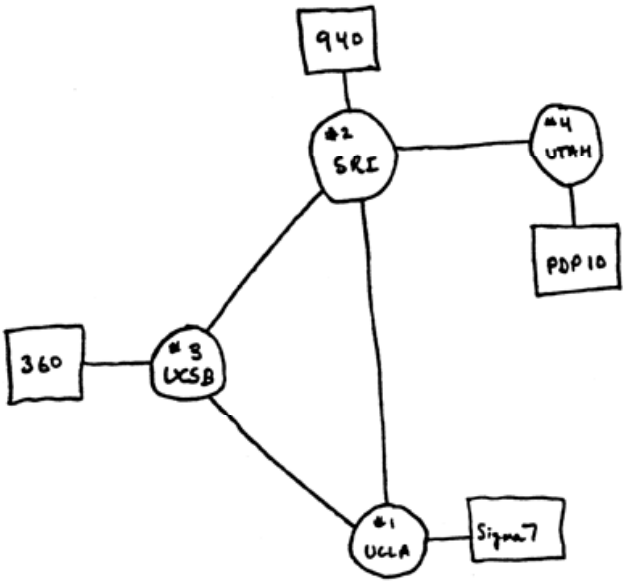
Structural Mining of Large-Scale Behavioral Data from the Internet

Thesis Defense

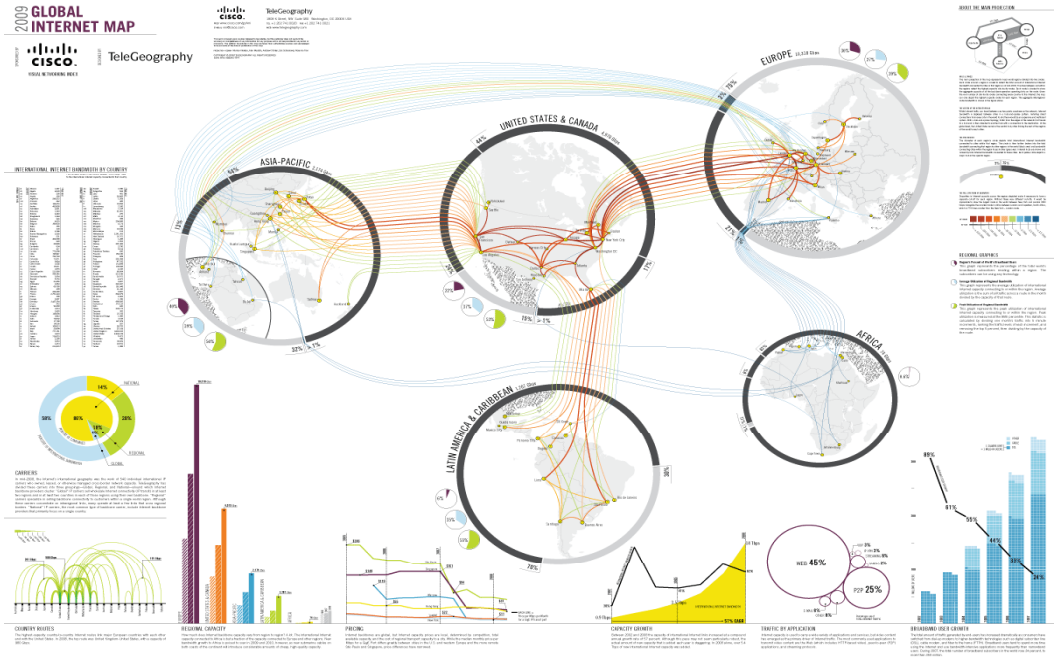
Mark Meiss

April 30, 2010

The Internet in 1969



The Internet in 2010



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

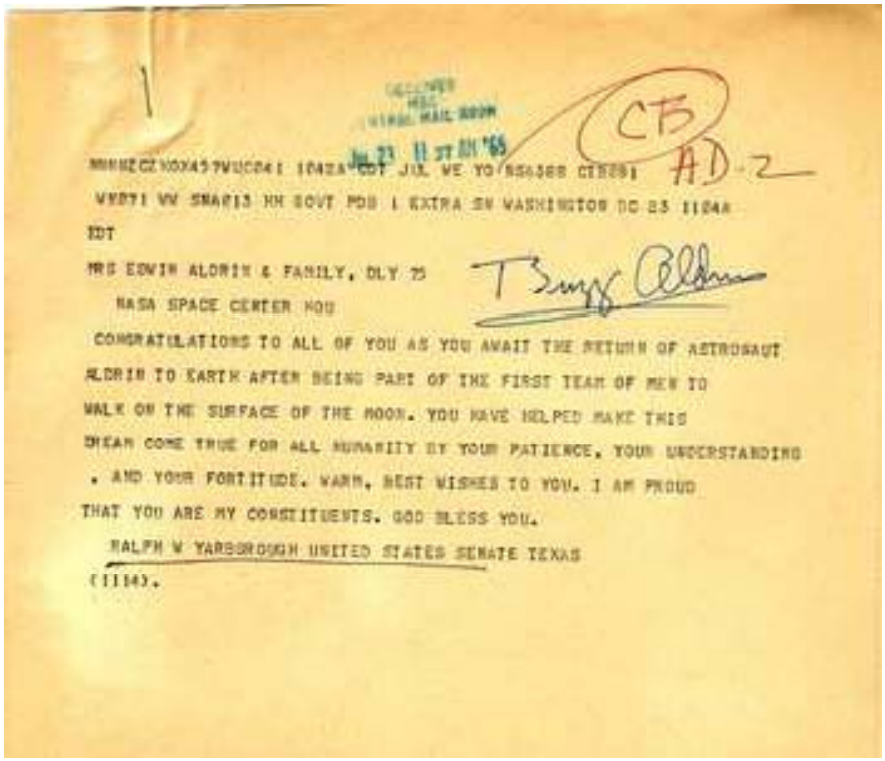
CC Family Guy: 420
Season 7 : Ep. 12 (21:53)
+ queue
More: Family Guy
Channel: Comedy

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

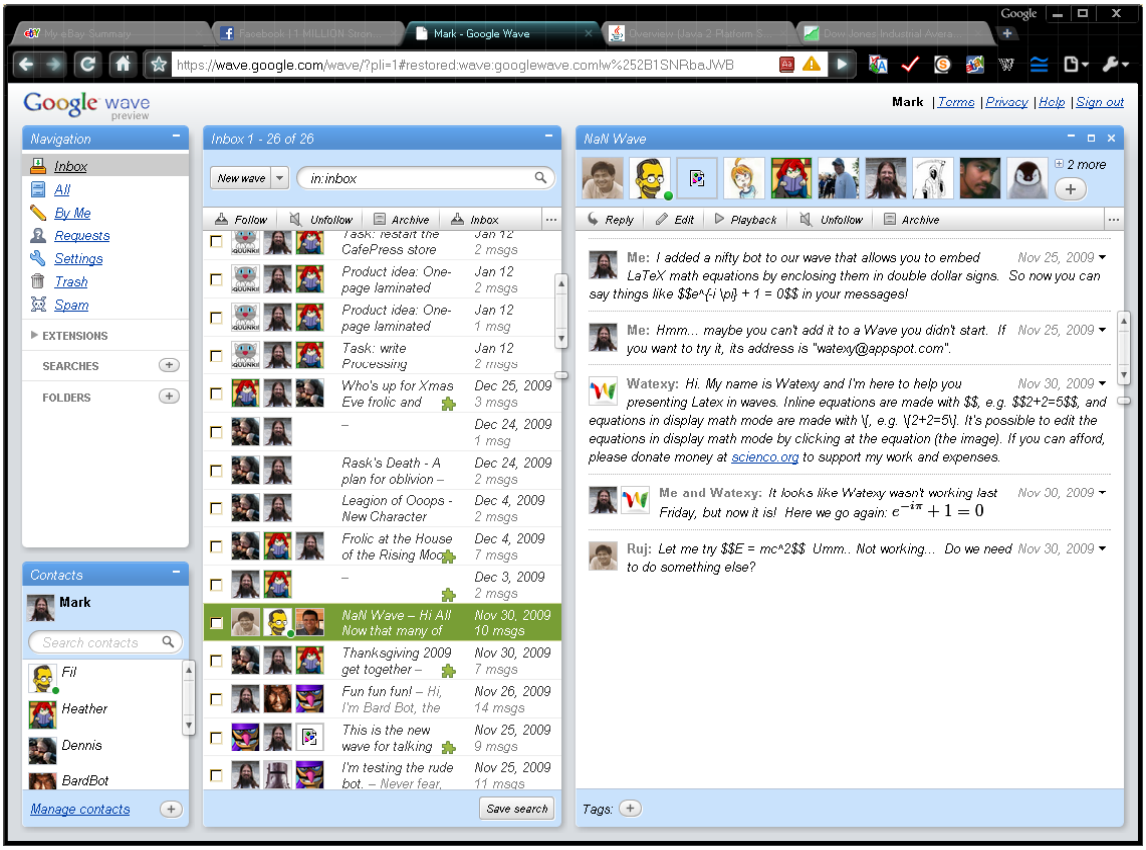


CB
AD-2

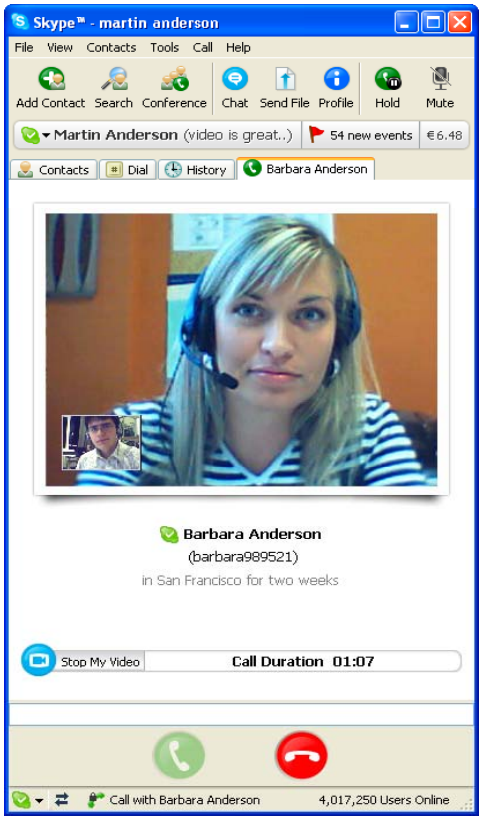
MRS EDWIN ALDRIN & FAMILY, DLY 25
NASA SPACE CENTER HOU

CONGRATULATIONS TO ALL OF YOU AS YOU AWAIT THE RETURN OF ASTRONAUT ALDRIN TO EARTH AFTER BEING PART OF THE FIRST TEAM OF MEN TO WALK ON THE SURFACE OF THE MOON. YOU HAVE HELPED MAKE THIS DREAM COME TRUE FOR ALL HUMANITY BY YOUR PATIENCE, YOUR UNDERSTANDING, AND YOUR FORTITUDE. WARM, BEST WISHES TO YOU. I AM PROUD THAT YOU ARE MY CONSTITUENTS. GOD BLESS YOU.

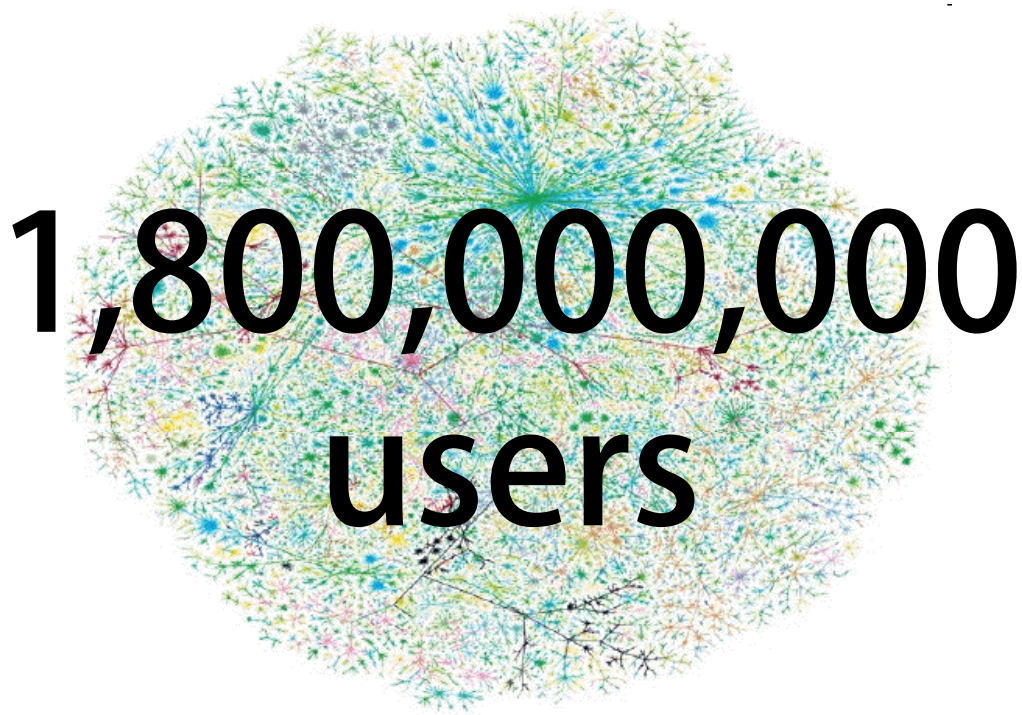
RALPH W YARBOROUGH UNITED STATES SENATE TEXAS


Mark - Google Wave
https://wave.google.com/wave/?pli=1#restored:wave:googlewave.com!w%252B1SNRbaJWB
Google wave preview
Mark | Terms | Privacy | Help | Sign out
Navigation
Inbox
All
By Me
Requests
Settings
Trash
Spam
EXTENSIONS
SEARCHES
FOLDERS
Contacts
Mark
Search contacts
Fil
Heather
Dennis
BardBot
Manage contacts
Inbox 1 - 26 of 26
New wave
in:inbox
Follow
Unfollow
Archive
Inbox
Task: restart the CafePress store Jan 12 2 msgs
Product idea: One-page laminated Jan 12 2 msgs
Product idea: One-page laminated Jan 12 1 msg
Task: write Processing Jan 12 2 msgs
Who's up for Xmas Eve frolic and Dec 25, 2009 3 msgs
– Dec 24, 2009 1 msg
Rask's Death - A plan for oblivion – Dec 24, 2009 2 msgs
Leagion of Ooops - New Character Dec 4, 2009 2 msgs
Frolic at the House of the Rising Moo Dec 4, 2009 7 msgs
– Dec 3, 2009 2 msgs
NaN Wave – Hi All Now that many of Nov 30, 2009 10 msgs
Thanksgiving 2009 get together – Nov 30, 2009 7 msgs
Fun fun fun! – Hi, I'm Bard Bot, the Nov 26, 2009 14 msgs
This is the new wave for talking Nov 25, 2009 9 msgs
I'm testing the rude bot. – Never fear, Nov 25, 2009 11 msgs
Save search
NaN Wave
2 more
Reply
Edit
Playback
Unfollow
Archive
Me: I added a nifty bot to our wave that allows you to embed LaTeX math equations by enclosing them in double dollar signs. So now you can say things like $$e^{-i \pi} + 1 = 0$$ in your messages! Nov 25, 2009
Me: Hmm... maybe you can't add it to a Wave you didn't start. If you want to try it, its address is "watexy@appspot.com". Nov 25, 2009
Watexy: Hi. My name is Watexy and I'm here to help you presenting Latex in waves. Inline equations are made with $$, e.g. $$2+2=5$$, and equations in display math mode are made with \[, e.g. \[2+2=5\]. It's possible to edit the equations in display math mode by clicking at the equation (the image). If you can afford, please donate money at scienco.org to support my work and expenses. Nov 30, 2009
Me and Watexy: It looks like Watexy wasn't working last Friday, but now it is! Here we go again: $e^{-i\pi} + 1 = 0$ Nov 30, 2009
Ruj: Let me try $$E = mc^2$$ Umm.. Not working... Do we need to do something else? Nov 30, 2009
Tags:


Spark Fun


Skype™ - martin anderson
File View Contacts Tools Call Help
Add Contact Search Conference Chat Send File Profile Hold Mute
Martin Anderson (video is great..)
54 new events
€6.48
Contacts Dial History Barbara Anderson
Barbara Anderson
(barbara989521)
in San Francisco for two weeks
Stop My Video
Call Duration 01:07
Call with Barbara Anderson
4,017,250 Users Online


SQL Server
IRC
FTP
World of Warcraft
USENET
WWW
Steam
WinMX
Skype
NFS
ConnectGateway
Battlenet
SSH
email
Flash
eDonkey
Shoutcast
DirectConnect
Tsunami
Nintendo WFC
Bittorrent
Gnutella
Botnets
Back Doors



Key Questions

- Can we make meaningful <u>inferences</u> about user behavior with available sources of data?
- What <u>implications</u> do patterns of network behavior have for its design and structure?
- How can behavioral data be used to <u>understand</u> users and <u>improve</u> network applications?

“how it acts” > “what it is”

Patterns > Payloads

Practicality

Privacy

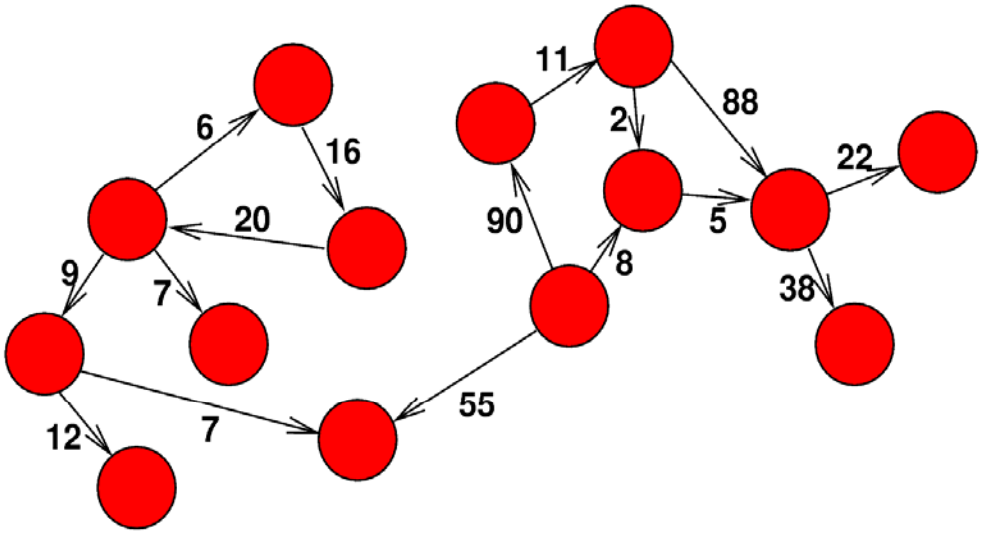
Roadmap

- Background
- Network flow analysis
- Web click analysis
- Conclusions

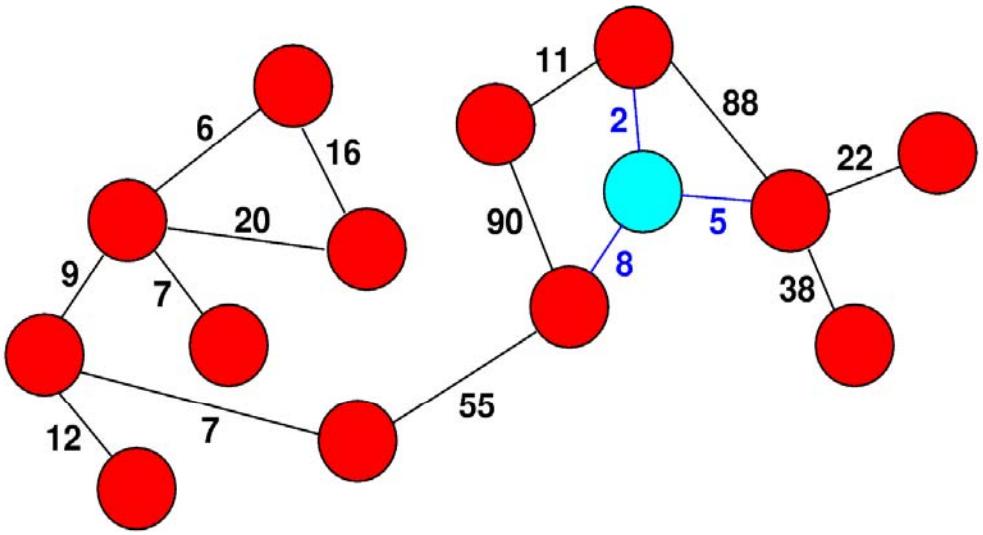
Roadmap

- Background
- Network flow analysis
- Web click analysis
- Conclusions

Weighted digraph

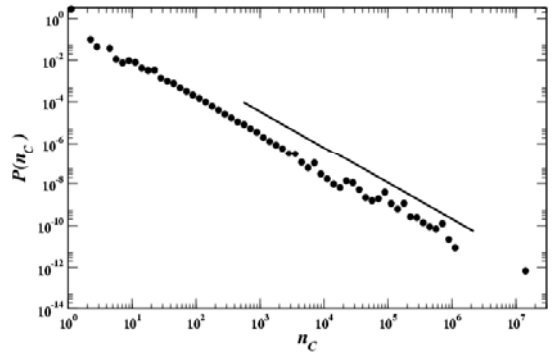


Degree and strength



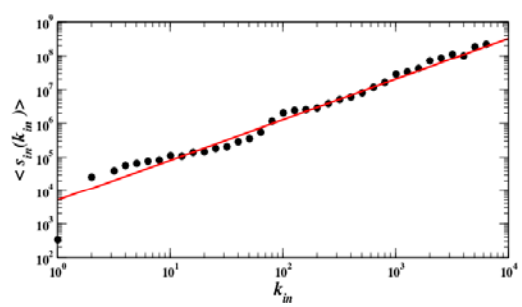
Distributions

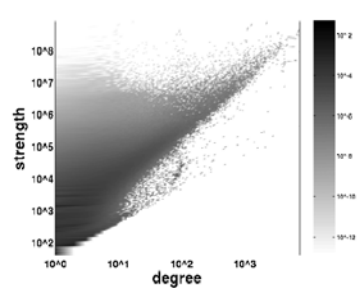
- We can calculate *probability density functions* for degree, strength, etc.
  - Area under curve is 1
  - Can calculate likelihood of a node being in some interval
- Shown here is an example of a very wide-tailed distribution
  - Best approximated by power law



Scaling relations

- We can also investigate how these distributions are *correlated*
- Shown here is a *degree vs. strength* plot.





Other network properties

- *Spectral analysis:* Looking at the eigen{values,vectors} of the connectivity matrix.
- *Clustering: Looking at the density of connections among neighbors of a node.*
- *Assortativity: Looking at whether high-degree nodes connect to other high-degree nodes.*

Roadmap

- Background
- Network flow analysis
- Web click analysis
- Conclusions

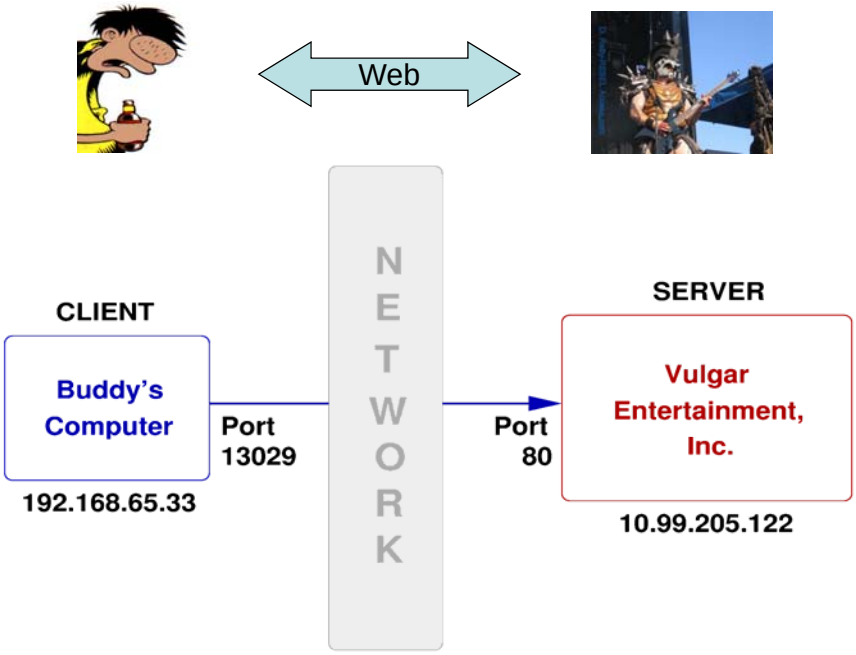
The Internet2/Abilene Network

- TCP/IP network connecting ***research and educational*** institutions in the U.S.
  - Over 200 universities and corporate research labs
- Also provides ***transit service*** between Pacific Rim and European networks

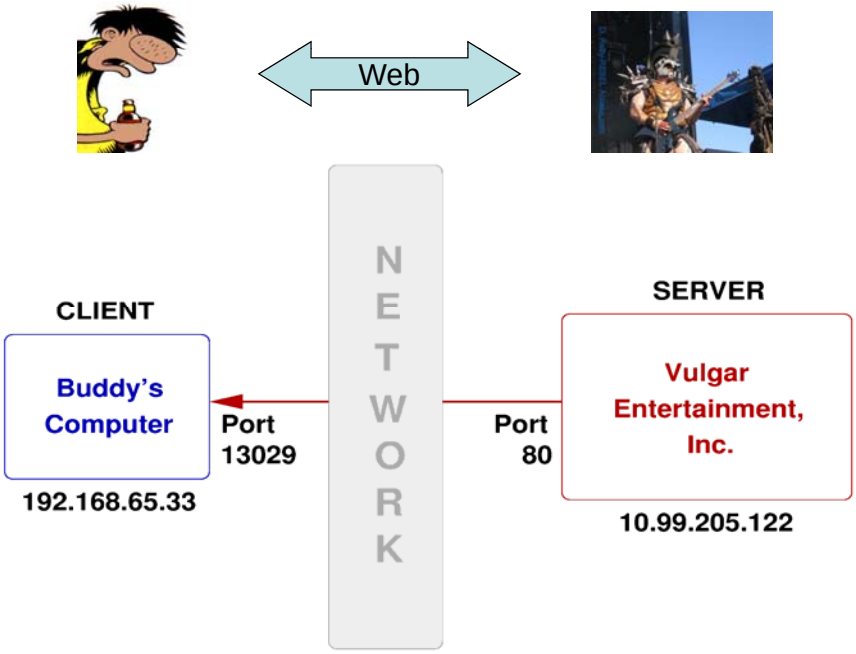
Why study Abilene?

- ***Wide-area network*** that includes both domestic and international traffic
- ***Heterogeneous user base*** including hundreds of thousands of undergraduates
- ***High capacity*** network (10-Gbps fiber-optic links) that has never been congested
- ***Research partnership*** gives access to (anonymized) traffic data unavailable from commercial networks
- ***Variety of traffic*** to both academic and commercial hosts

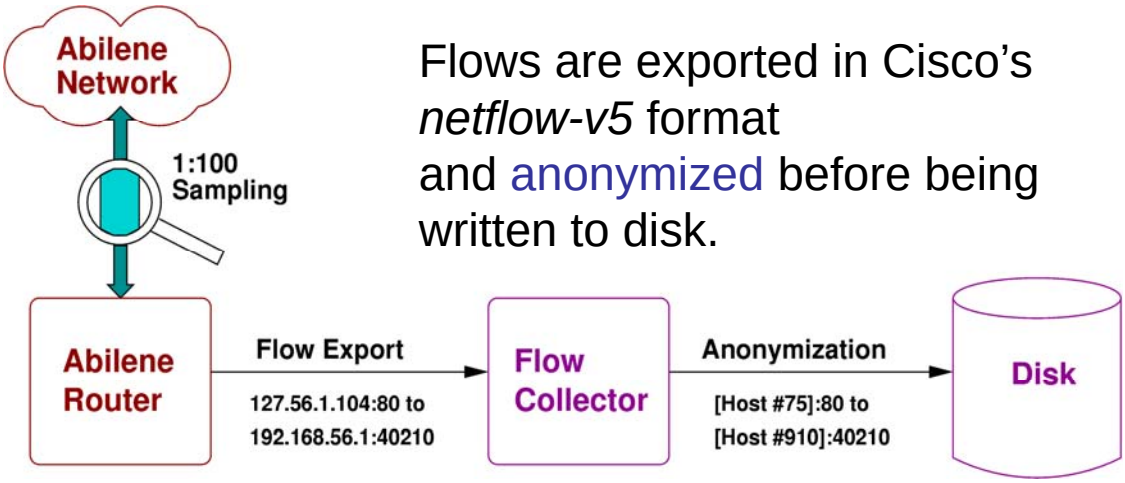
Introducing the “flow”



Introducing the “flow”



Flow collection



Flows are exported in Cisco's *netflow-v5* format and anonymized before being written to disk.

Data dimensions

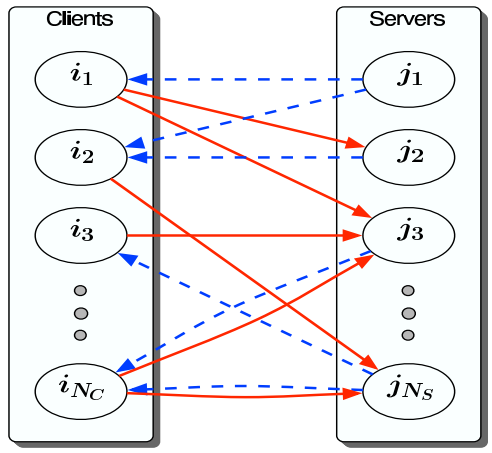
- In a typical day:
  - Over 200 terabytes of data exchanged
  - Almost 1 billion flow records
  - Over 40 gigabytes on disk
  - Over 20 million unique hosts involved

What can you do with a flow?

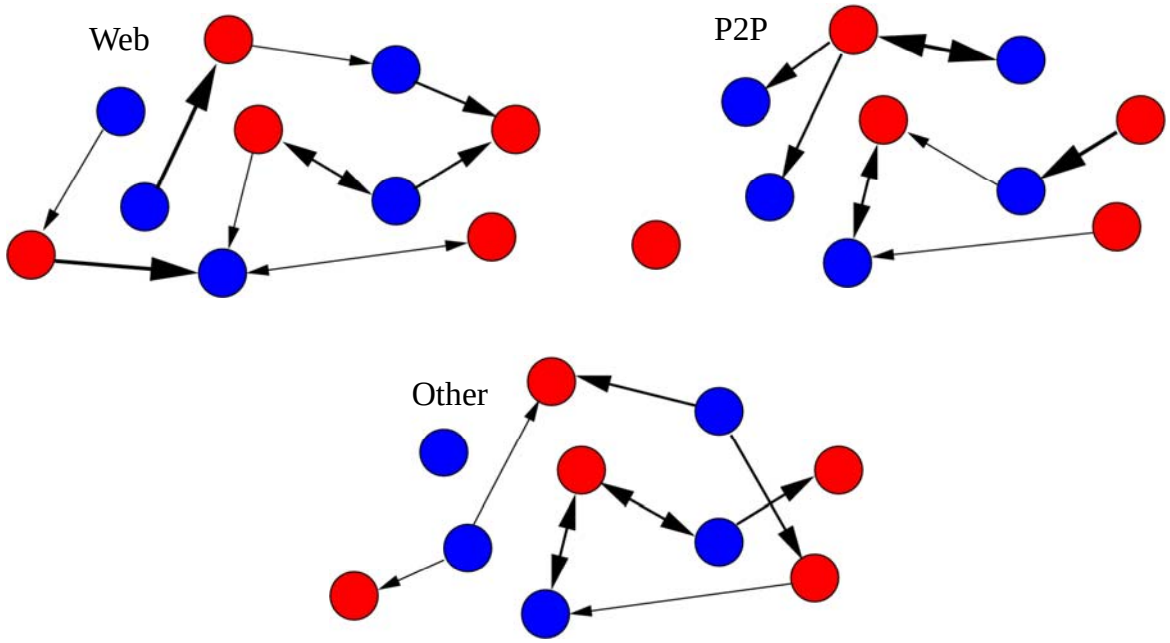
- Standard answer:
  - Treat a flow as a record in a relational database
    - *Who talked to port 1337?*
    - *What proportion of our traffic is on port 80?*
    - *Who is scanning for vulnerable systems?*
    - *Which hosts are infected with this worm?*

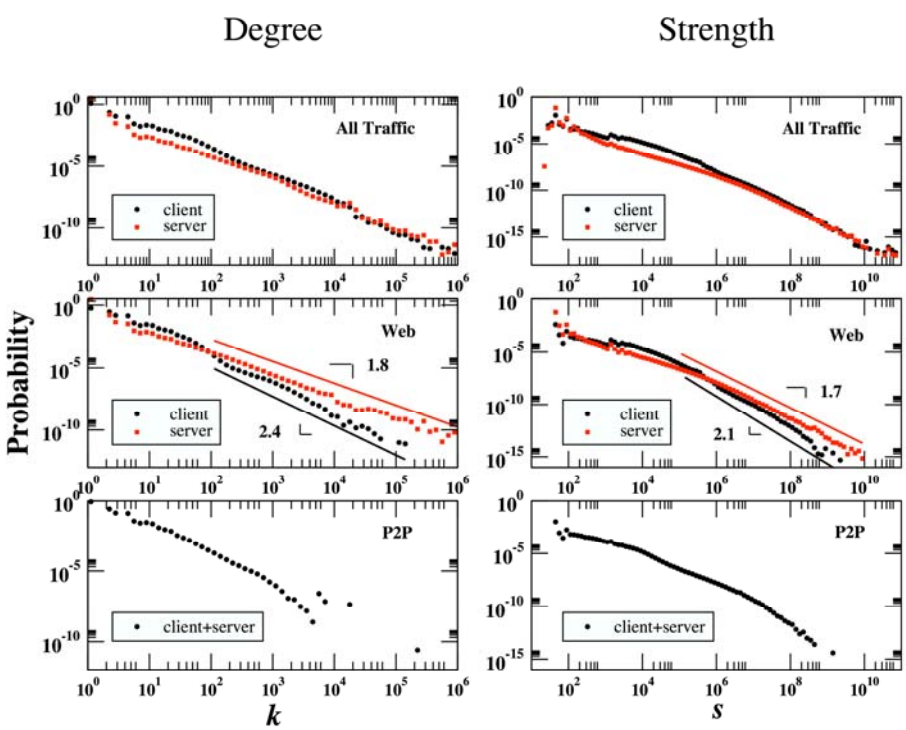
What can you do with a flow?

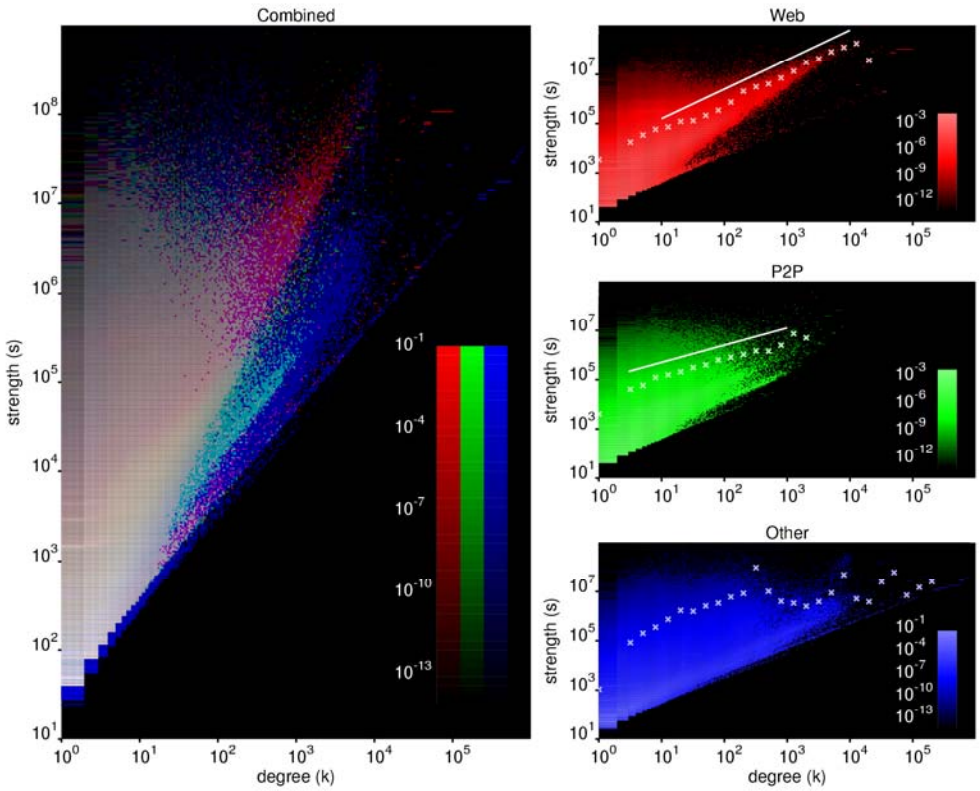
- Graph-centric approach:
  - Treat a flow as a directed, weighted edge

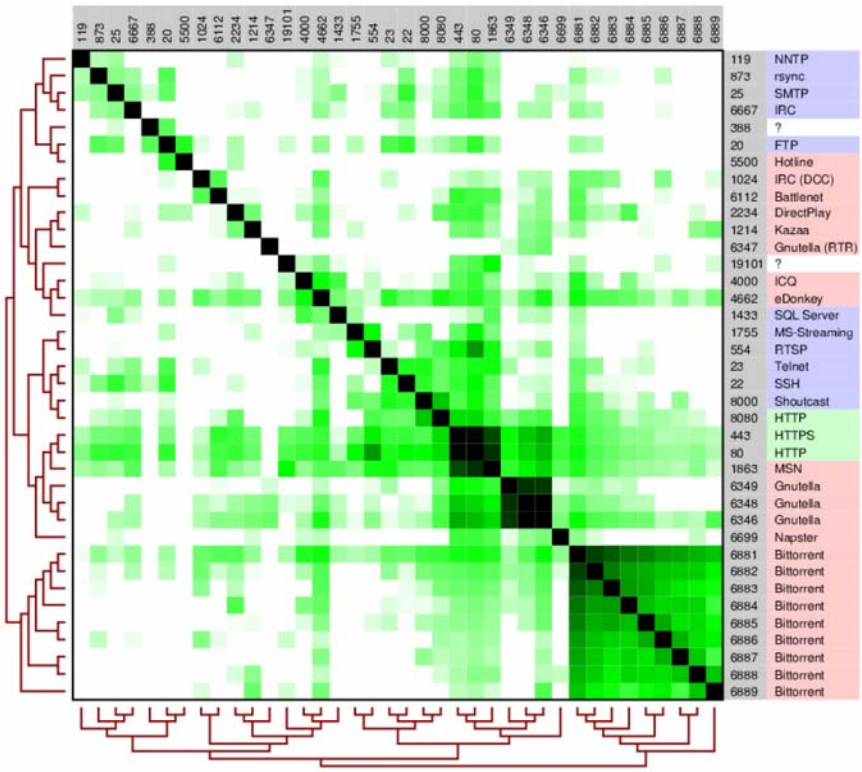


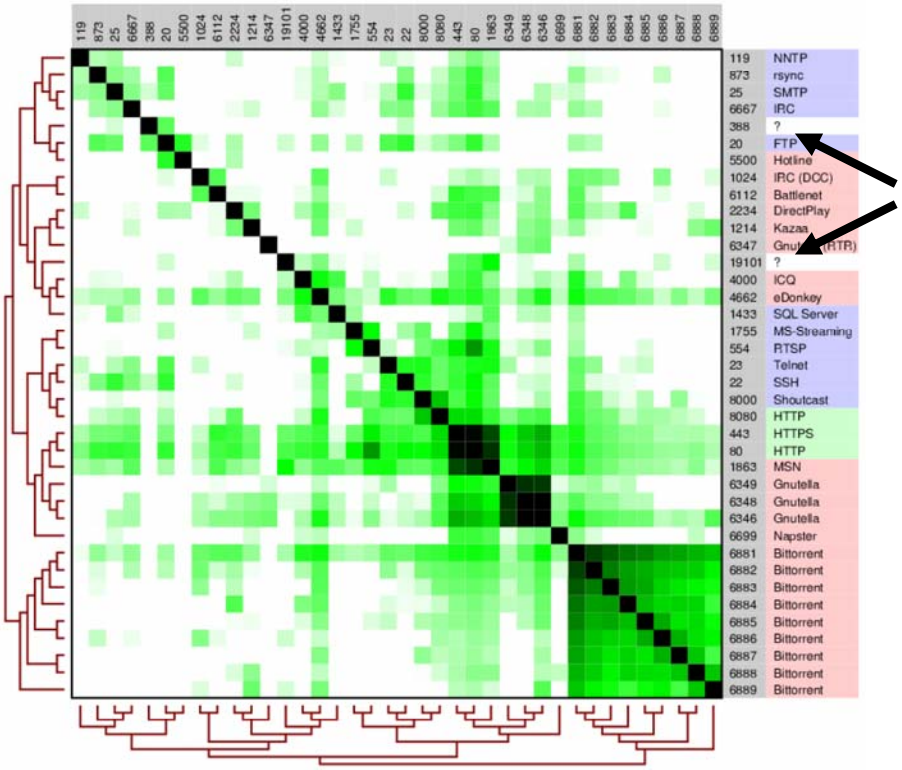
Multiple digraphs

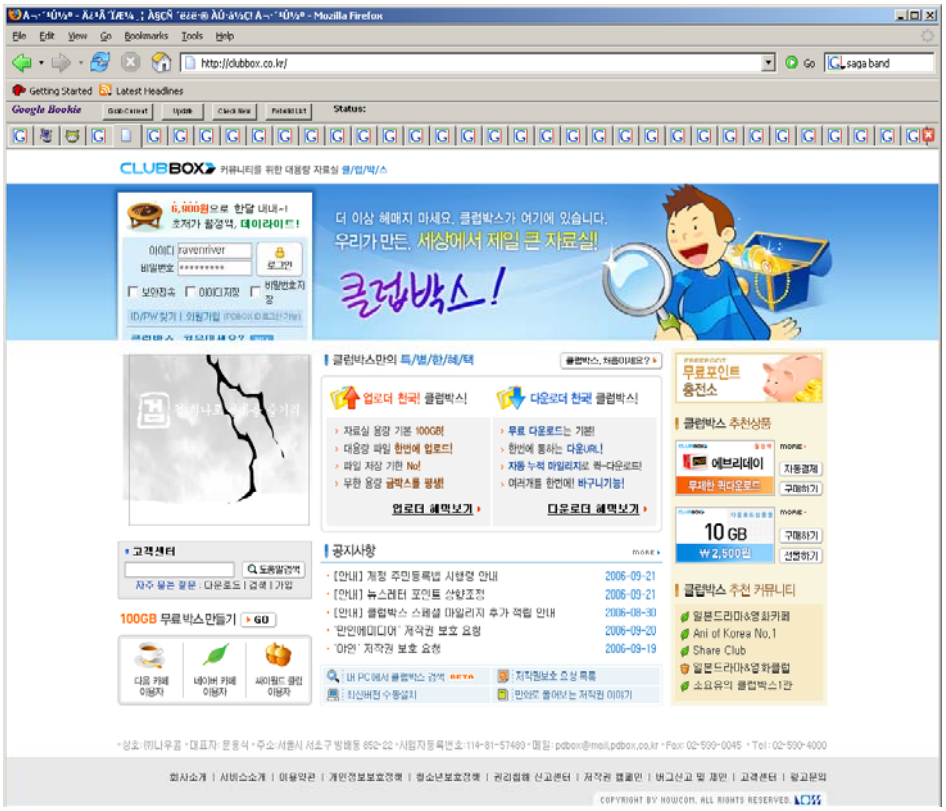





Combined
Web
P2P
Other
strength (s)
degree (k)


119 NNTP
873 rsync
25 SMTP
6667 IRC
388 ?
20 FTP
5500 Hotline
1024 IRC (DCC)
6112 Battlenet
2234 DirectPlay
1214 Kazaa
6347 Gnutella (RTR)
19101 ?
4000 ICQ
4662 eDonkey
1433 SQL Server
1755 MS-Streaming
554 RTSP
23 Telnet
22 SSH
8000 Shoutcast
8080 HTTP
443 HTTPS
80 HTTP
1863 MSN
6349 Gnutella
6348 Gnutella
6346 Gnutella
6699 Napster
6881 Bittorrent
6882 Bittorrent
6883 Bittorrent
6884 Bittorrent
6885 Bittorrent
6886 Bittorrent
6887 Bittorrent
6888 Bittorrent
6889 Bittorrent

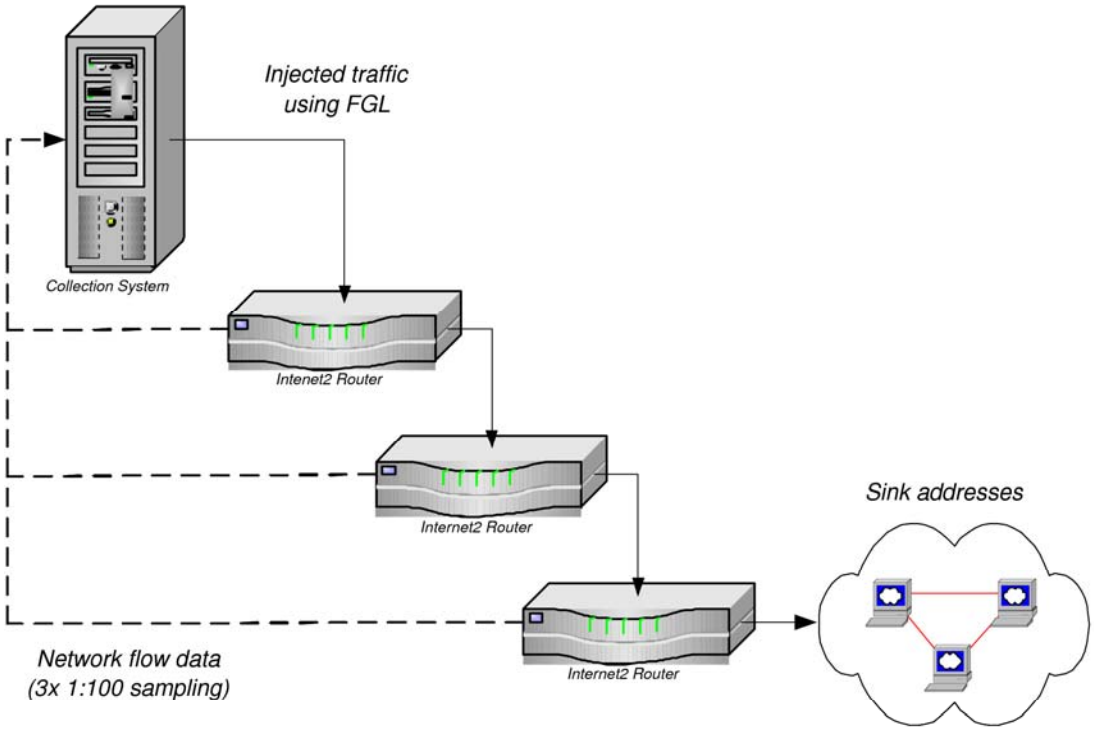

119 NNTP
873 rsync
25 SMTP
6667 IRC
388 ?
20 FTP
5500 Hotline
1024 IRC (DCC)
6112 Battlenet
2234 DirectPlay
1214 Kazaa
6347 Gnu (HTR)
19101 ?
4000 ICQ
4662 eDonkey
1433 SQL Server
1755 MS-Streaming
554 RTSP
23 Telnet
22 SSH
8000 Shoutcast
8080 HTTP
443 HTTPS
80 HTTP
1863 MSN
6349 Gnutella
6348 Gnutella
6346 Gnutella
6699 Napster
6881 Bittorrent
6882 Bittorrent
6883 Bittorrent
6884 Bittorrent
6885 Bittorrent
6886 Bittorrent
6887 Bittorrent
6888 Bittorrent
6889 Bittorrent

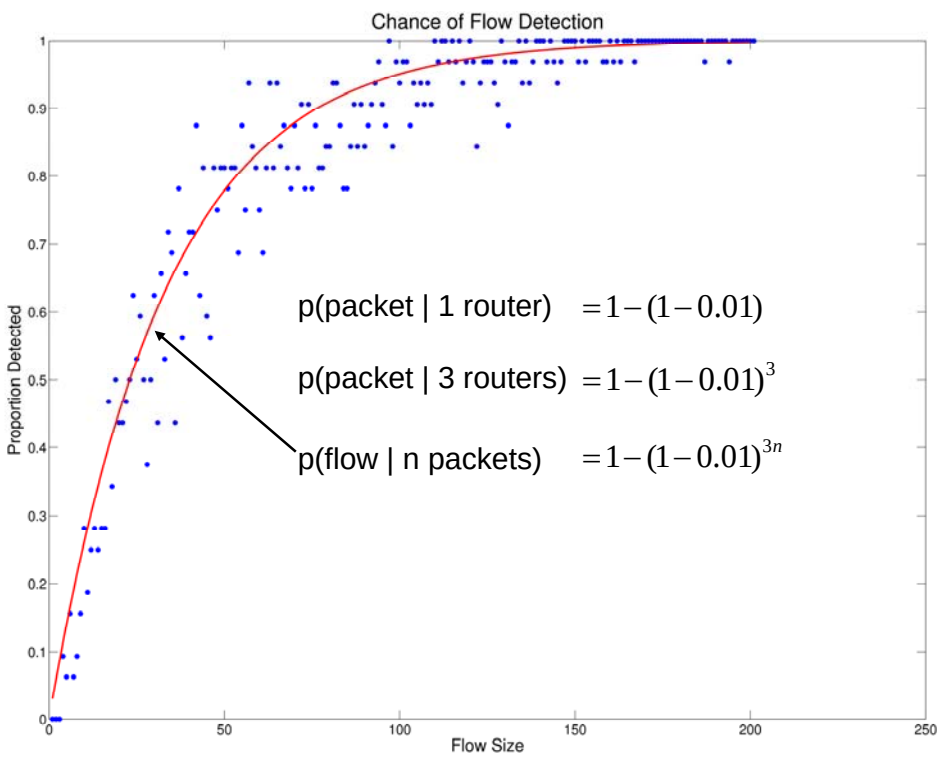

http://clubbox.co.kr/
CLUBBOX
더 이상 헤매지 마세요. 클럽박스가 여기에 있습니다.
우리가 만든, 세상에서 제일 큰 자료실!
클럽박스!
공지사항
COPYRIGHT BY NOWCOM. ALL RIGHTS RESERVED.

Port	Predicted	Actual	Match?
388	traditional file transfer	weather data transfer	yes
19101	P2P chat or file transfer	individual file shares	yes
9080	P2P with central index	team collaboration	yes
8090	Windows P2P w/ Web svc.	Weblog server	yes
5020	Windows P2P file transfer	BBFTP file transfer	partial
42899	P2P file sharing or trojan	*(unknown)*	unknown
8301	P2P file sharing or trojan	several trojans	partial
1025	trojan	many different trojans	yes
20000	P2P, probably BitTorrent	BitTorrent	yes
59174	P2P file sharing or trojan	*(unknown)*	unknown
20001	P2P file sharing or trojan	several trojans	partial
15002	P2P file sharing or trojan	biology collab. tool	partial
16881	P2P, probably BitTorrent	BitTorrent	yes
9000	P2P file sharing or trojan	several trojans	partial
3124	Windows P2P file transfer	Web proxy (Windows)	yes
39281	P2P file sharing or trojan	grid-based computing	partial

Where do flows come from?

- Architectural features of Internet routers allow them to export flow data
- Routers can't summarize all the data
  - Packets are sampled to construct the flows
  - Typical sampling rate is around 1:100





p(packet | 1 router) $= 1-(1-0.01)$

p(packet | 3 routers) $= 1-(1-0.01)^3$

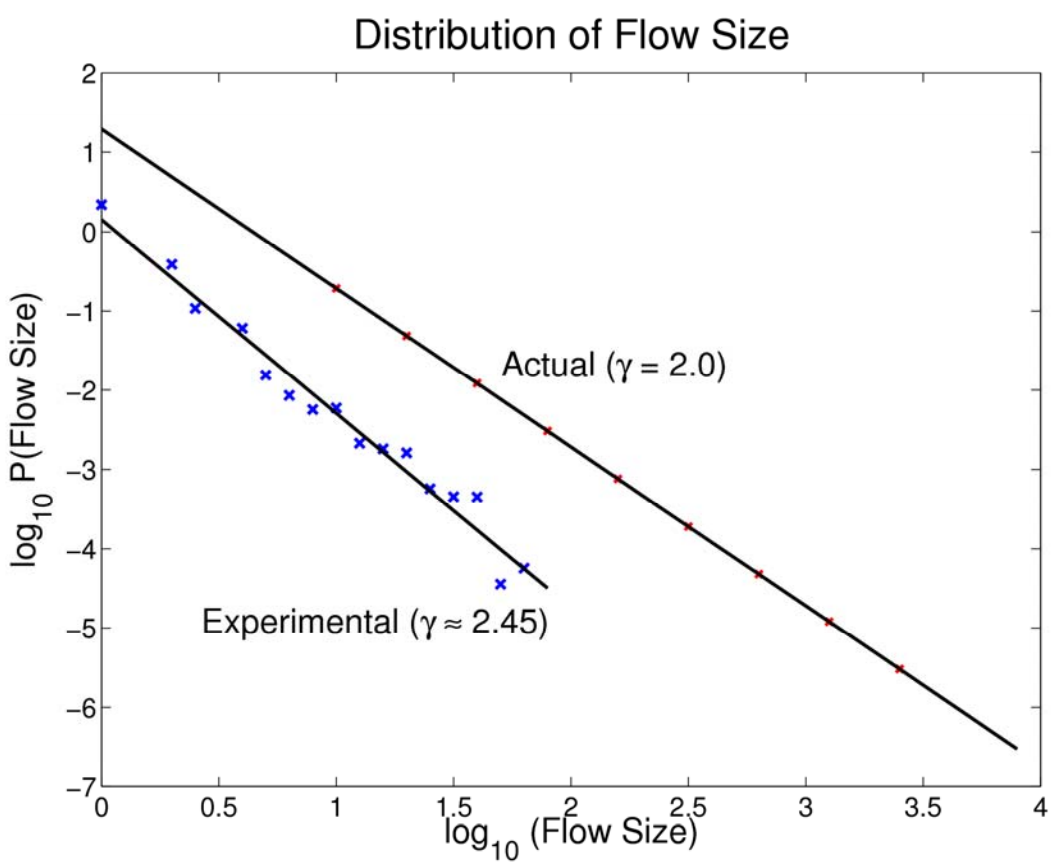
p(flow | n packets) $= 1-(1-0.01)^{3n}$

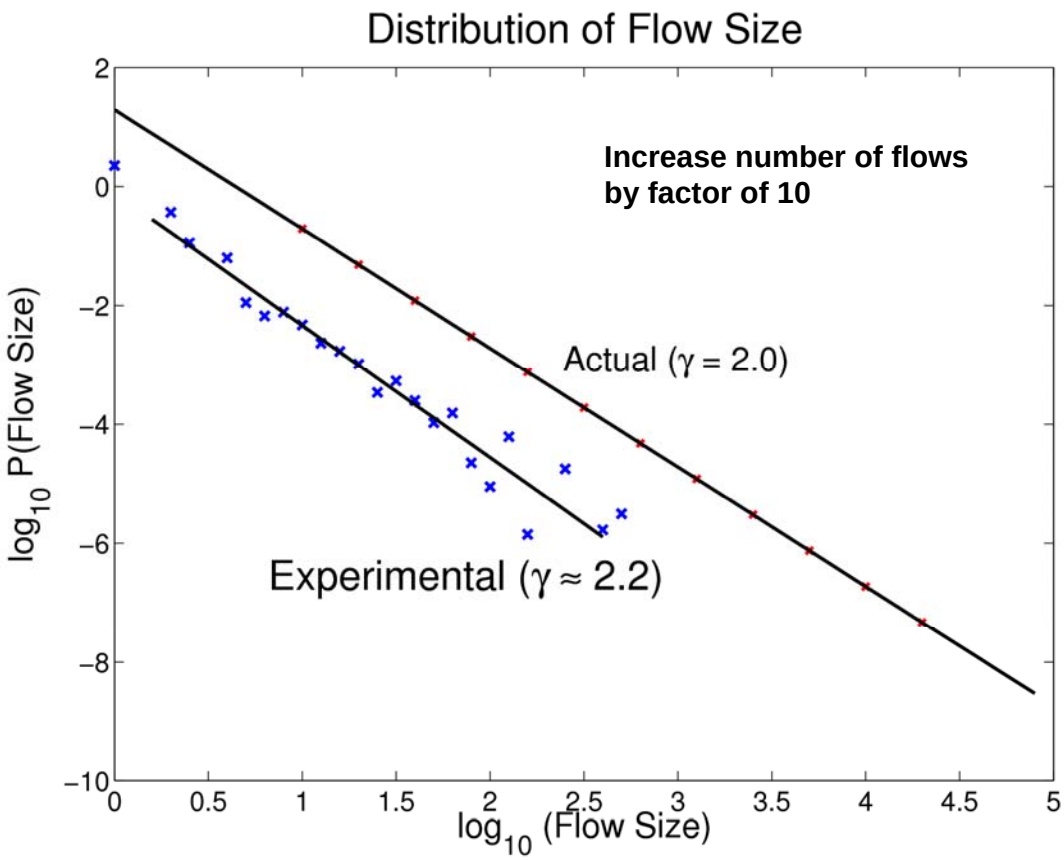
Distribution Recovery

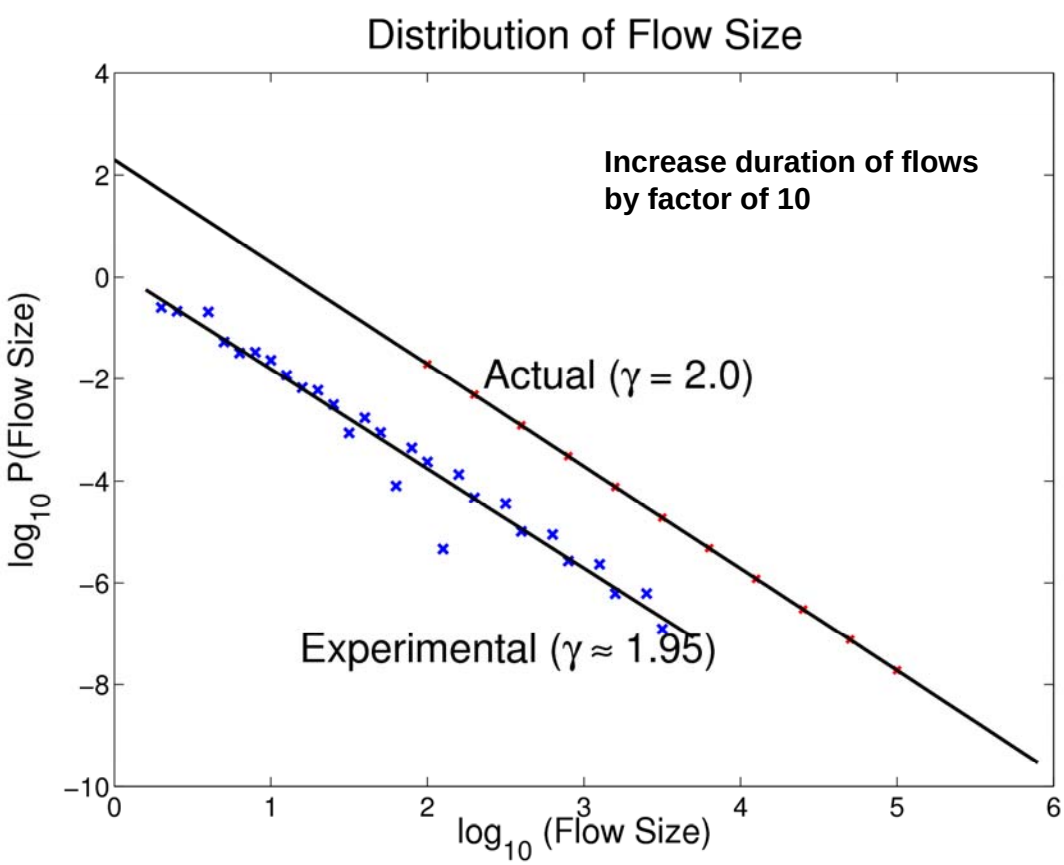
Try to recover a power law, exponent = 2.

Send to each of 10 hosts:

- 256 10-packet flows
- 128 20-packet flows
- 64 40-packet flows
- (etc.)




Distribution of Flow Size
Increase number of flows
by factor of 10
Actual (γ = 2.0)
Experimental (γ ≈ 2.2)
log10 P(Flow Size)
log10 (Flow Size)
2
0
−2
−4
−6
−8
−10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


Distribution of Flow Size
Increase duration of flows
by factor of 10
Actual (γ = 2.0)
Experimental (γ ≈ 1.95)
log10 P(Flow Size)
log10 (Flow Size)
4
2
0
−2
−4
−6
−8
−10
0
1
2
3
4
5
6

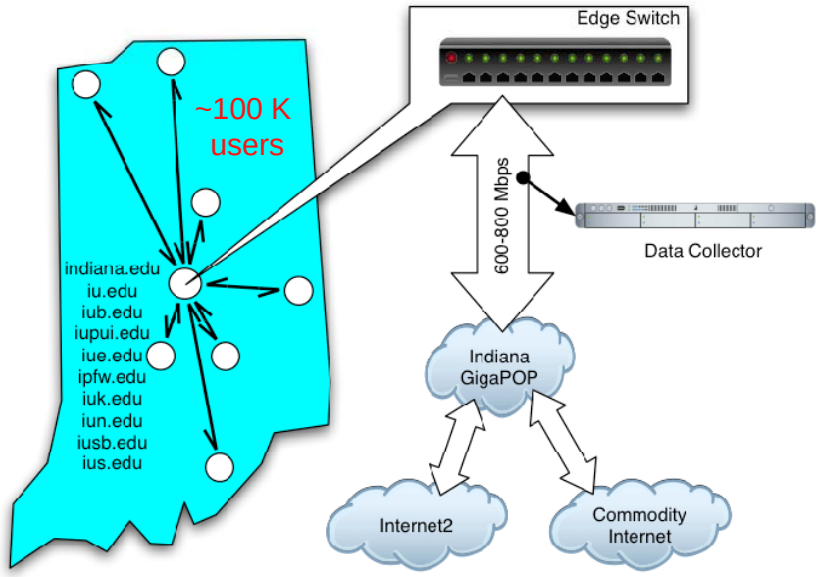
Results

- Nonlinear chance of flow detection.
- Very small flows lead to an overestimate of the exponent.
- With large flows, a range of exponents can be recovered reliably.
- *Aggregation* is necessary for accurate results.

Roadmap

- Background
- Network flow analysis
- Web click analysis
- Conclusions

Source of Click Data



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*

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***

We have a Web request

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

Web Requests

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

from this client

- *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*

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*

going from this URL

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***

to this one

Web Requests

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

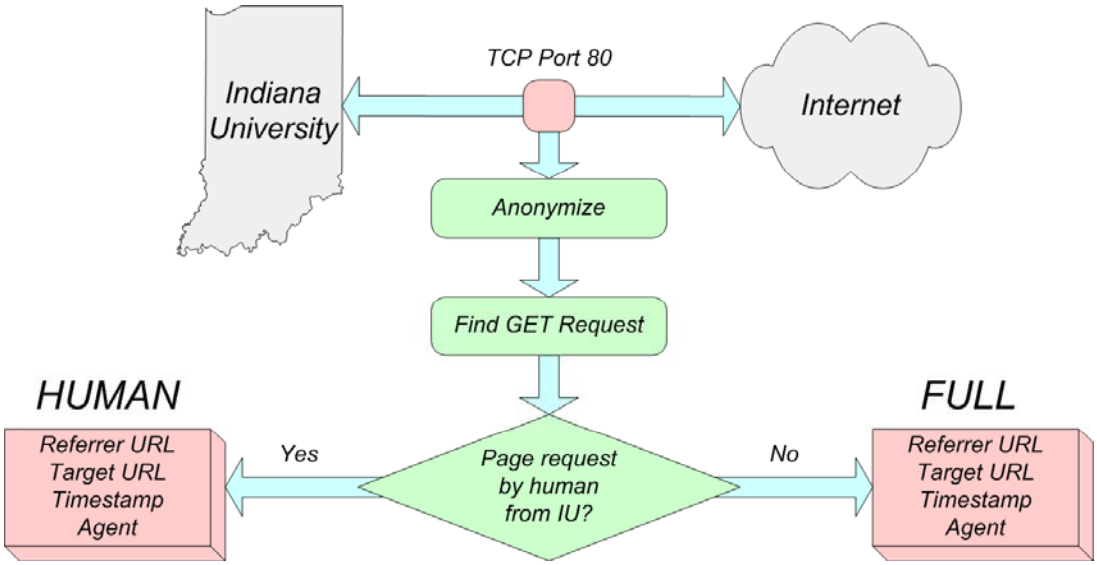
- *Source IP: 192.168.39.190*
- *Dest. IP: 127.100.251.3*

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

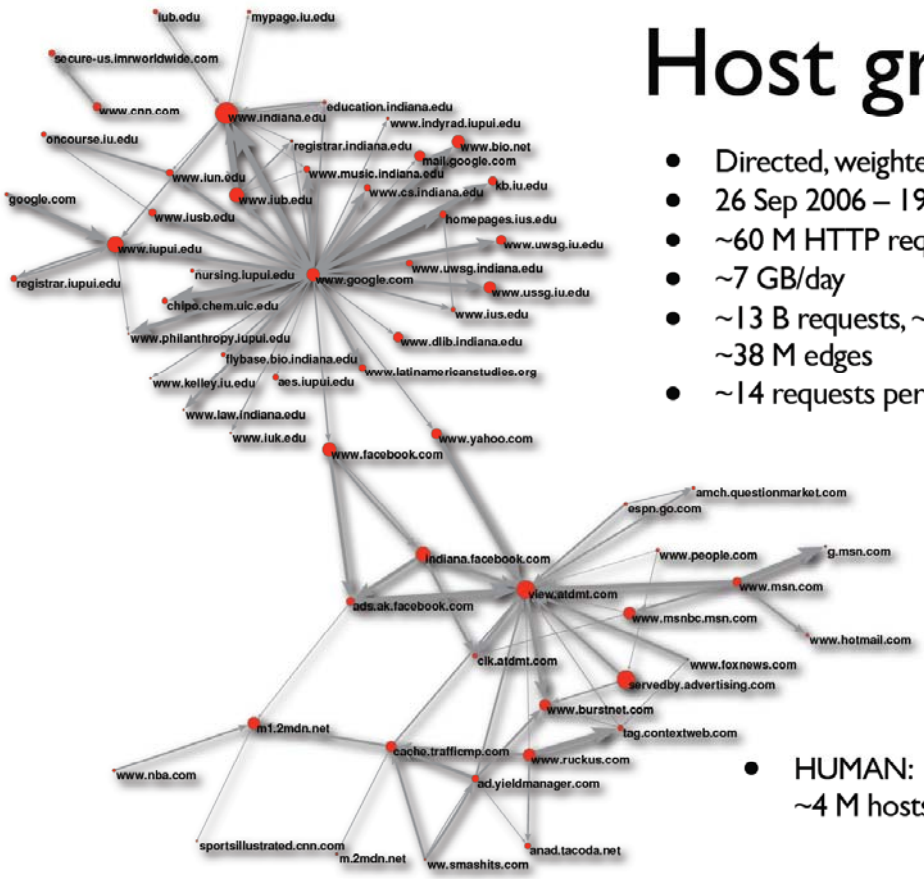
using this agent

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

Click Collection



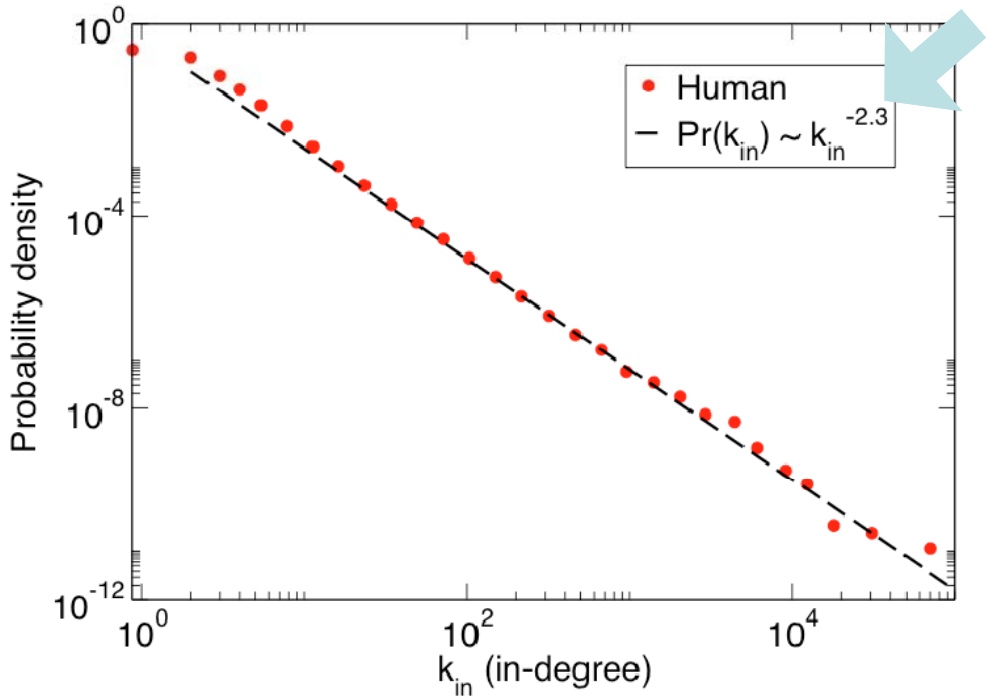
Host graphs



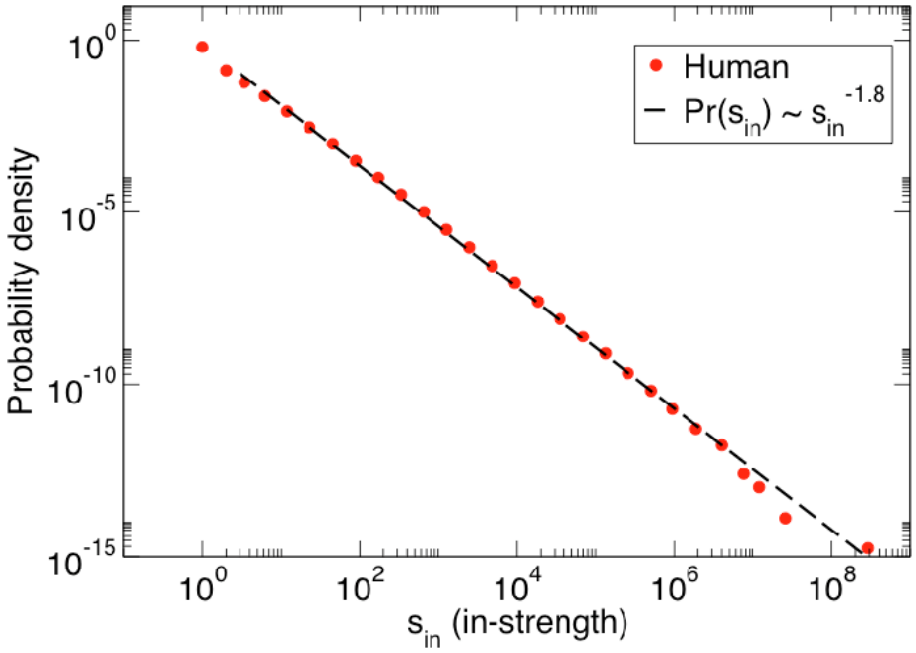
- Directed, weighted networks
- 26 Sep 2006 – 19 May 2007
- ~60 M HTTP requests per day
- ~7 GB/day
- ~13 B requests, ~8 M hosts, ~38 M edges
- ~14 requests per human click

- HUMAN: ~1 B requests, ~4 M hosts, ~11 M edges

Structural properties: *Degree (Link Count)*



Structural properties: *Strength (Site Traffic)*

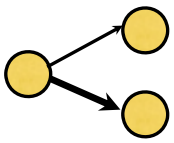


Evaluation of PageRank

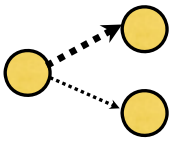
- PR is a ***stationary distribution of visit frequency*** by a modified random walk
- Compare with ***actual site traffic*** (in-strength)
- From an application perspective, we care about the resulting ***ranking of sites***

PageRank Assumptions

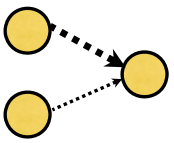
1. Equal probability of following each link from any given node



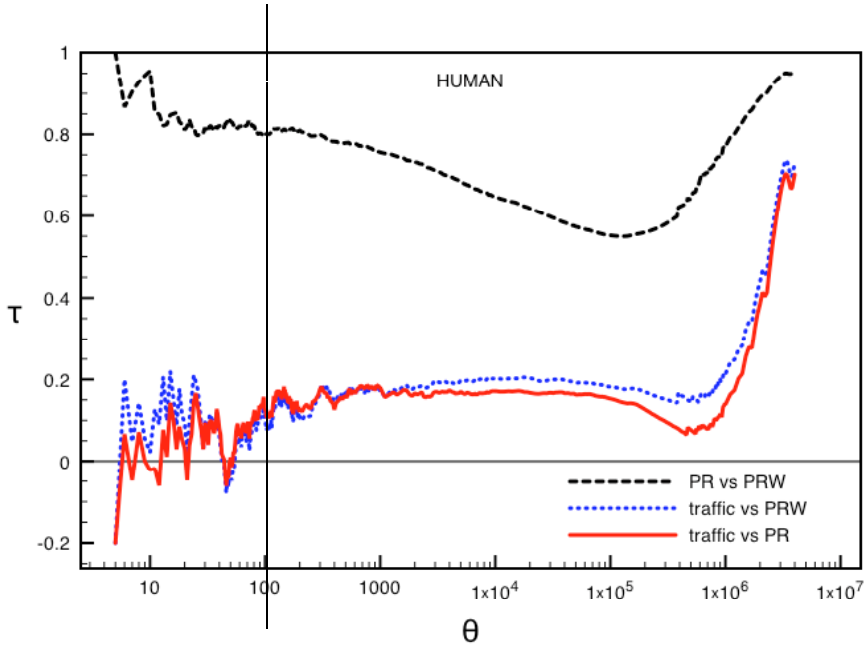
2. Equal probability of teleporting ***to*** each of the nodes



3. Equal probability of teleporting ***from*** each of the nodes



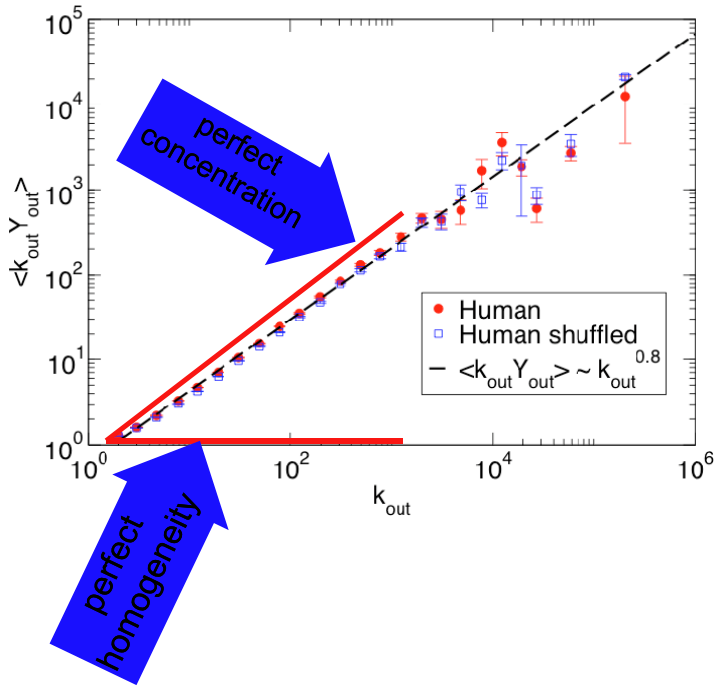
#1: Kendall's Rank Correlation



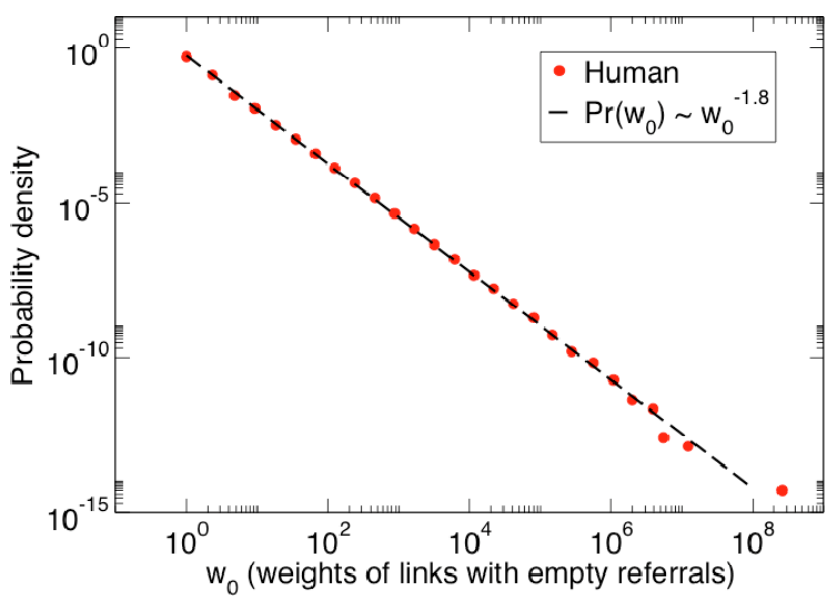
Local Link Heterogeneity

$$Y_i = \sum_j \left(\frac{w_{ij}}{s_{out}(i)} \right)^2$$

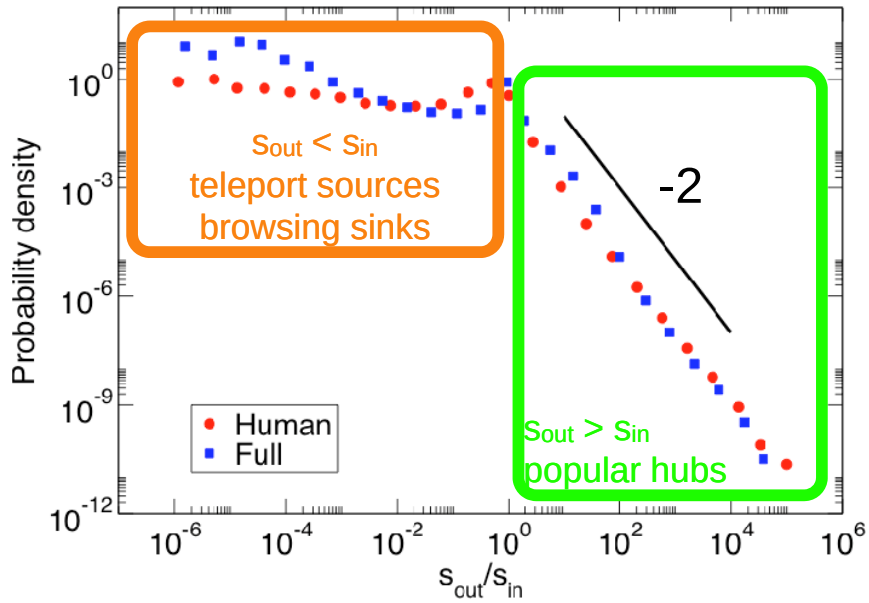
HH Index of concentration or disparity



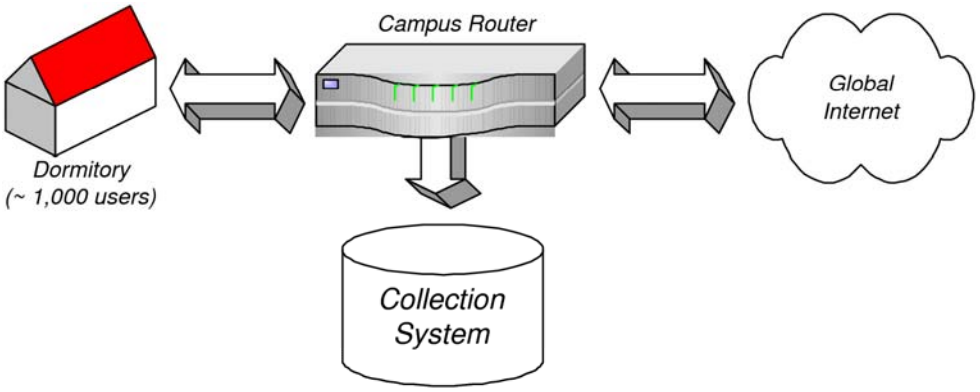
#2: Teleportation Target Heterogeneity



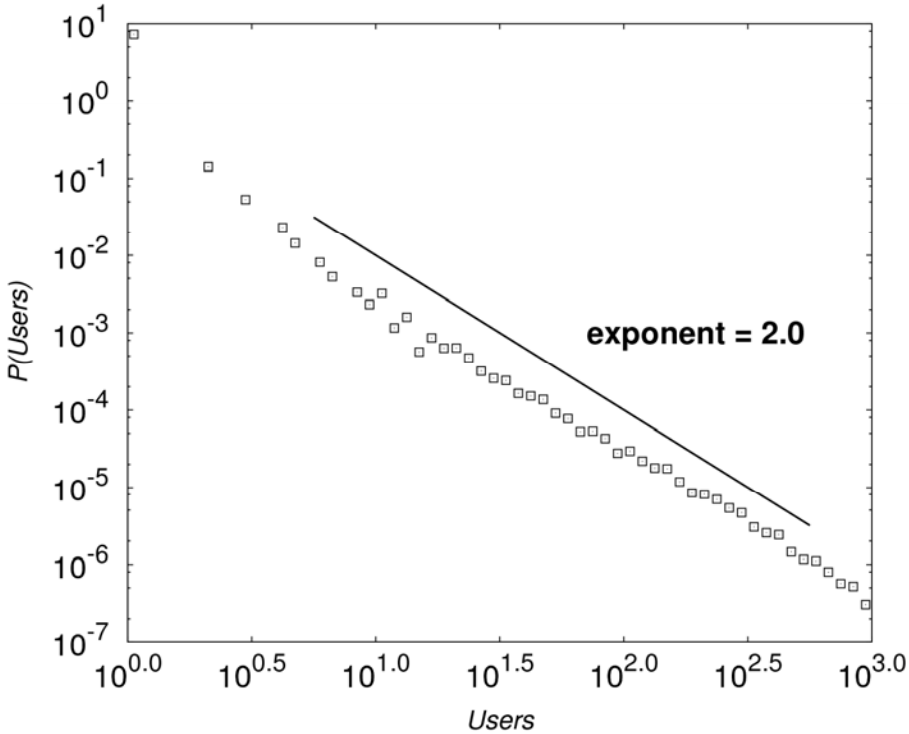
#3: Teleportation Source Heterogeneity *("hubness")*



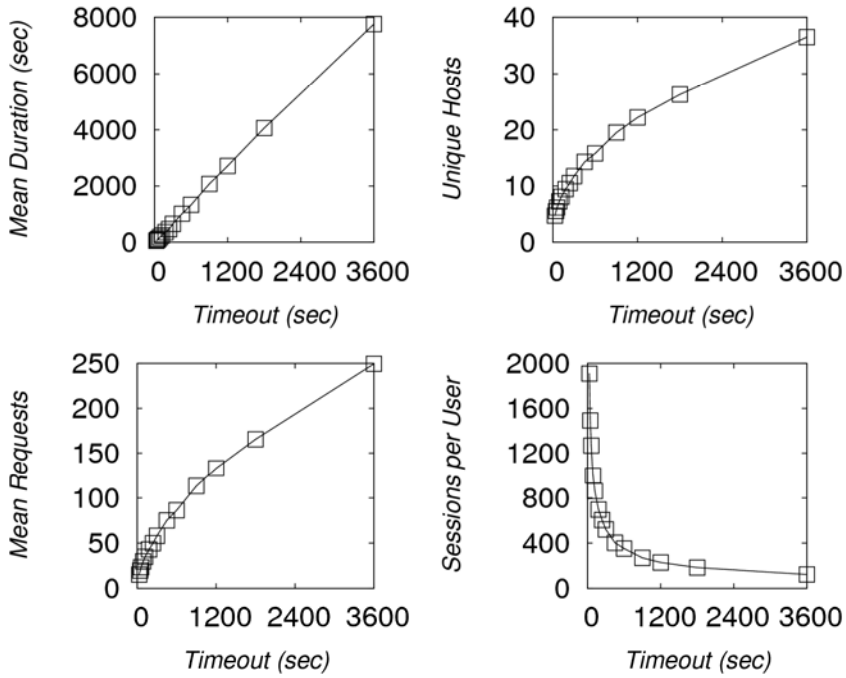
Click data with retention of user identity



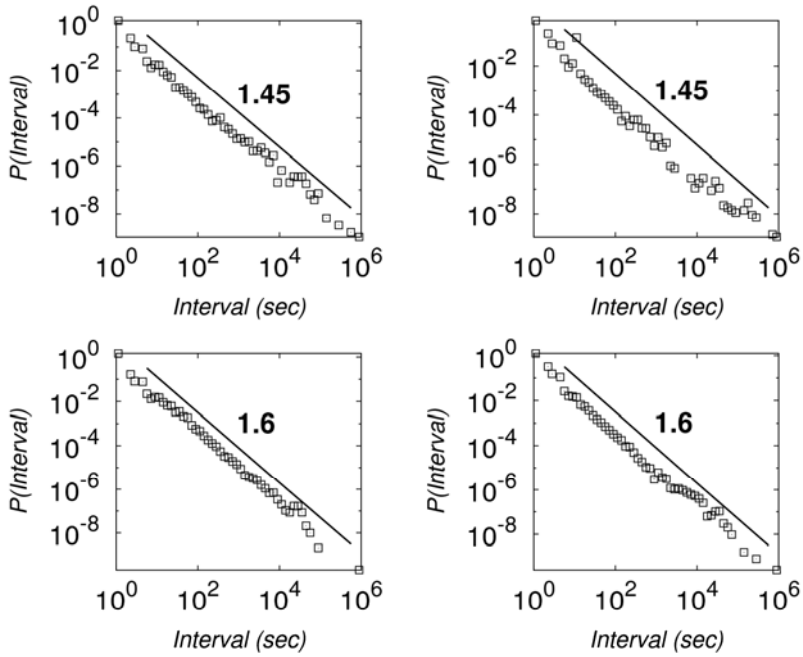
Popularity is unbounded!



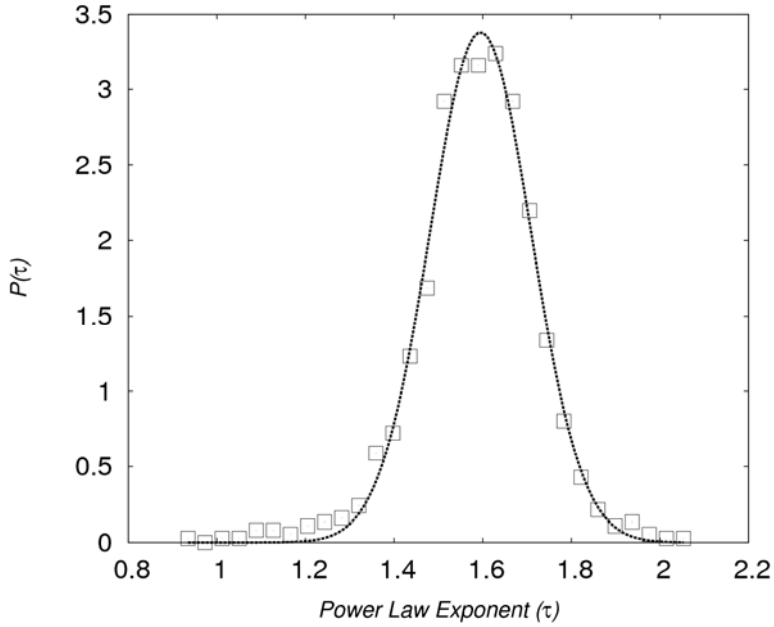
Session properties depend on timeout.

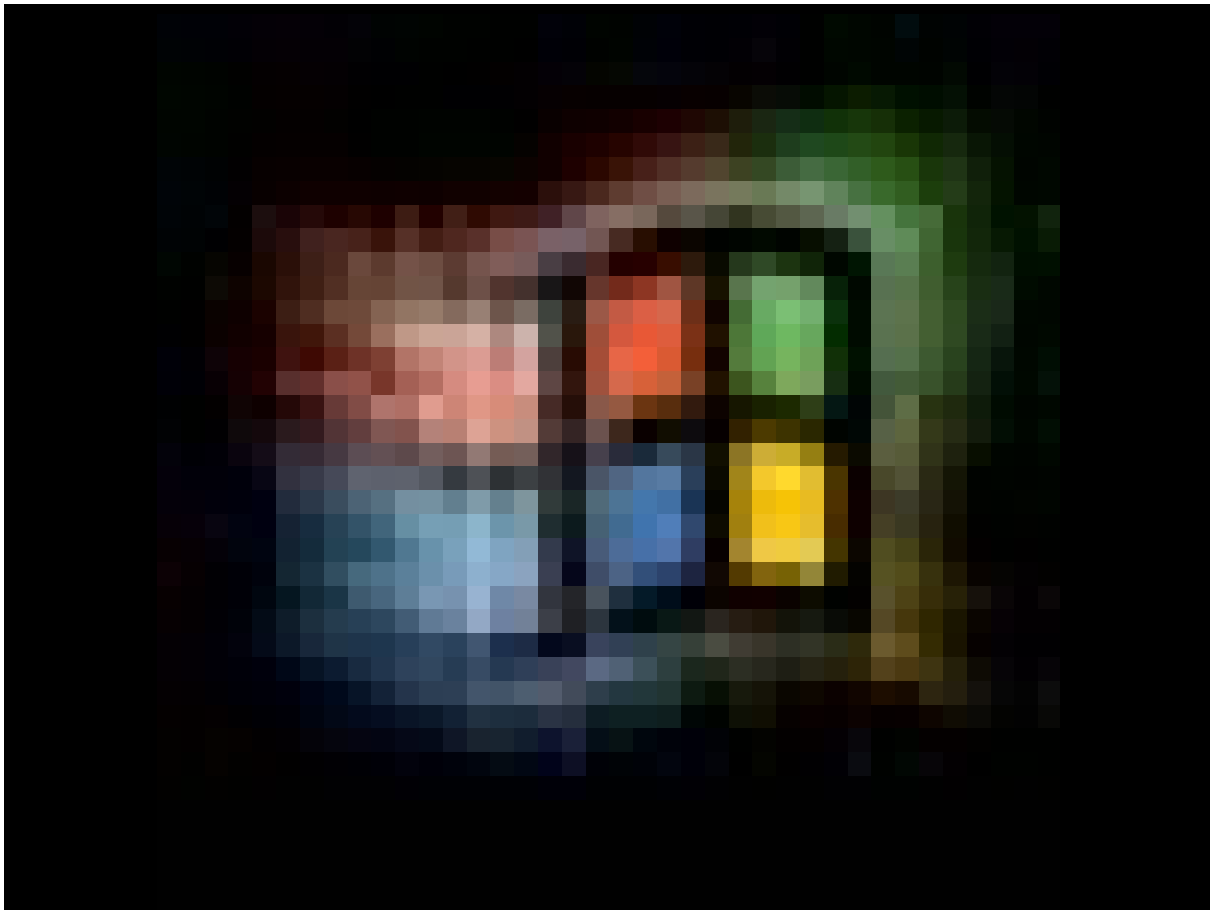


Where's the "sweet spot" ?

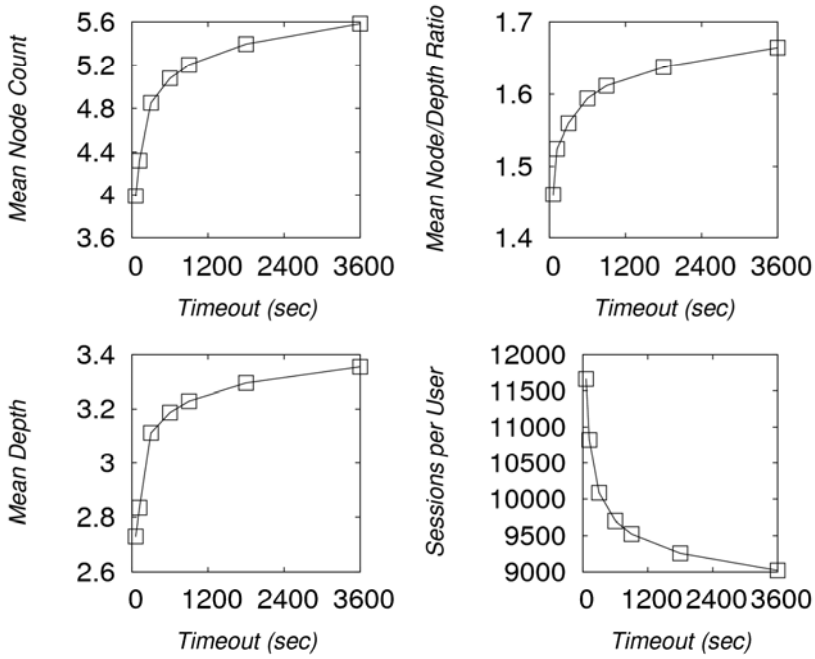


All users are abnormal.

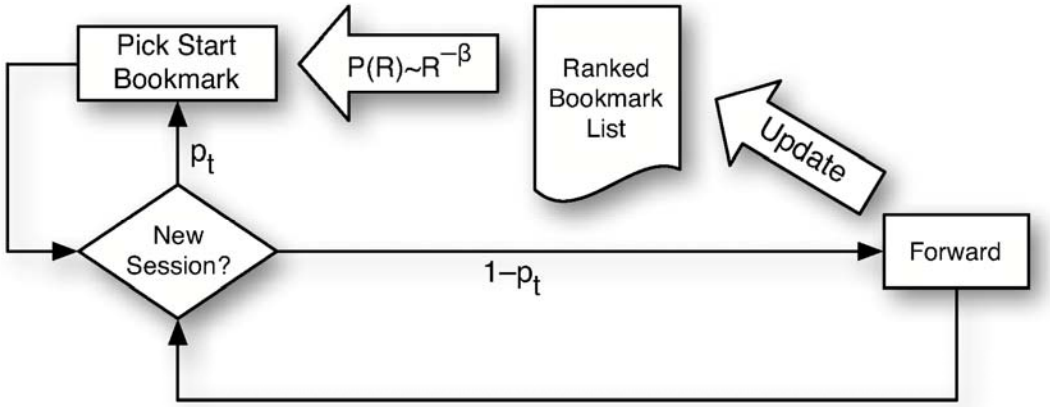


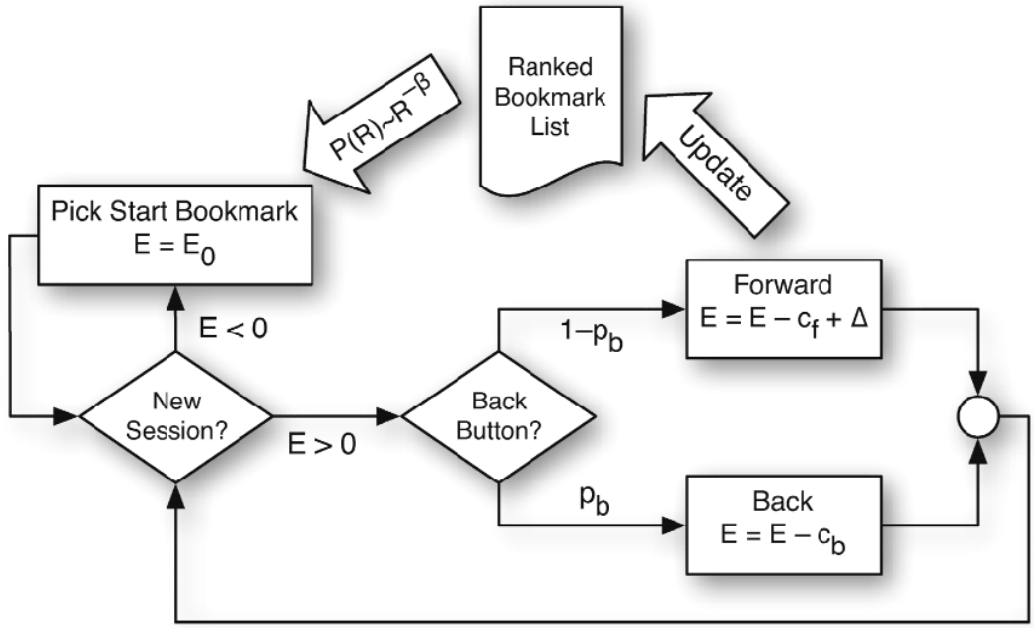


Timeout dependence is much weaker.

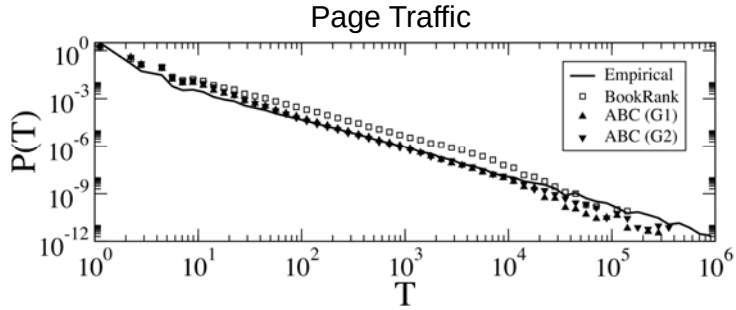



Teleport to random page
p_t
New Session?
$1-p_t$
Forward

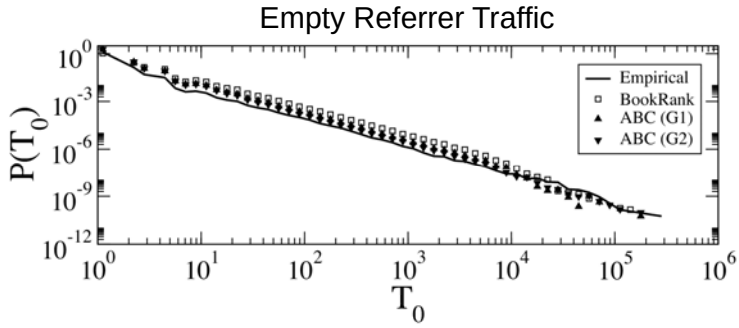

Pick Start Bookmark
$P(R) \sim R^{-\beta}$
Ranked Bookmark List
Update
p_t
New Session?
$1-p_t$
Forward



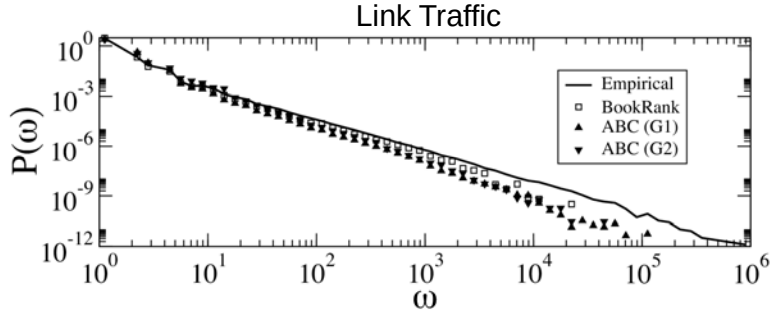
Page Traffic



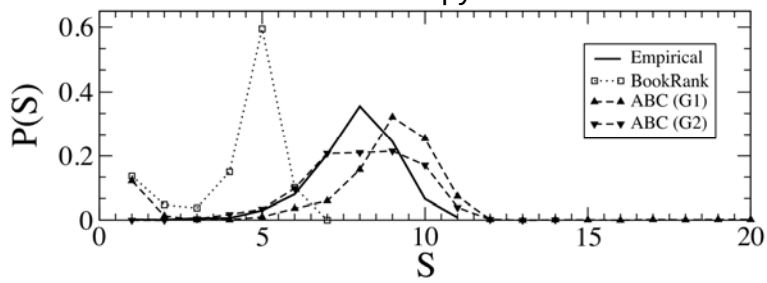
Empty Referrer Traffic



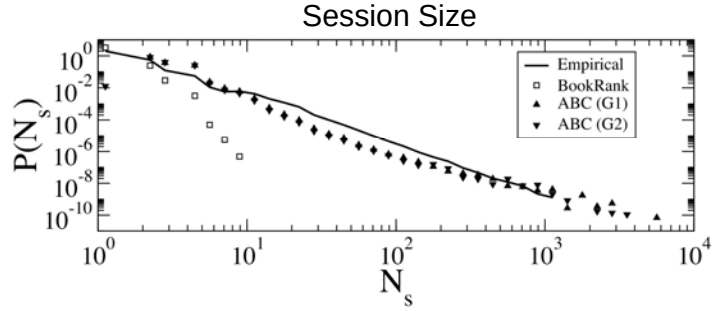
Link Traffic



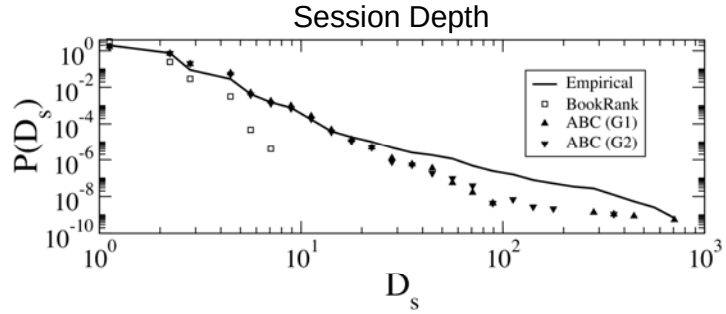
Entropy

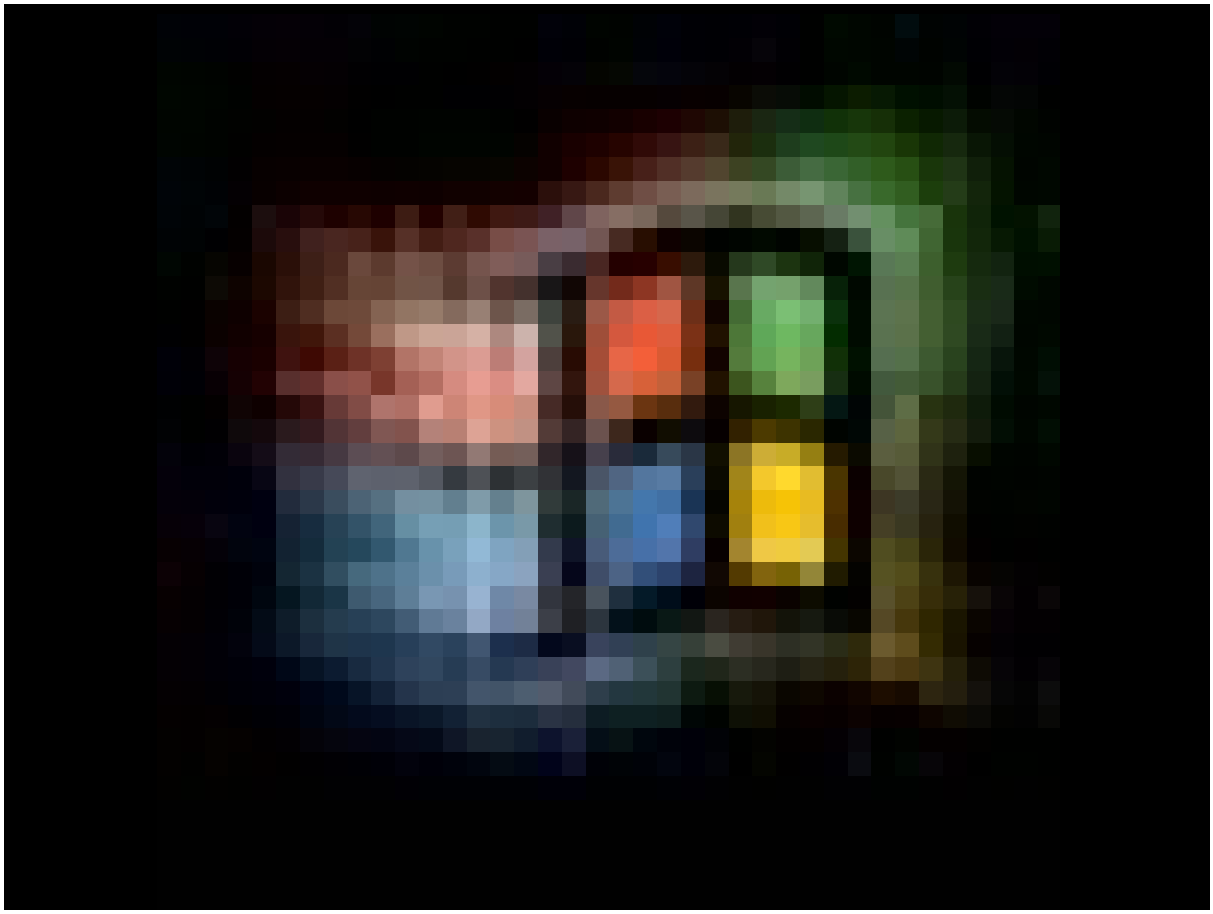


Session Size



Session Depth





Roadmap

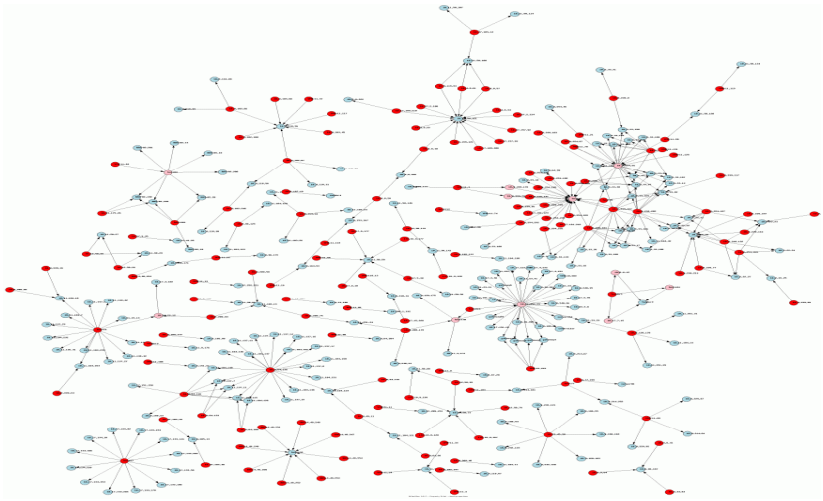
- Background
- Network flow analysis
- Web click analysis
- Conclusions

Conclusions

- Internet behavior is characterized by extreme heterogeneity.

Conclusions

- Behavioral analysis offers advantages over packet inspection.



Conclusions

- We observe heterogeneity because users are *idiosyncratic*, not pathologically eclectic.

Future Directions

- Relationship between traffic & substrate
- Community detection
- Characterization of links
- Validation of HITS
- Time-series analysis

THANKS!

- Filippo Menczer
- Alessandro Vespigani
- Katy Börner
- Minaxi Gupta
- Kay Connelly
- Steven Wallace
- Gregory Travis
- David Ripley
- Edward Balas
- Camillo Viecco
- Jean Camp
- J. Duncan
- Alessandro Flammini
- Santo Fortunato
- Bruno Gonçalves
- José Ramasco
- Damon Beals
- Dave Hershberger
- John Stigall
- Heather Roinestad

Questions & Comments