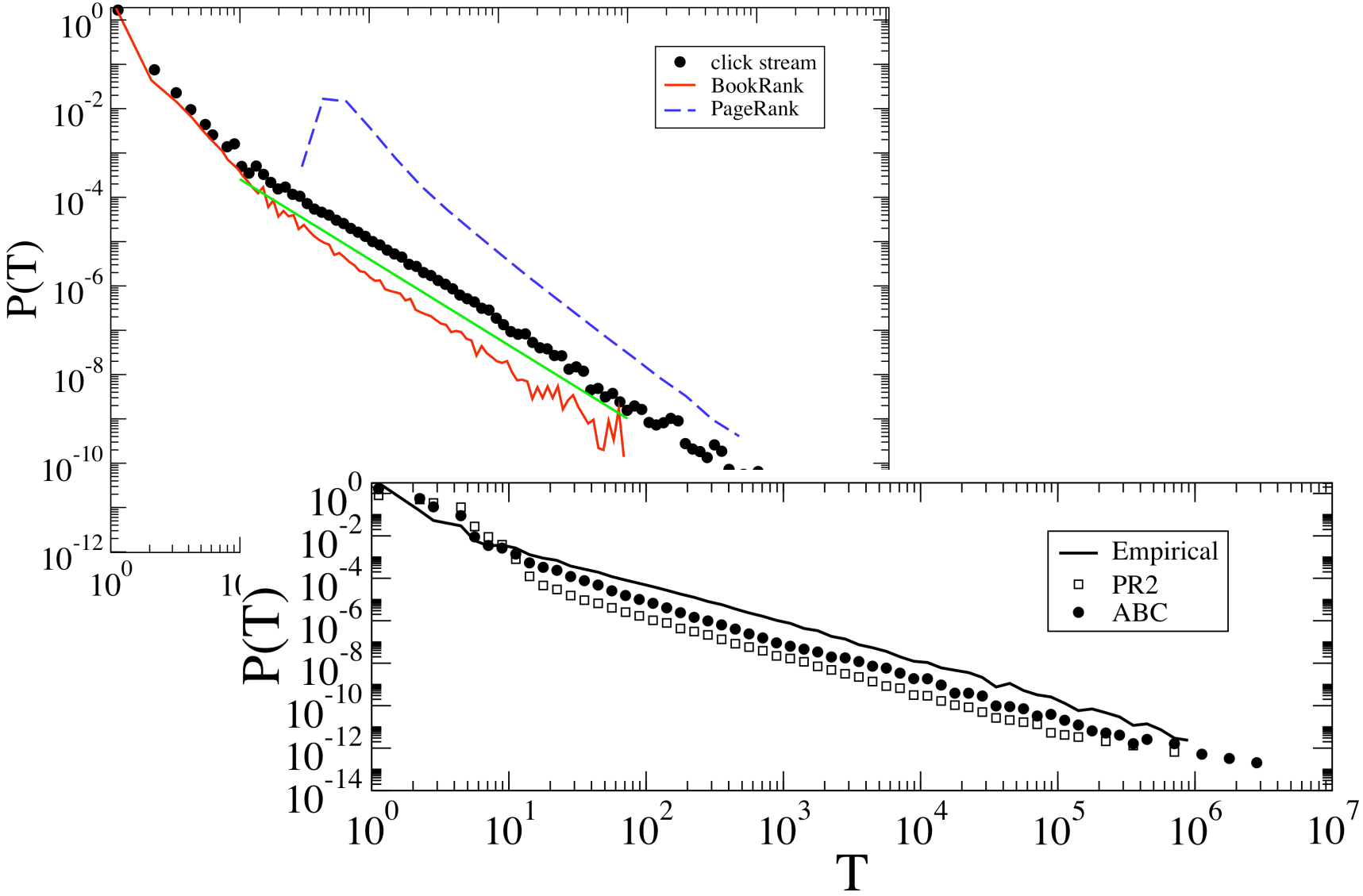
- Each user keeps a list of pages visited (bookmarks)
- They are ordered according to the number of times he/she visited the pages (rank r)
- Each time a user starts a new session, starting page selected from the bookmark list with prob $\sim r^{-a}$
- Back bottom, p

Models: bookmarks + topicality (ABC)

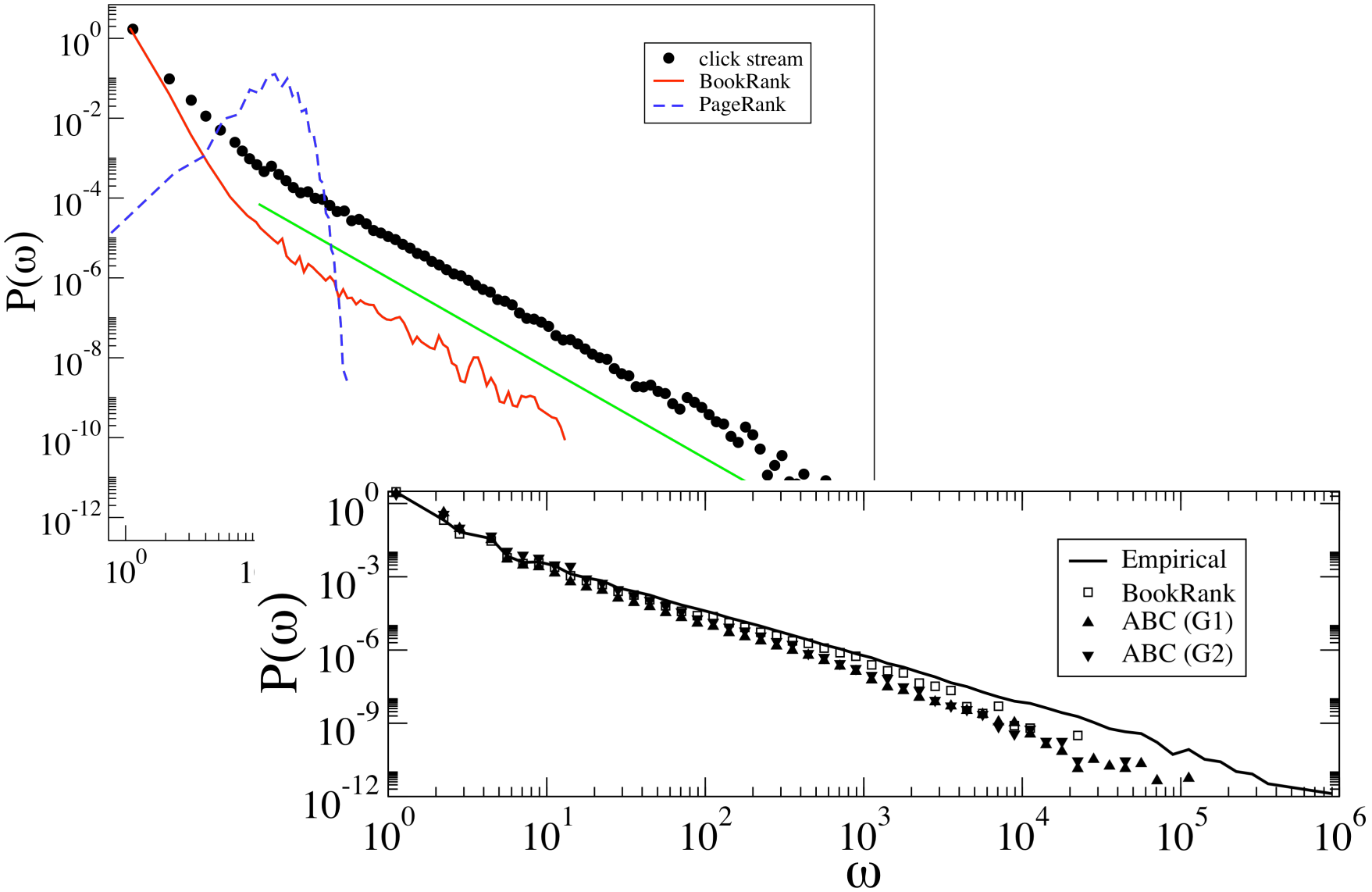
In order to reproduce the single user traces we needed to add yet another ingredient related to pages topicality:

- Each trace starts with a E level
- There is a cost for each action $E_t = E_{t-1} - C$
- For each new page visited the $E_t = (1-\Delta\ \eta)\ E_{t-1}$
- If $E_t < 0$, new session

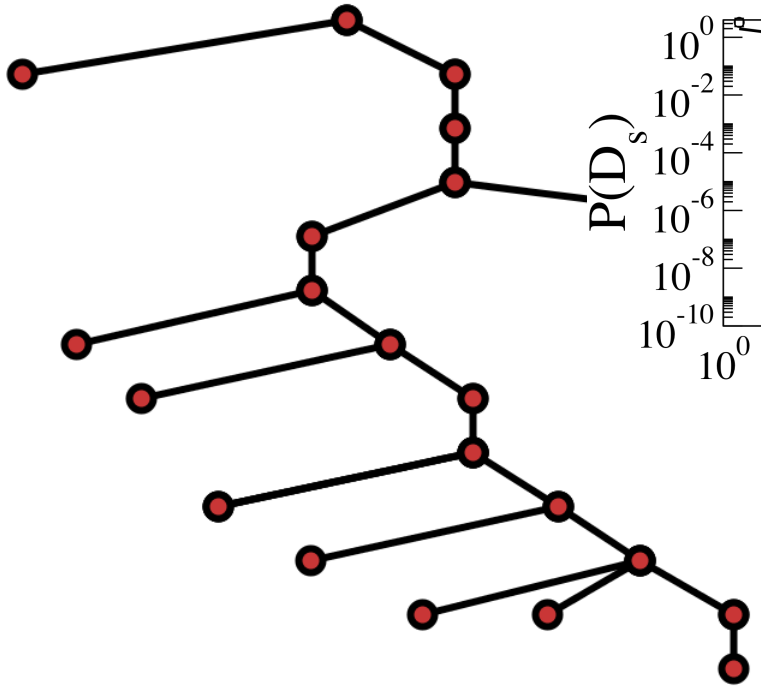
Simulation vs empirical data

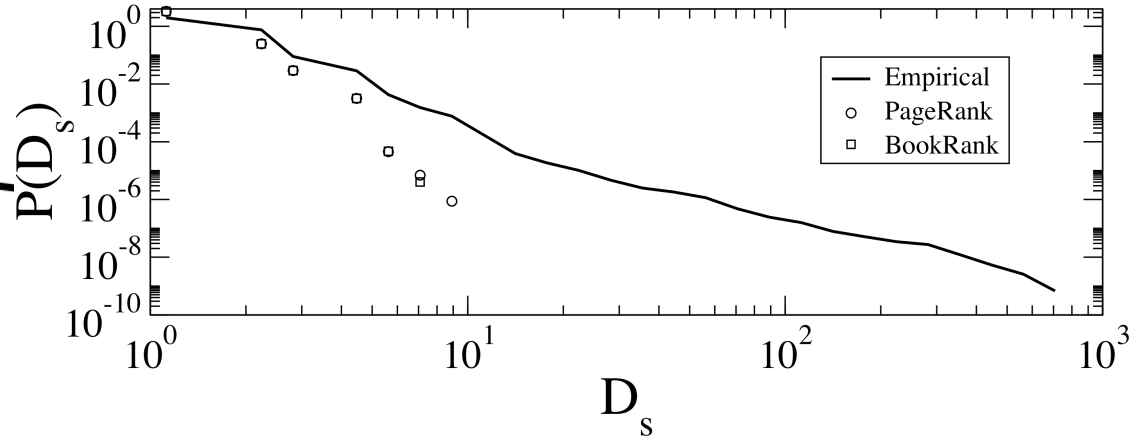


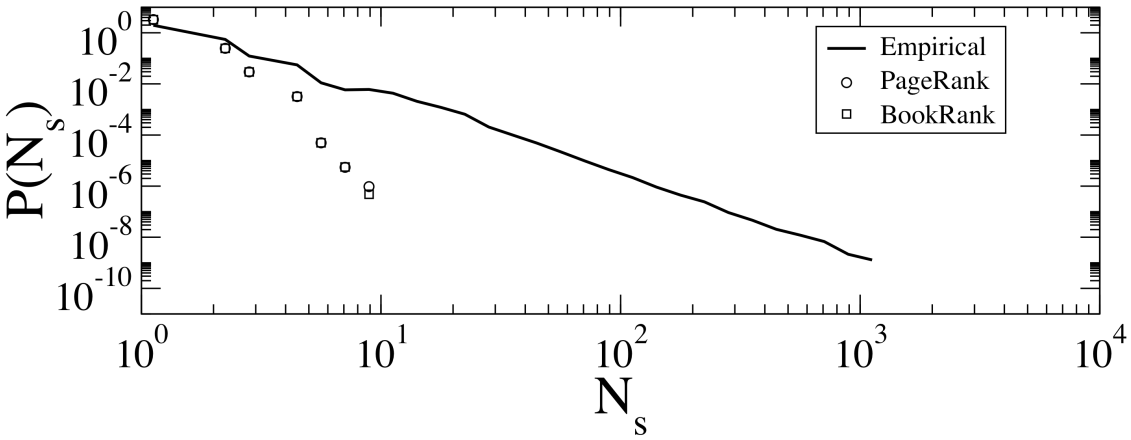
Simulation vs empirical data



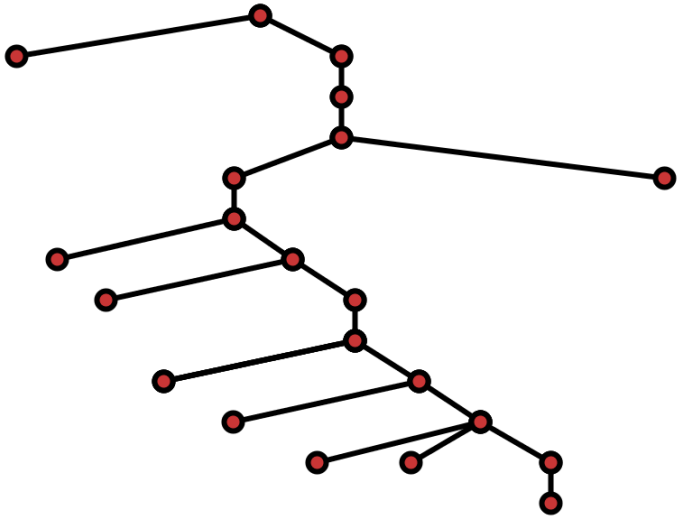
Simulation vs empirical data



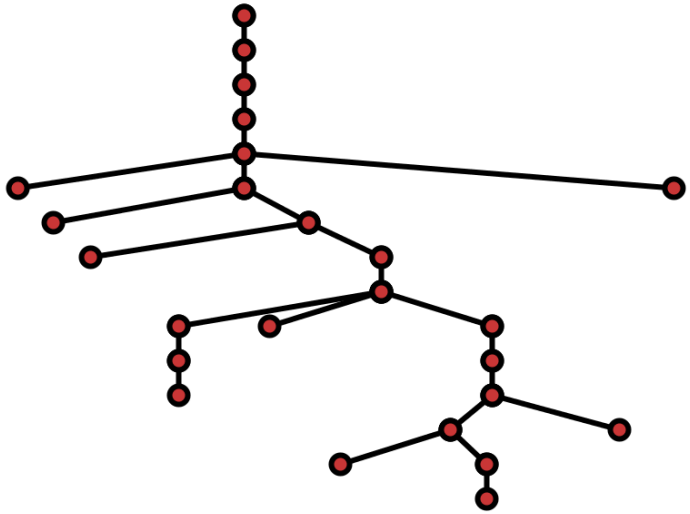




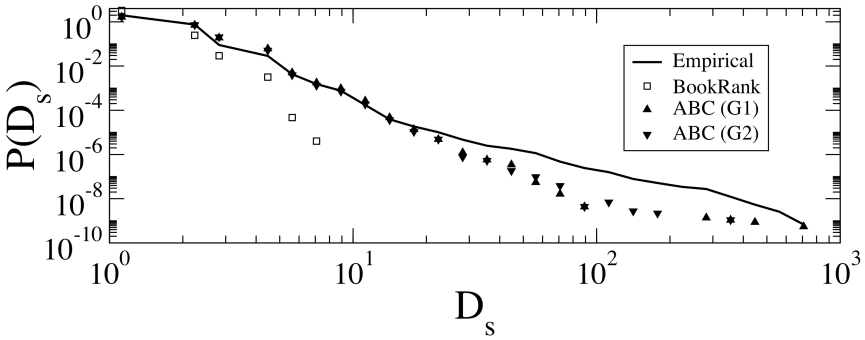
Simulation vs empirical data

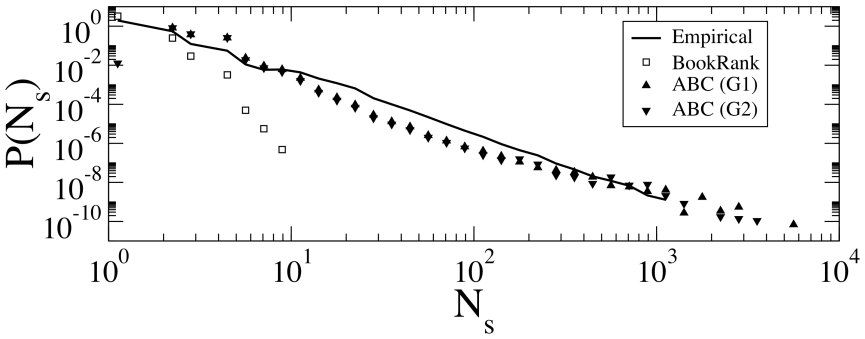


Data



Model

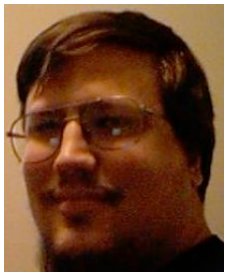




Conclusions

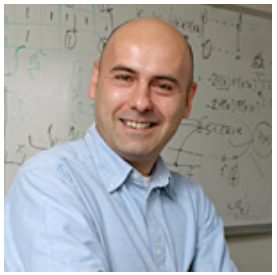
- We have studied the Web navigation traces of a large number of users.
- Some of the features seem to be relatively universal despite natural user-user variability.
- We have proposed a family of models able to reproduce deeper and deeper characteristics of the users' navigation patterns.
- How far should we go? Do this last simple model implement topicality satisfactorily? And what about real time dynamics? …

Collaborators & papers



Mark Meiss

Bruno Gonçalves



Sandro Flammini

Fil Menczer

- Human dynamics unveiled through Web Analytics, Phys. Rev. E **78** , 026123 (2008)

- Remembering what we like: Toward an agent-based model of Web traffic, WSDM 2009 Late Breaking Results

- What's in a Session: Tracking Individual Behavior on the Web, Hypertext 2009

- Agents, Bookmarks and Clicks: A topical model of Web traffic, Hypertext 2010