The Web Click Network

Mark Meiss

Advanced Network Management Lab
Pervasive Technology Labs at Indiana University

Filippo Menczer

Santo Fortunato

Alessandro Flammini

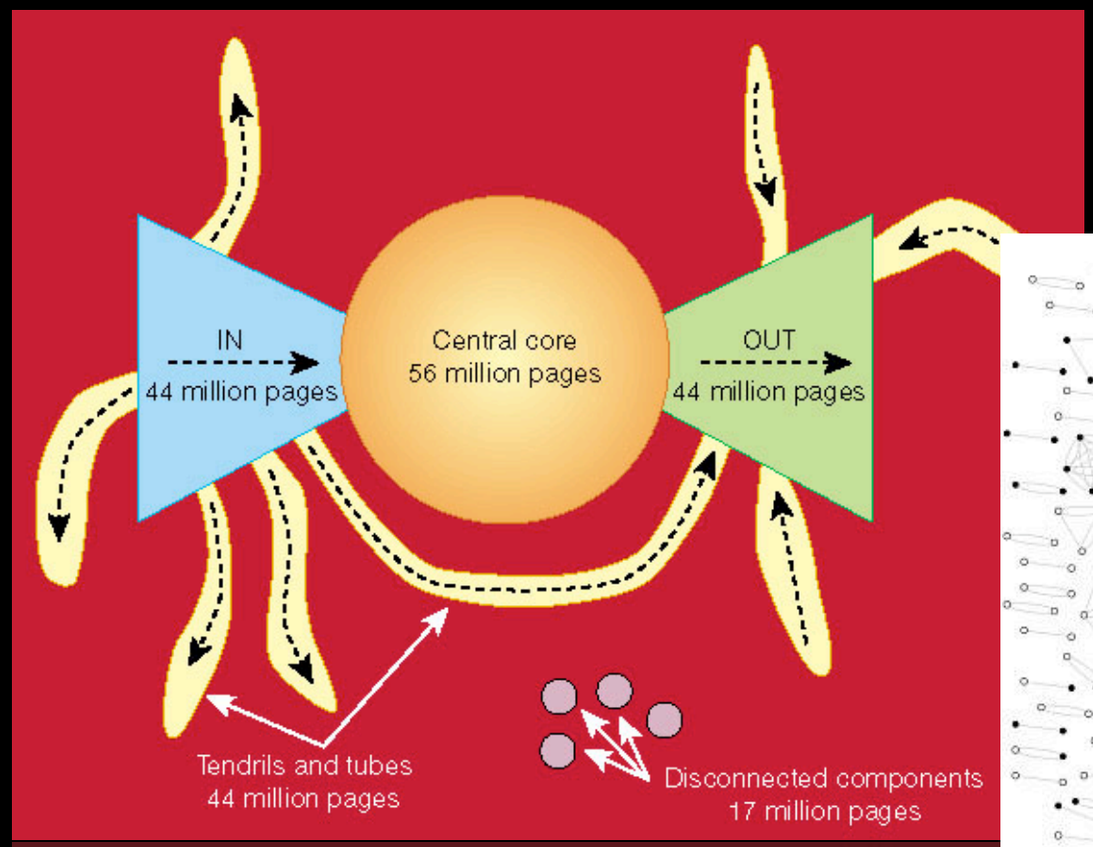
Alessandro Vespignani

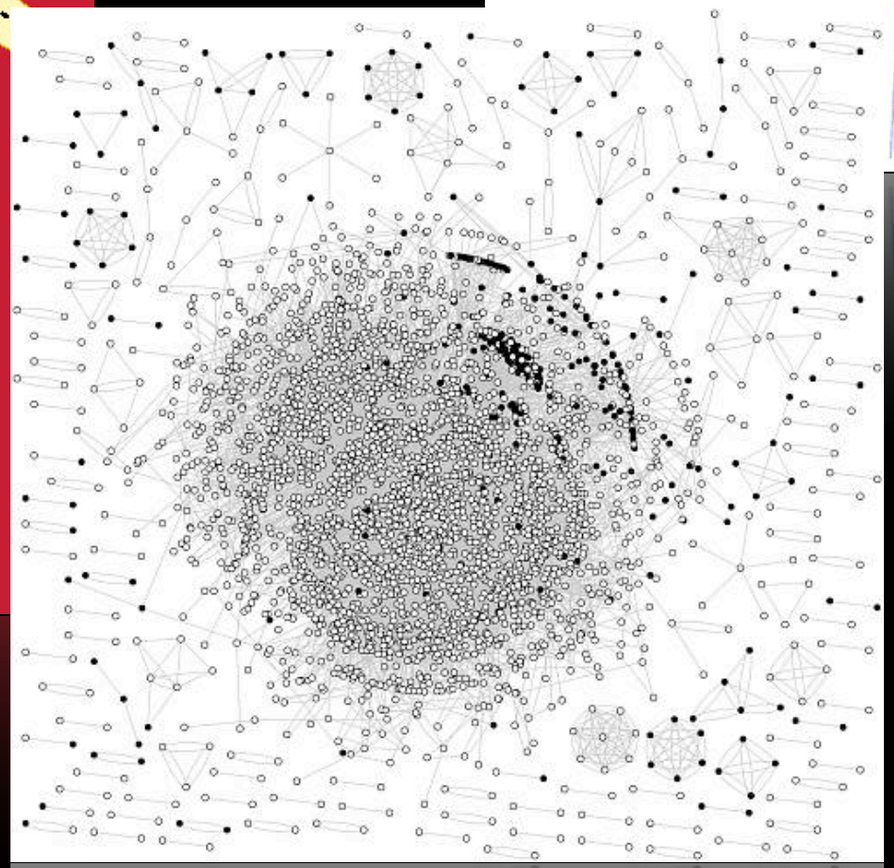
Indiana University School of informatics

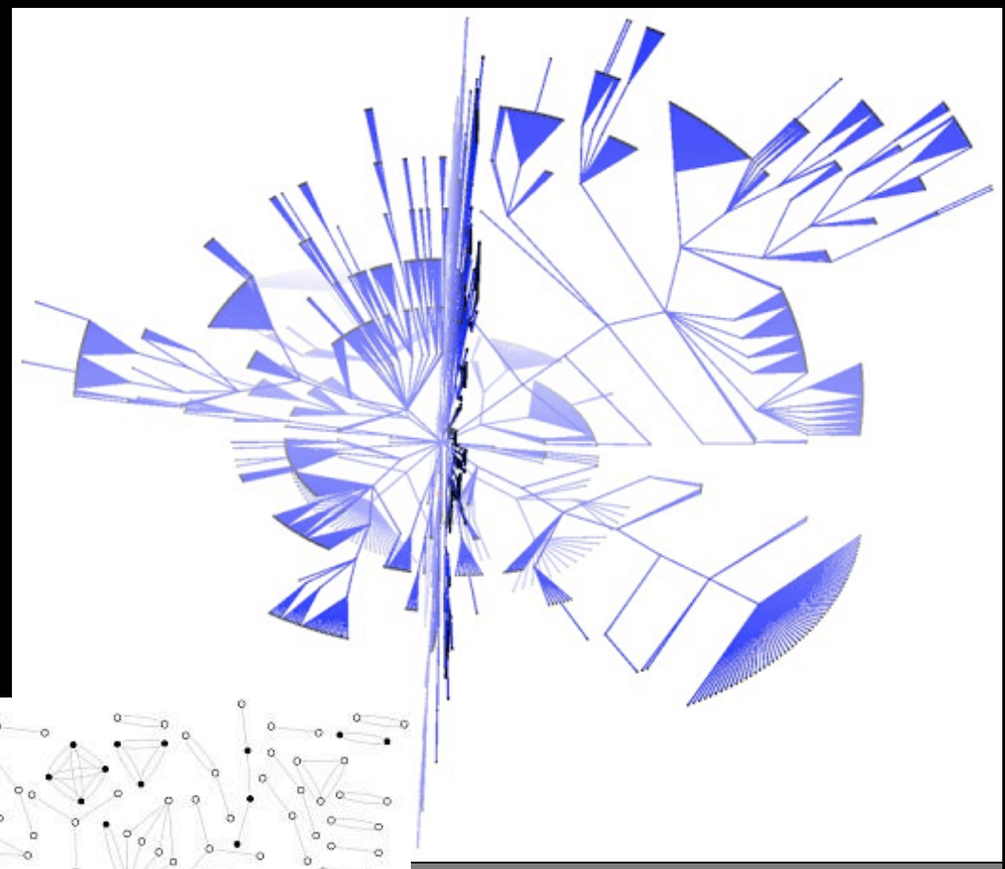
CNLL
Progetto Lagrange

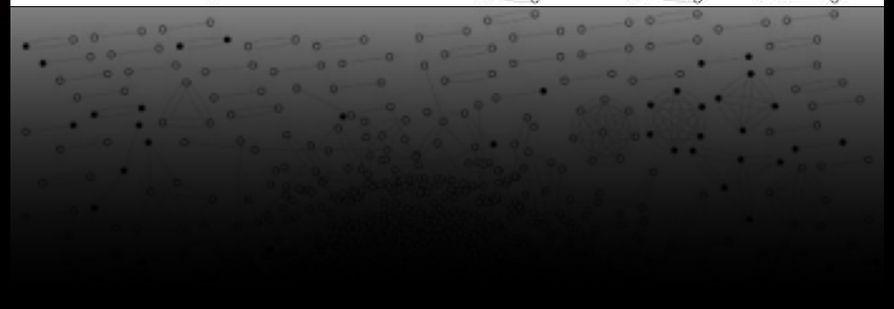
INSTITUTE FOR SCIENTIFIC INTERCHANGE FOUNDATION

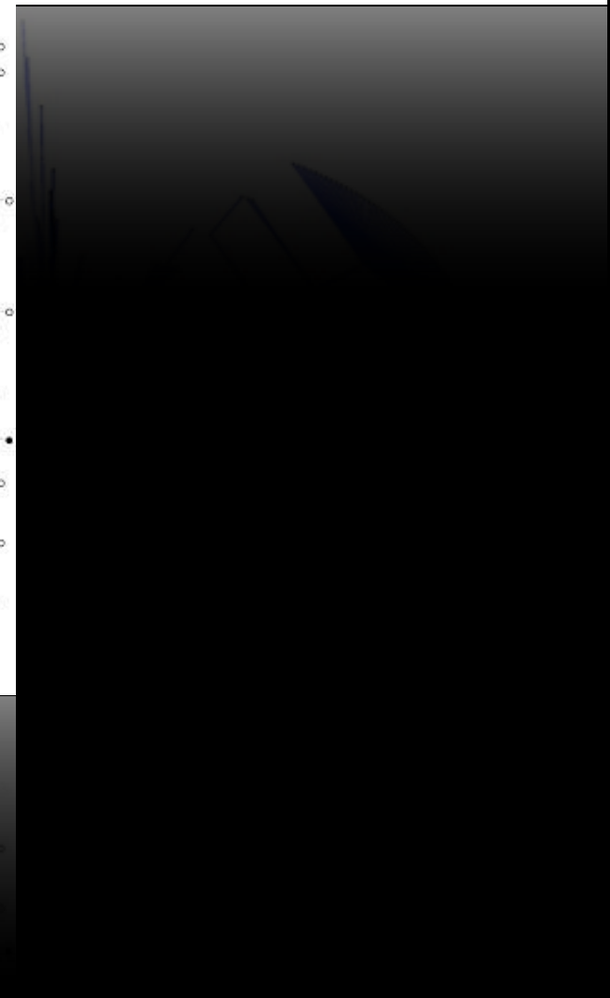
Static Web link graph



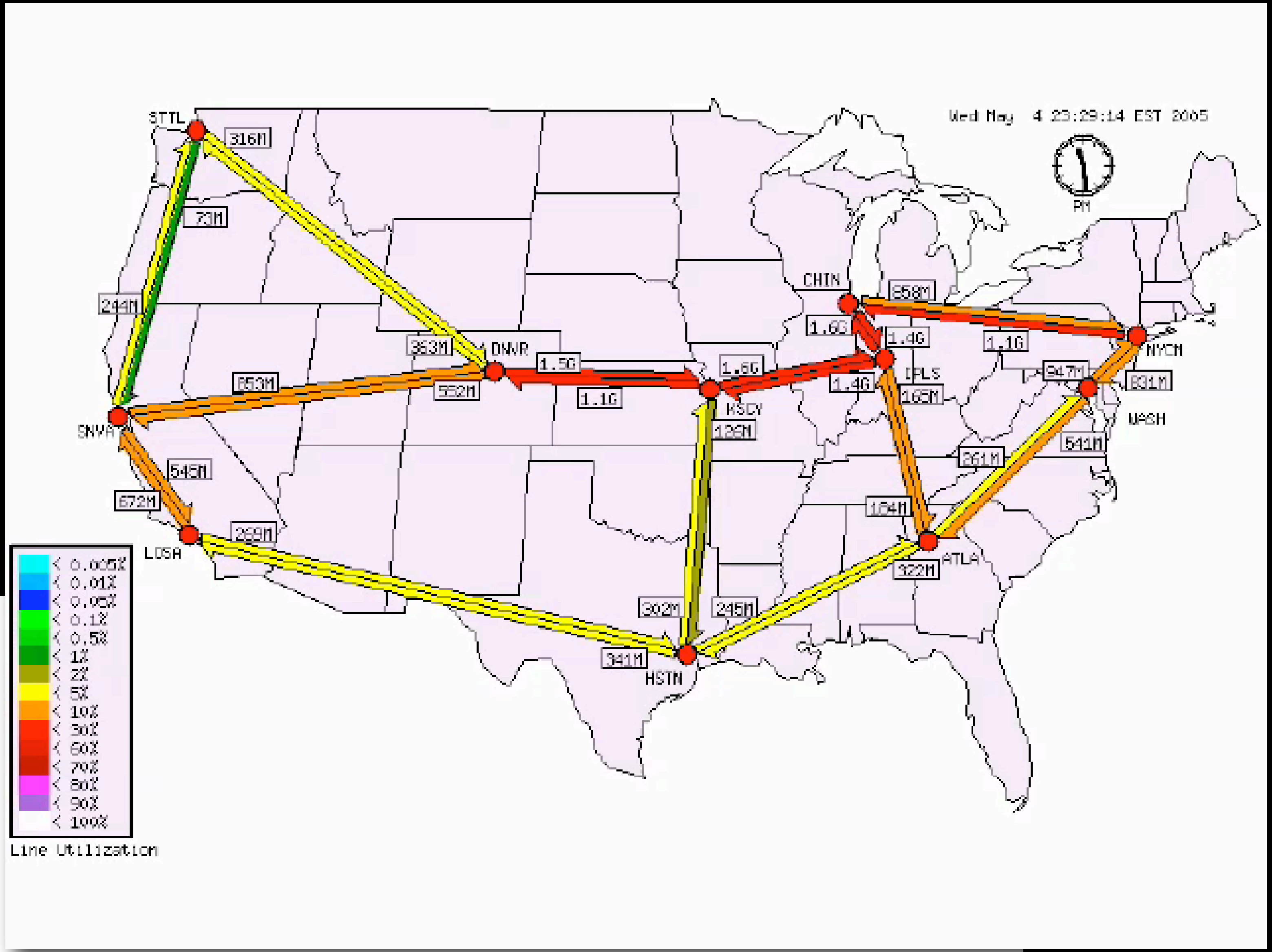


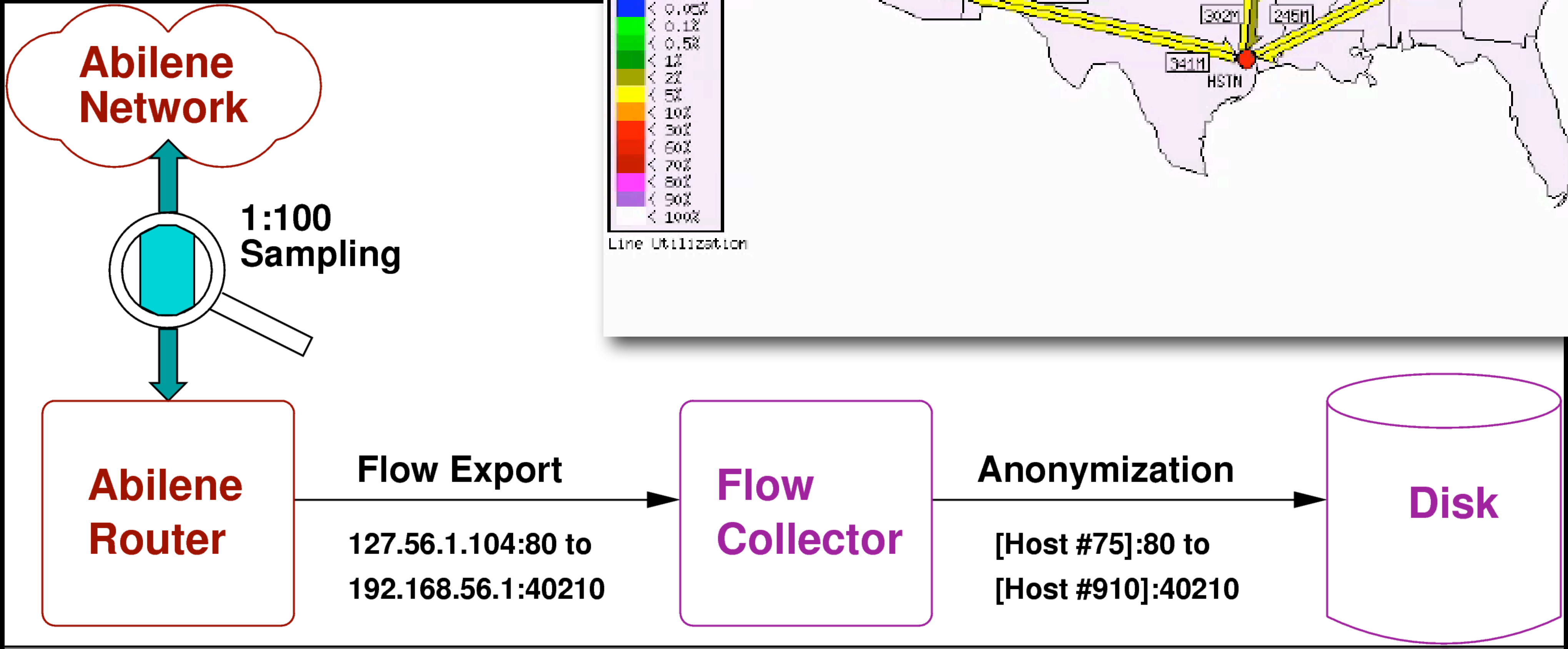




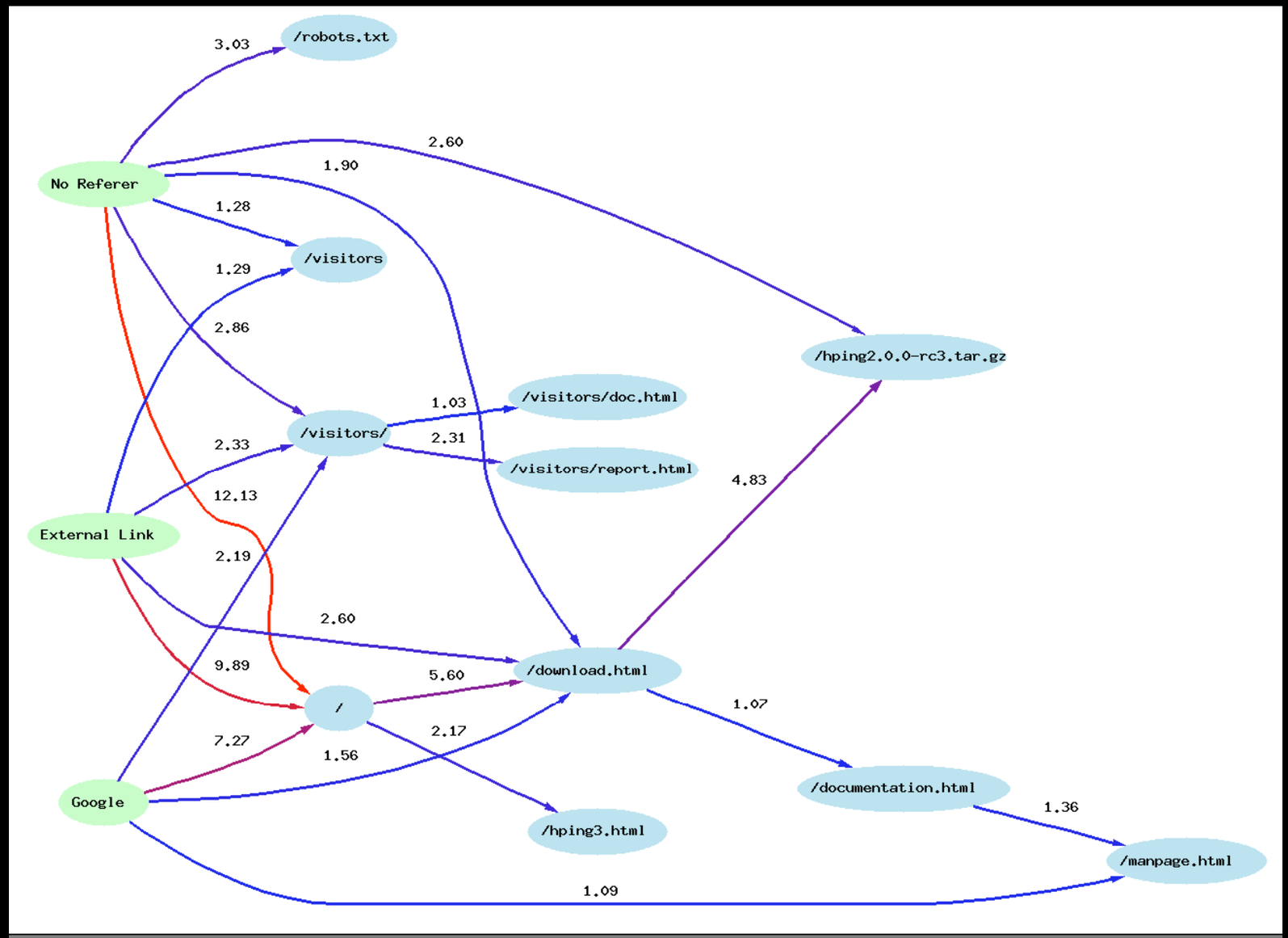


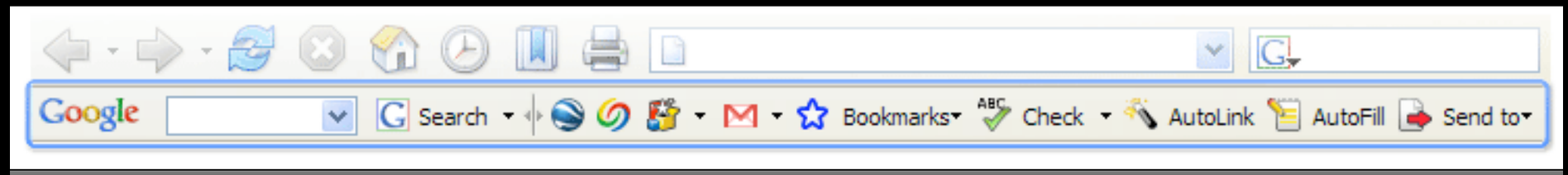
NetFlow


Wed May 4 23:29:14 EST 2005
STTL
DNVR
SNVA
LOSA
HSTN
KSCY
CHIN
IPLS
NYCM
WASH
ATLA
Line Utilization

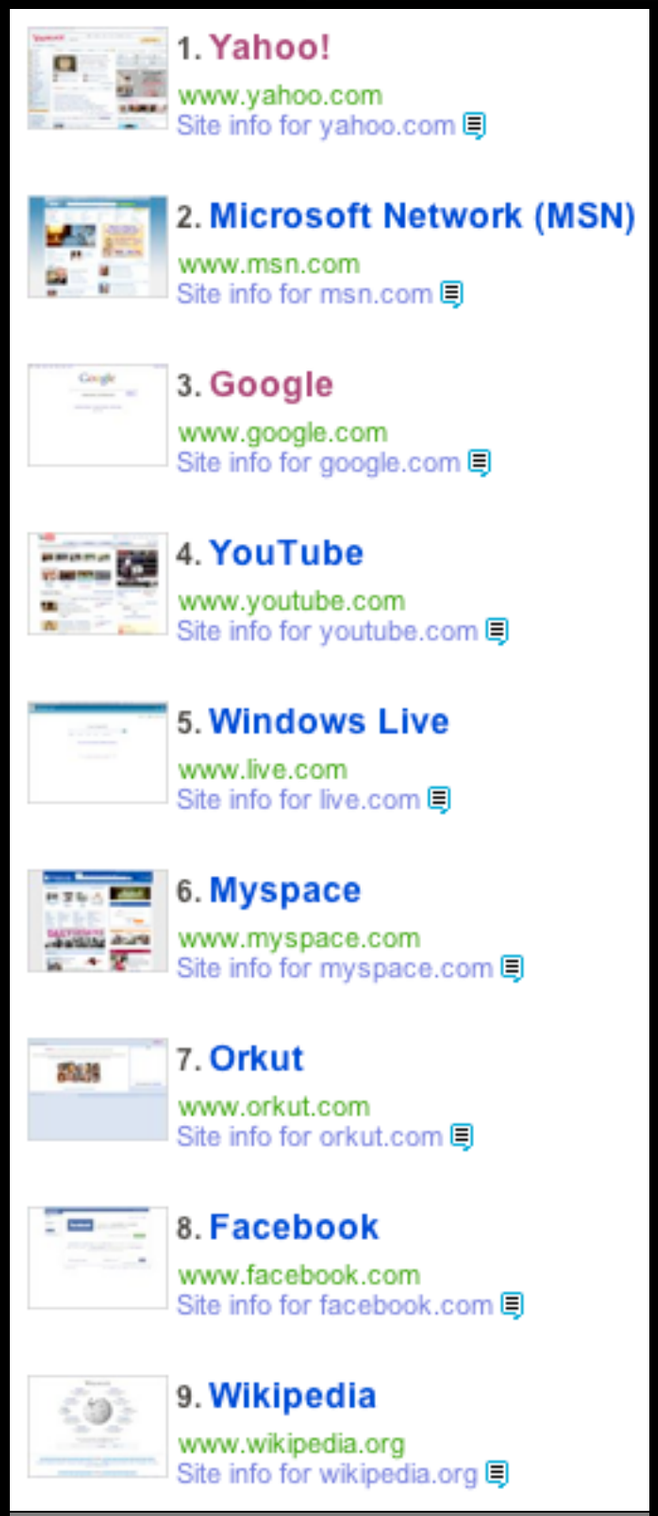

Abilene Network
1:100 Sampling
Abilene Router
Flow Export
127.56.1.104:80 to
192.168.56.1:40210
Flow Collector
Anonymization
[Host #75]:80 to
[Host #910]:40210
Disk

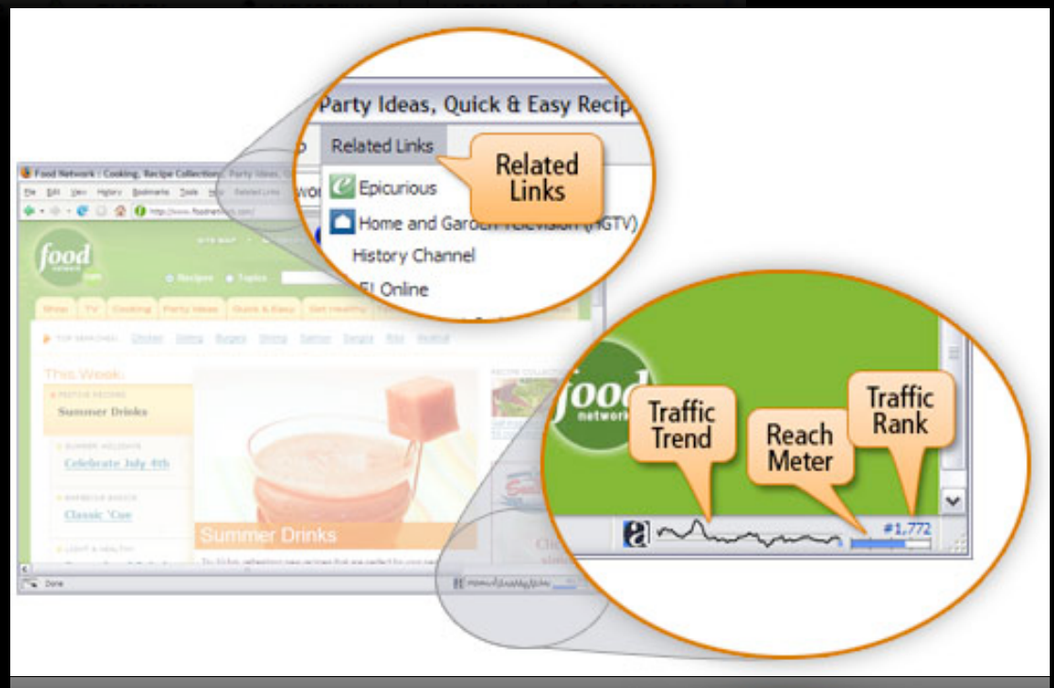
Web server logs

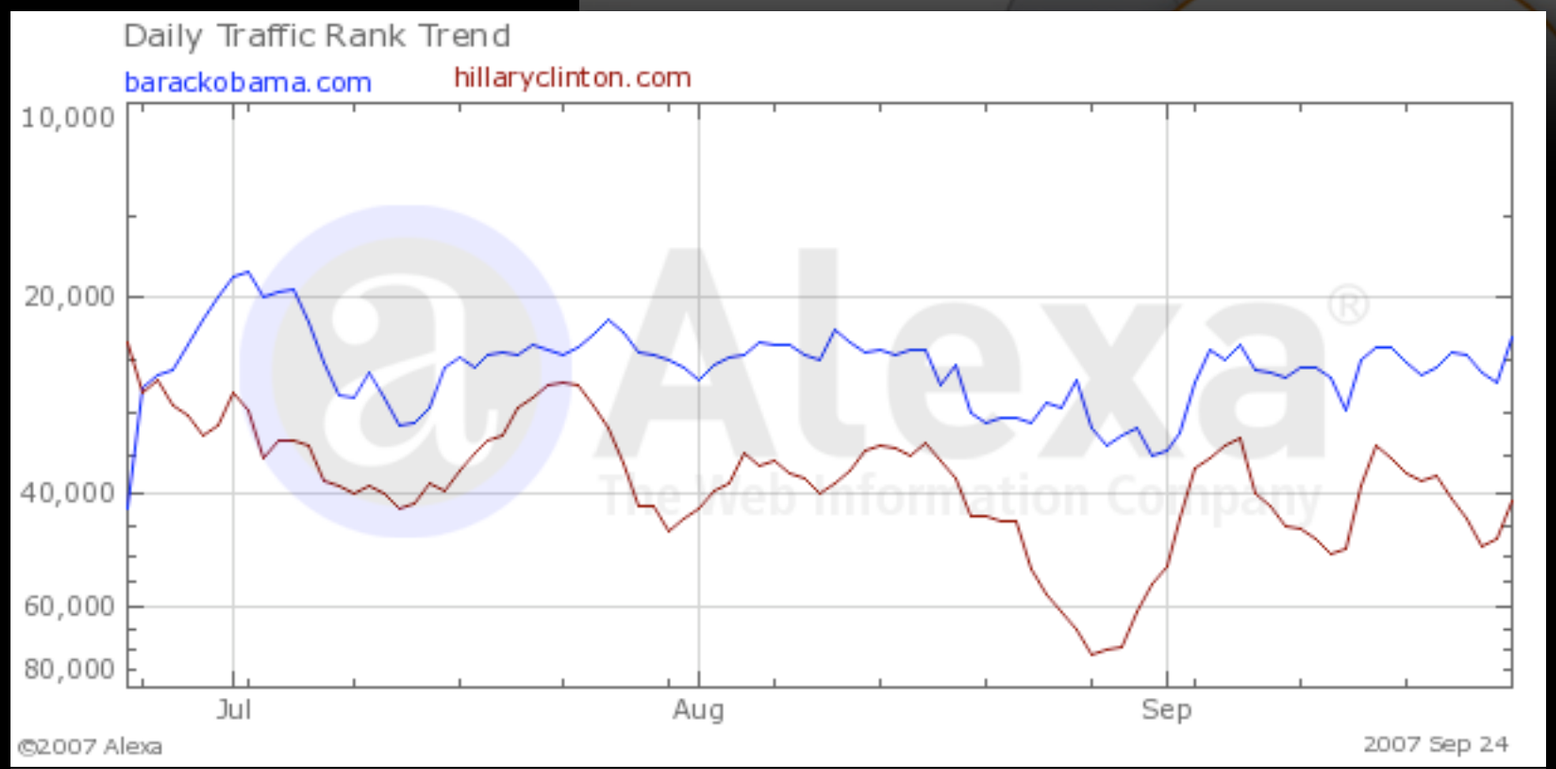


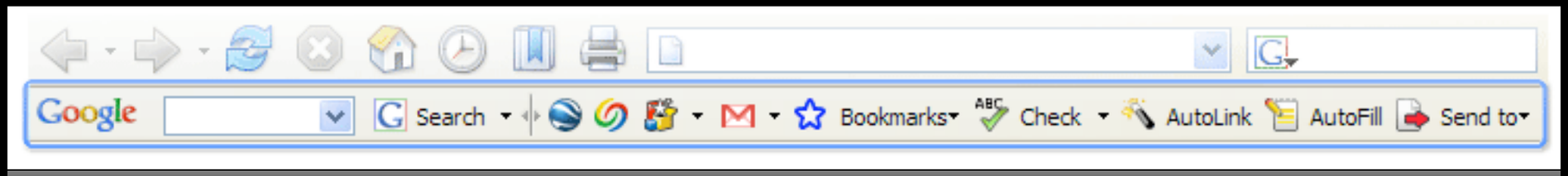

Google
Search
Bookmarks
Check
AutoLink
AutoFill
Send to

Toolbars


1. Yahoo!
www.yahoo.com
Site info for yahoo.com
2. Microsoft Network (MSN)
www.msn.com
Site info for msn.com
3. Google
www.google.com
Site info for google.com
4. YouTube
www.youtube.com
Site info for youtube.com
5. Windows Live
www.live.com
Site info for live.com
6. Myspace
www.myspace.com
Site info for myspace.com
7. Orkut
www.orkut.com
Site info for orkut.com
8. Facebook
www.facebook.com
Site info for facebook.com
9. Wikipedia
www.wikipedia.org
Site info for wikipedia.org


Party Ideas, Quick & Easy Recip
Related Links
Epicurious
Home and Garden Television (HGTV)
History Channel
Related Links
Traffic Trend
Reach Meter
Traffic Rank
#1,772
Summer Drinks


Daily Traffic Rank Trend
barackobama.com
hillaryclinton.com
10,000
20,000
40,000
60,000
80,000
Jul
Aug
Sep
©2007 Alexa
2007 Sep 24



1. **Yahoo!**
www.yahoo.com
Site info for yahoo.com

2. **Microsoft Network (MSN)**
www.msn.com
Site info for msn.com

3. **Google**
www.google.com
Site info for google.com

4. **YouTube**
www.youtube.com
Site info for youtube.com

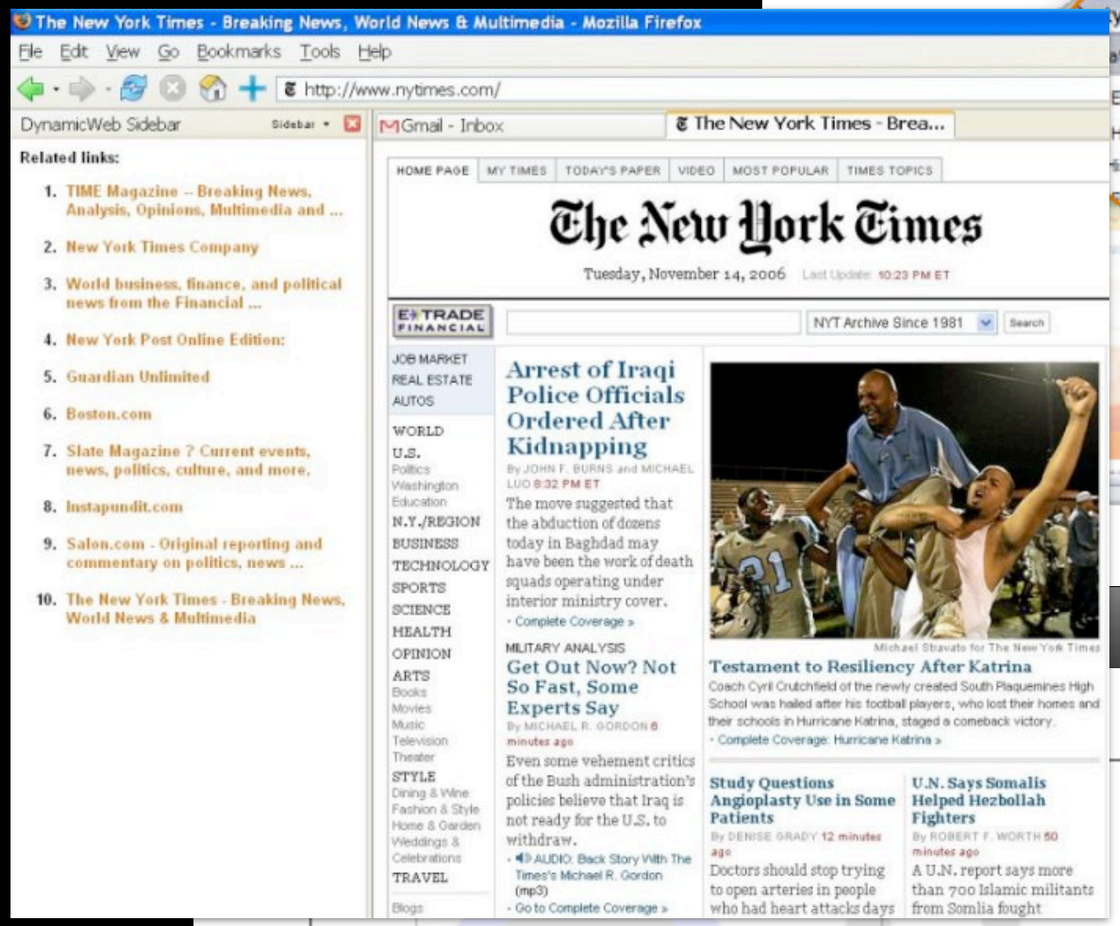
5. **Windows Live**
www.live.com
Site info for live.com

6. **Myspace**
www.myspace.com
Site info for myspace.com

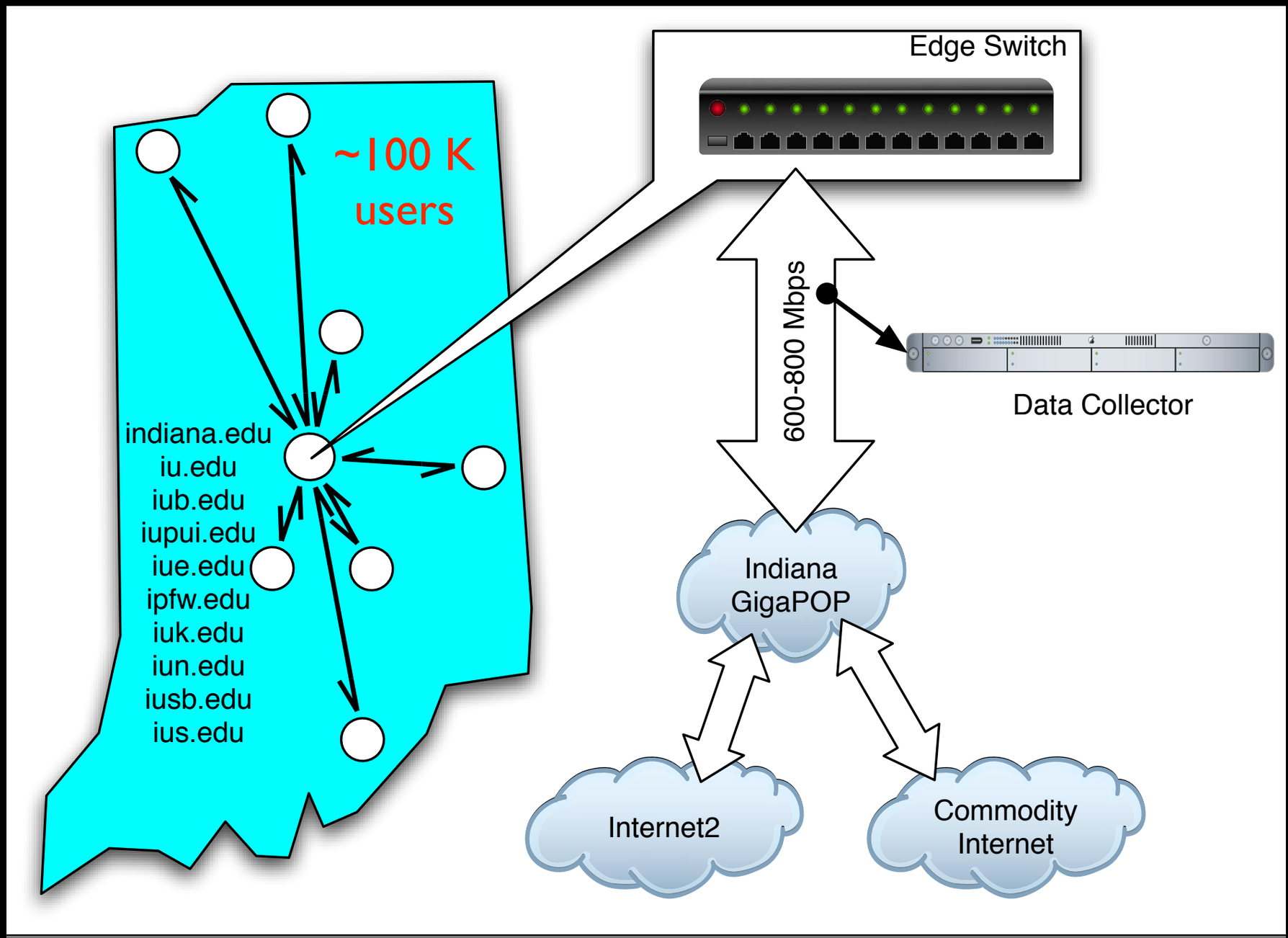
7. **Orkut**
www.orkut.com
Site info for orkut.com

8. **Facebook**
www.facebook.com
Site info for facebook.com

9. **Wikipedia**
www.wikipedia.org
Site info for wikipedia.org



ISP


Edge Switch
~100 K
users
indiana.edu
iu.edu
iub.edu
iupui.edu
iue.edu
ipfw.edu
iuk.edu
iun.edu
iusb.edu
ius.edu
600-800 Mbps
Data Collector
Indiana
GigaPOP
Internet2
Commodity
Internet

The Internet is for...

The Internet is for...

Clicky: Search Results

http://steinbeck.ucs.indiana.edu/~mmeiss/clickbrowser/

Google

Clicky: Search Results for 'porn'

7470 results found.

Sorted by descending order of *instr*

Host	k_{in}	k_{out}	s_{in}	s_{out}
tgp.pornaccess.com	626	106	112726	105097
free.freepornofreeporn.com	49	0	99907	0
thumbs.famouspornstars.com	7	0	84016	0
html.freepornofiles.com	81	6	48254	54338
photo.pornotube.com	26	1	46747	76
www.totallypornstars.net	13	0	43857	0
www.adultpornstarguide.com	9	0	34965	0
www.pornpimps.com	11	9	29752	30618
www.youngpornmovies.com	56	123	27788	27179
www.wildpornreviews.com	248	33	23781	17248
www.pornotube.com	46	14	22619	54389
www.bravoporn.com	17	13	21901	24446
www.megapornstarvids.com	33	108	21895	27394
playah.itsyourporn.com	3	4	21630	21405
products.adultlegalporn2go.com	4	4	20894	26663
www.8teenporn.com	33	88	19482	19535
img.porno-pics-free.com	4	0	19125	0
www.pornstarfinder.net	67	168	18540	18154
galleries2.porn365.com	17	0	17339	0
www.pornmoviepage.com	42	75	16852	17035
www.porno-vids.com	80	13	15578	14265
www.porneskimo.com	41	214	14464	29019
pic.freeporn.hu	27	0	14220	0
www.jointheporn.com	124	8	11560	10585
images.porninspector.com	4	1	11353	44
www.tgpornstars.com	22	105	11248	10831
www.duckyporn.com	45	338	11213	13314
www.pornstarscope.com	38	66	11206	13357
images.porndvddirect.com	4	0	11019	0
www.hpornstars.com	46	37	10874	10561
usemyporn.com	2	19	10651	15706
www.famouspornstars.com	63	287	10565	82240
scripts.sunporno.com	4	0	10550	0

Clicky

Clicky

Clicky: Search Results

http://steinbeck.ucs.indiana.edu/~mmeiss/clickbrowser/

Clicky: Search Results for 'porn'

7470 results found.

Sorted by descending order of *instr*

Host	k_{in}	k_{out}	s_{in}	s_{out}
tgp.pornaccess.com	626	106	112726	105097
free.freepornofreeporn.com	49	0	99907	0
thumbs.famouspornstars.com	7	0	84016	0
html.freepornofiles.com	81	6	48254	54338
photo.pornotube.com	26	1	46747	76
www.totallypornstars.net	13	0	43857	0
www.adultpornstarguide.com	9	0	34965	0
www.pornpimps.com	11	9	29752	30618
www.youngpornmovies.com	56	123	27788	27179
www.wildpornreviews.com	248	33	23781	17248
www.pornotube.com	46	14	22619	54389
www.bravoporn.com	17	13	21901	24446
www.megapornstarvids.com	33	108	21895	27394
playah.itsyourporn.com	3	4	21630	21405
products.adultlegalporn2go.com	4	4	20894	26663
www.8teenporn.com	33	88	19482	19535
img.porno-pics-free.com	4	0	19125	0
www.pornstarfinder.net	67	168	18540	18154
galleries2.porn365.com	17	0	17339	0
www.pornmoviepage.com	42	75	16852	17035
www.porno-vids.com	80	13	15578	14265
www.porneskimo.com	41	214	14464	29019
pic.freeporn.hu	27	0	14220	0
www.jointheporn.com	124	8	11560	10585
images.porninspector.com	4	1	11353	44
www.tgpornstars.com	22	105	11248	10831
www.duckyporn.com	45	338	11213	13314
www.pornstarscope.com	38	66	11206	13357
images.porndvddirect.com	4	0	11019	0
www.hpornstars.com	46	37	10874	10561
usemyporn.com	2	19	10651	15706
www.famouspornstars.com	63	287	10565	82240
scripts.sunporno.com	4	0	10550	0

Clicky: Viewing Node 'tgp.pornaccess.com'

http://steinbeck.ucs.indiana.edu/~mmeiss/clickbrowser/

Clicky: Viewing Node 'tgp.pornaccess.com'

(Do another search.)

***Incoming* Links for tgp.pornaccess.com:**
(112726 clicks from 626 hosts)

Host	Weight
tgp.pornaccess.com	100533
-	5240
www.tiava.com	386
www.gigagalleries.com	305
www.empire18.com	247
www.xnxx.com	195
frogsex.com	187
www.miamatures.com	165
www.catlist.com	158
www.fuckk.com	153
tiava.com	151
www.searchgalleries.com	142
www.porno-pics-free.com	123
search.askjolene.com	114
www.onlymovies.com	114
www.easygals.com	108
www.boneme.com	95
www.altavista.com	94
www.lolitampegs.com	87
www.milfmovs.com	74
www.movietitan.com	71
www.miateens.com	69
www.freeones.com	66
goatlist.com	65
www.theboobsmovies.com	64
www.searchvids.com	62
www.gaymoviedome.com	61
www.sexoasis.com	61
www.lodita.com	59
www.searchgals.com	54
bigtits-cinema.com	47
www.shaggle.com	41

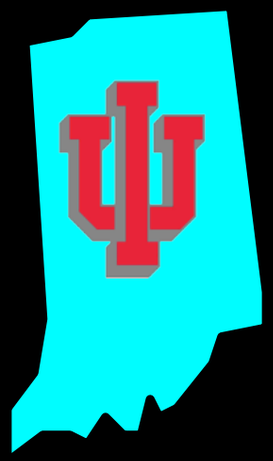
***Outgoing* Links for tgp.pornaccess.com:**
(105097 clicks to 106 hosts)

Host	Weight
tgp.pornaccess.com	100533
tgpvideos.pornaccess.com	3499
sweetliltranny.promo.pornaccess.com	610
track.pornaccess.com	132
vip.pornaccess.com	74
www.pornaccess.com	69
track.gaypornaccess.com	12
shemalemovies.pornaccess.com	11
amateurmovies.test.pornaccess.com	8
allnylonmovies.test.pornaccess.com	7
bustyisland.pornaccess.com	4
1001ultimatetits.pornaccess.com	4
maturetaboo.pornaccess.com	4
momshardvideo.promo.pornaccess.com	4
allnylonmovies.pornaccess.com	4
bustymomvideos.pornaccess.com	3
security-updater.com	3
sluttylittlebabysitters.pornaccess.com	3
momshardvideo.test.pornaccess.com	3
teachersandteenagers.pornaccess.com	3
grannyridesagain.pornaccess.com	3
em.gad-network.com	2
shemalessurprise.pornaccess.com	2
olderchicksfuckingyoungerdicks.pornaccess.com	2
247latexsex.pornaccess.com	2
sexinpublicplaces.pornaccess.com	2
joggs.pornaccess.com	2
babysitters.pornaccess.com	2
teacherspet.pornaccess.com	2
officesexmovies.pornaccess.com	2
hornyoldernymphos.pornaccess.com	2
dadsandtwinks.gaypornaccess.com	2

... But seriously... Outline

- Data collection
- Structural properties
- Behavioral patterns
- Temporal patterns
- PageRank validation

Data collection



HTTP (80)
30% @ peak

Internet

anonymizer

GET

requests from
IU only

Host
Path
Referer
User-Agent
Timestamp

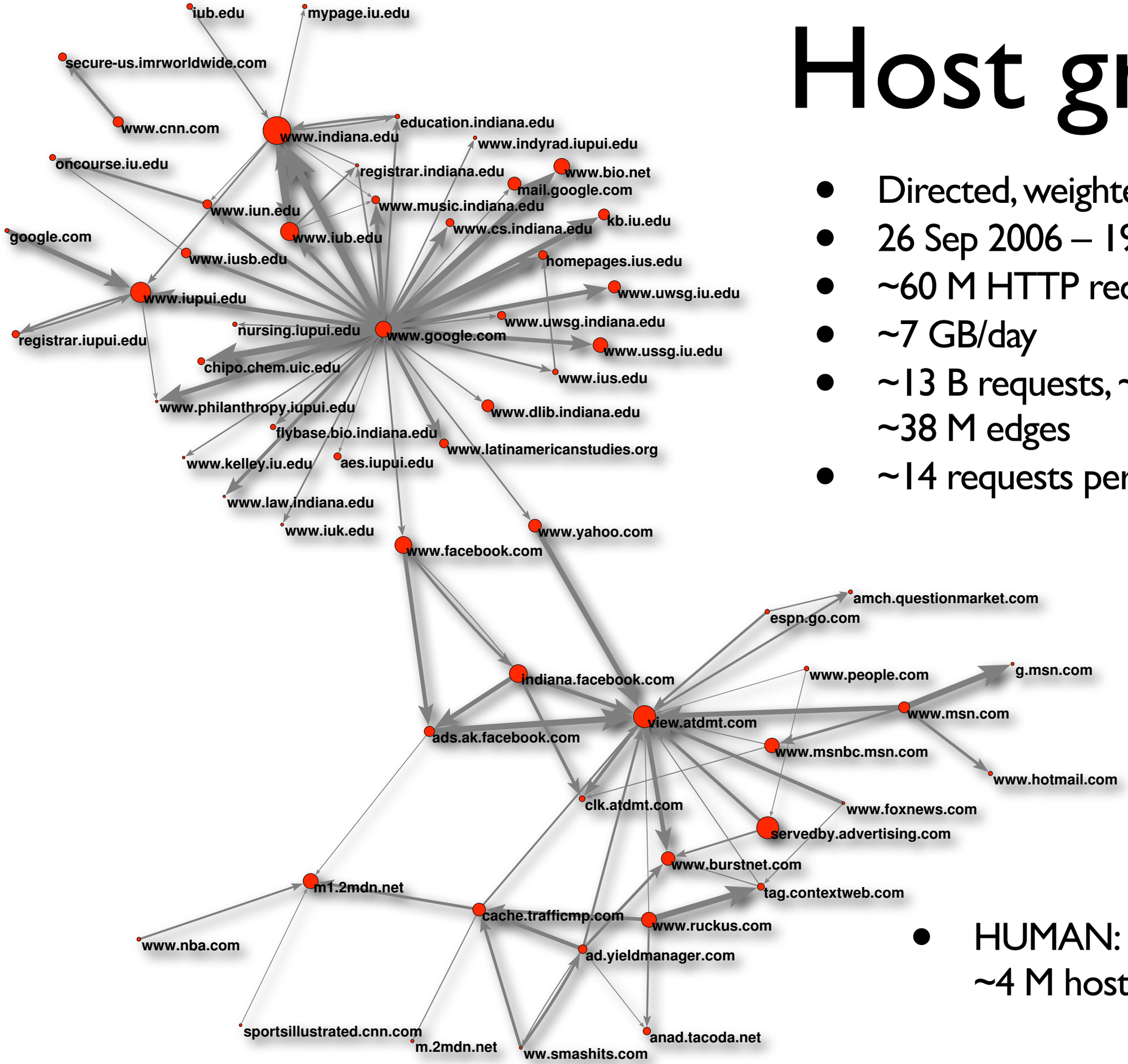
HUMAN
h..p..r..a..t

FULL
h..p..r..a..t

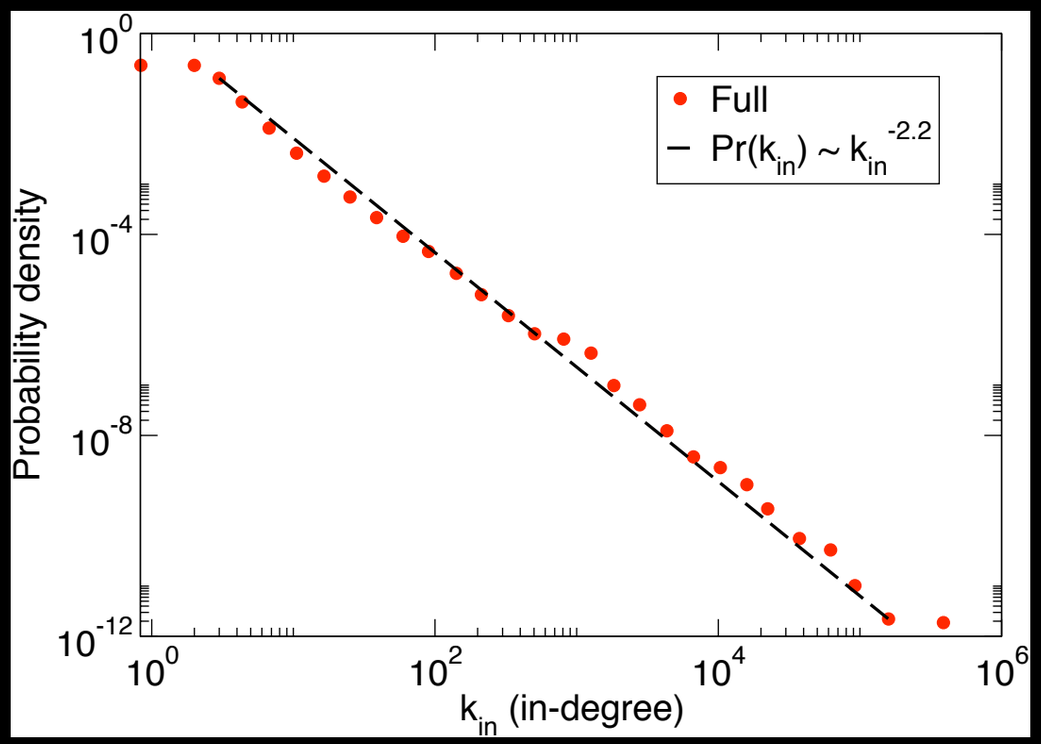
Host graphs

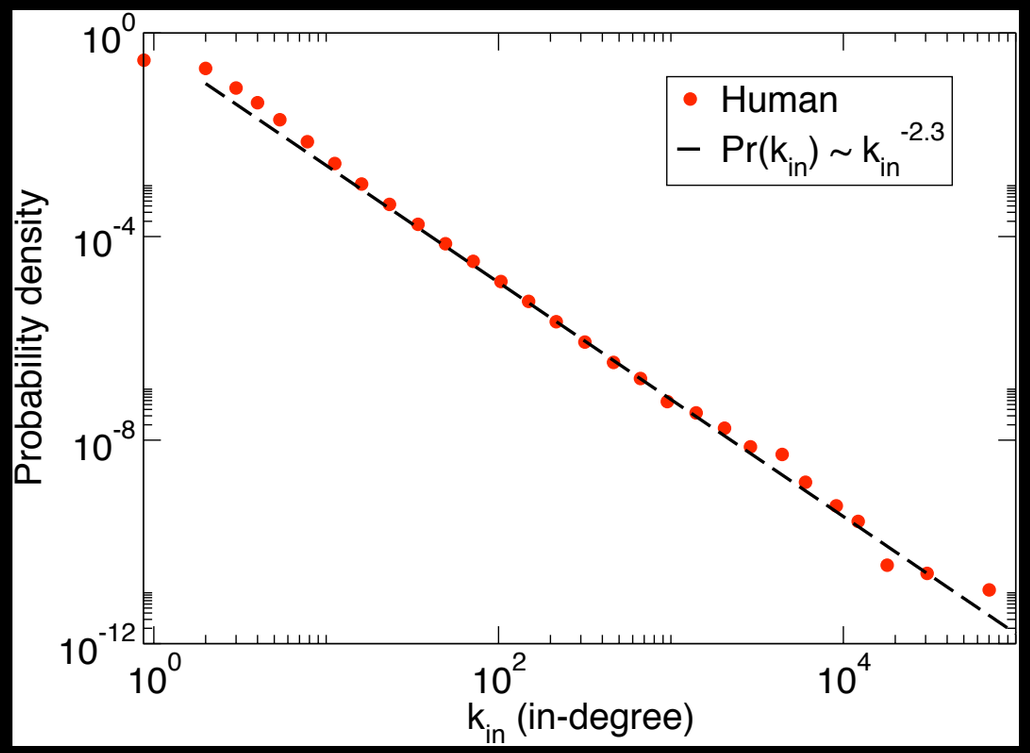
- Directed, weighted networks
- 26 Sep 2006 – 19 May 2007
- ~60 M HTTP requests per day
- ~7 GB/day
- ~13 B requests, ~8 M hosts, ~38 M edges
- ~14 requests per human click

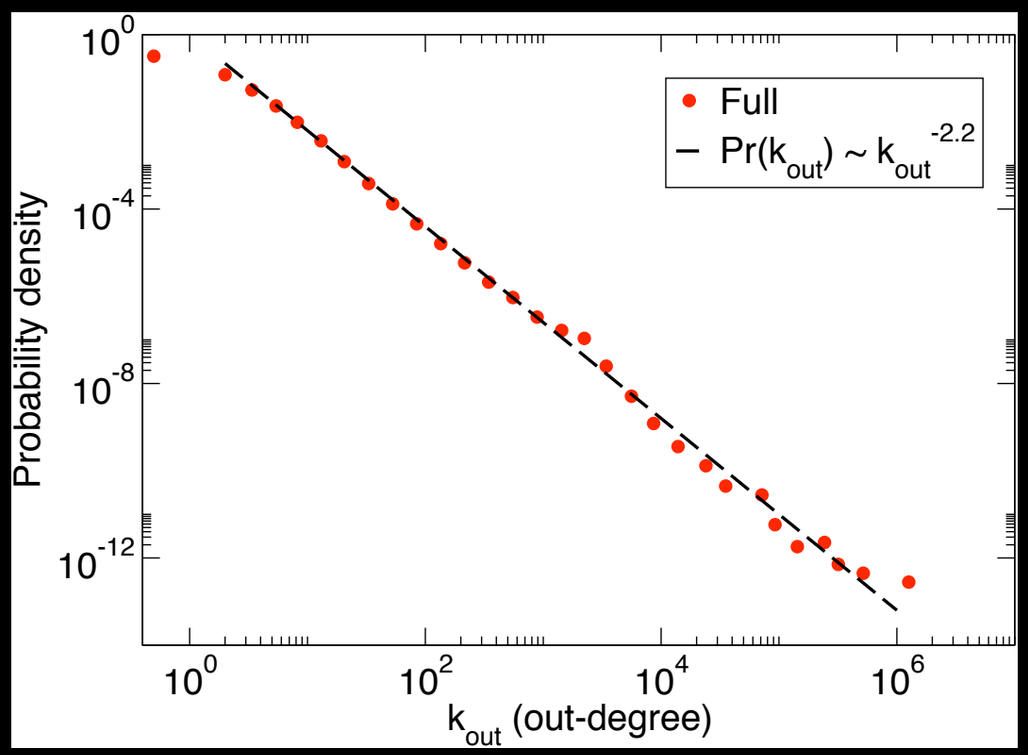
- HUMAN: ~1 B requests, ~4 M hosts, ~11 M edges

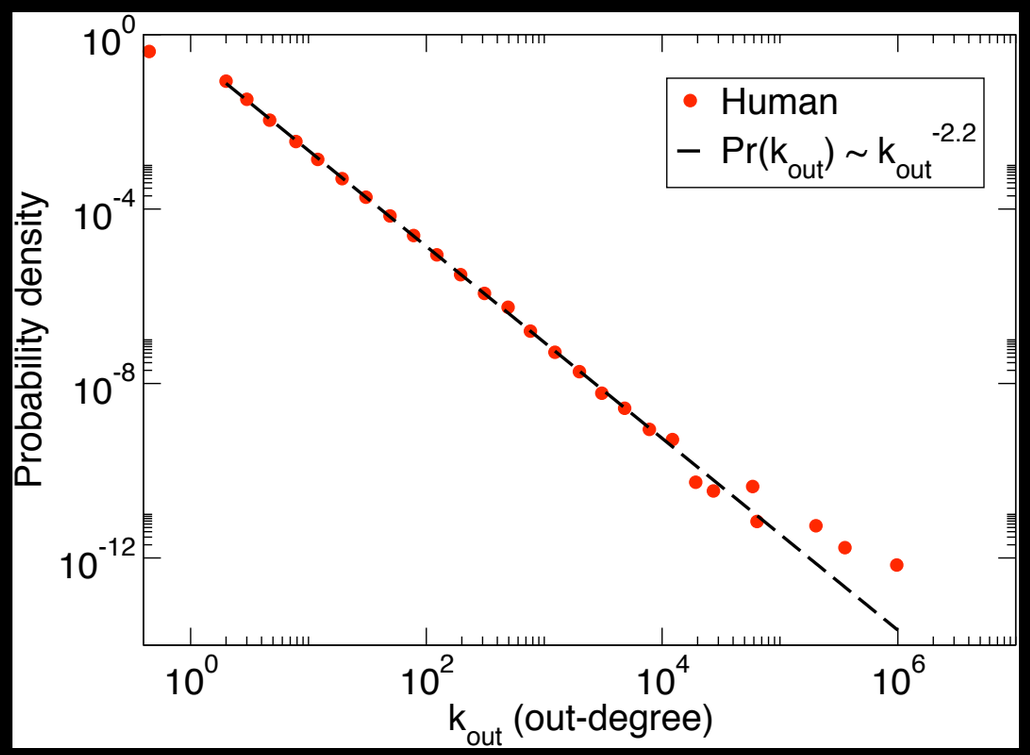


Structural properties: degree

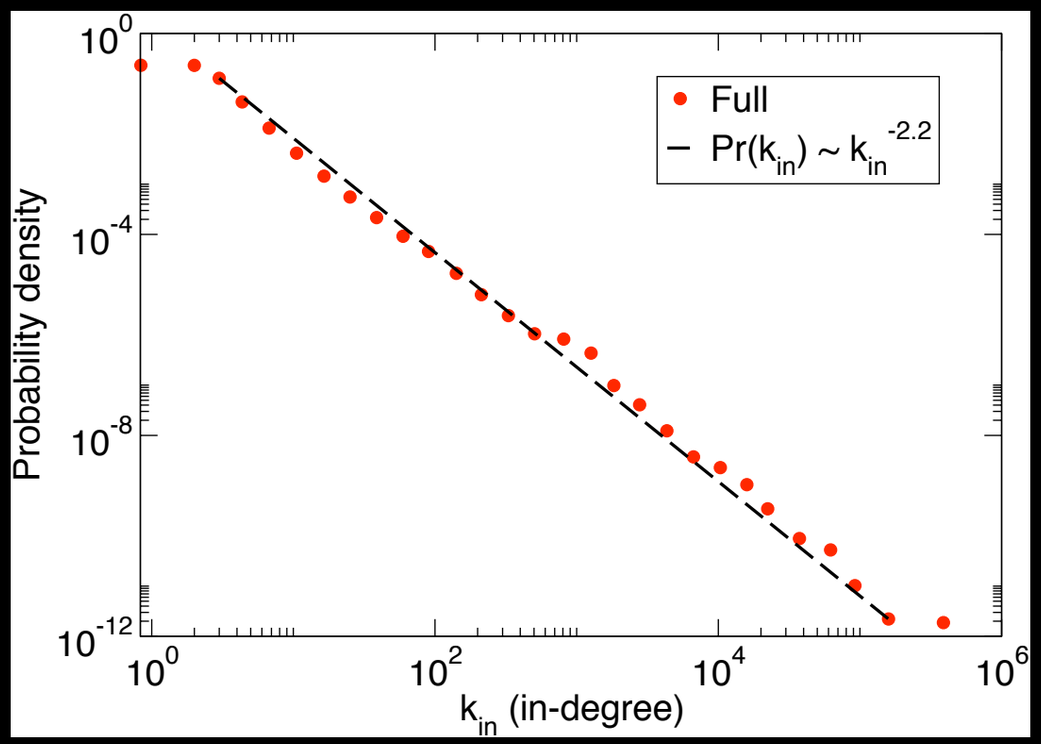


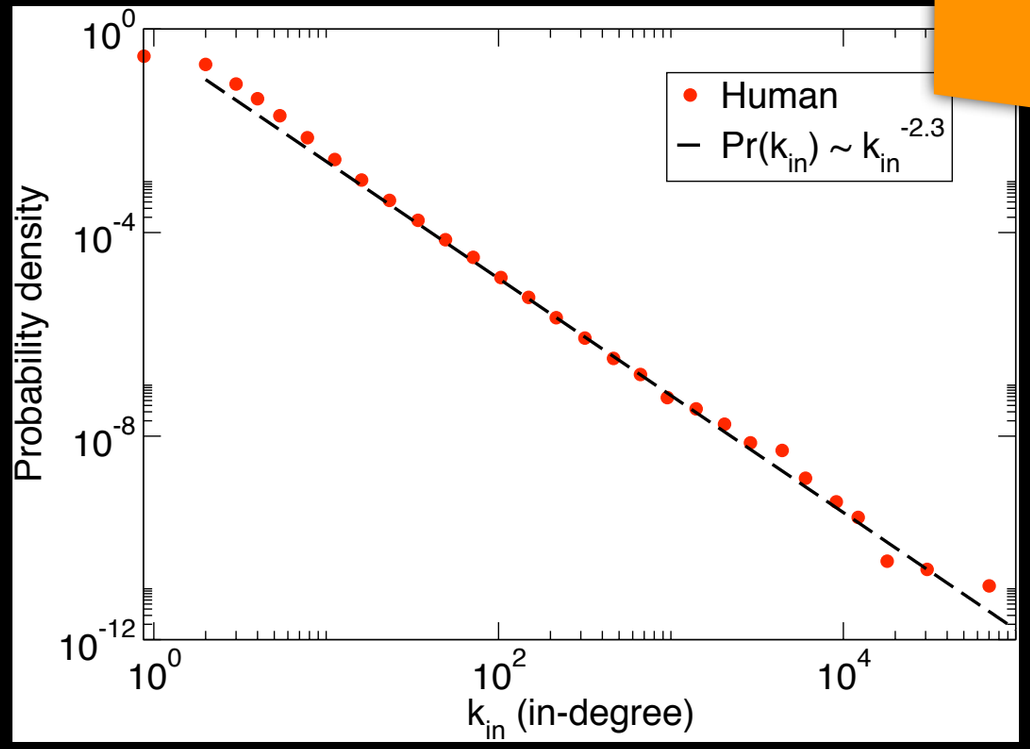


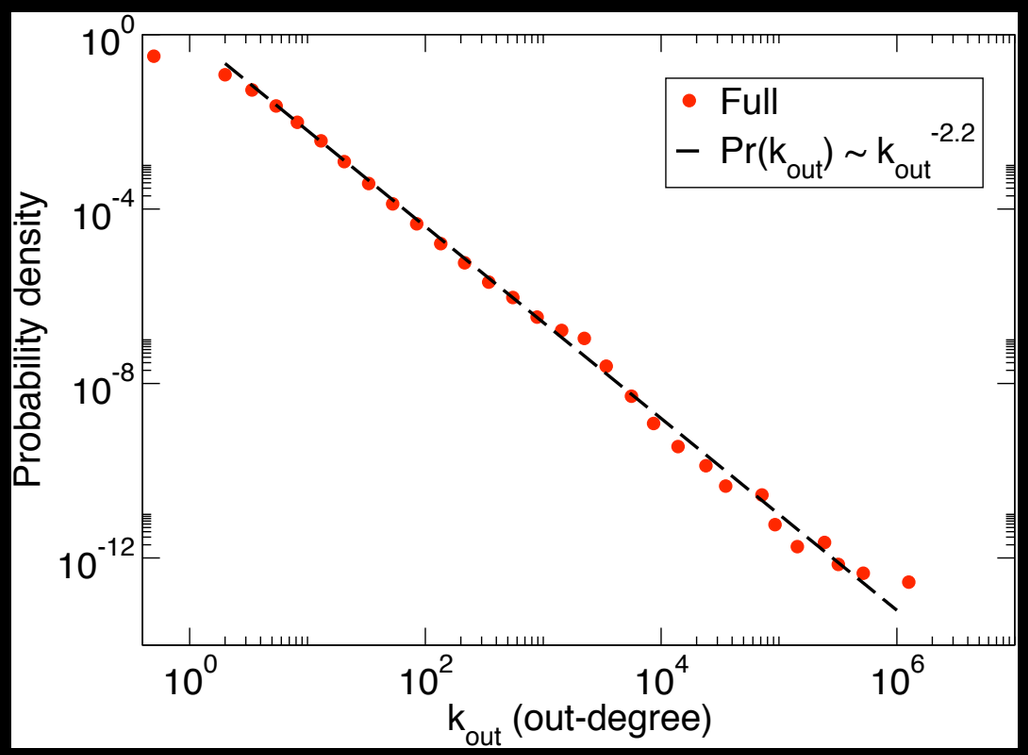


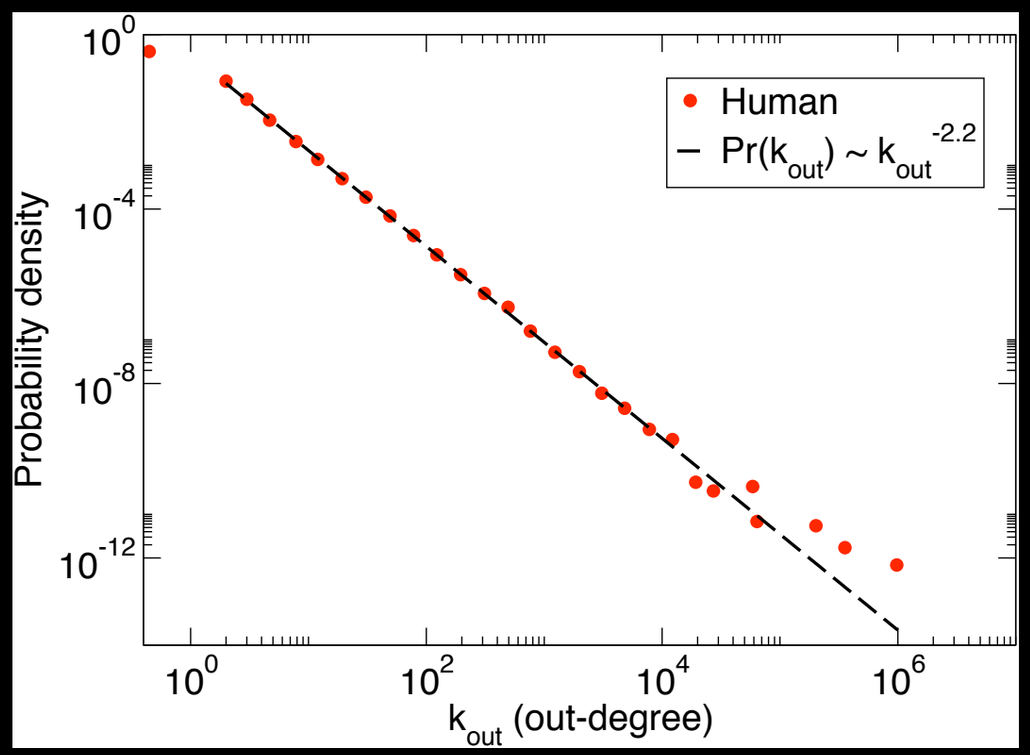


Structural properties: degree

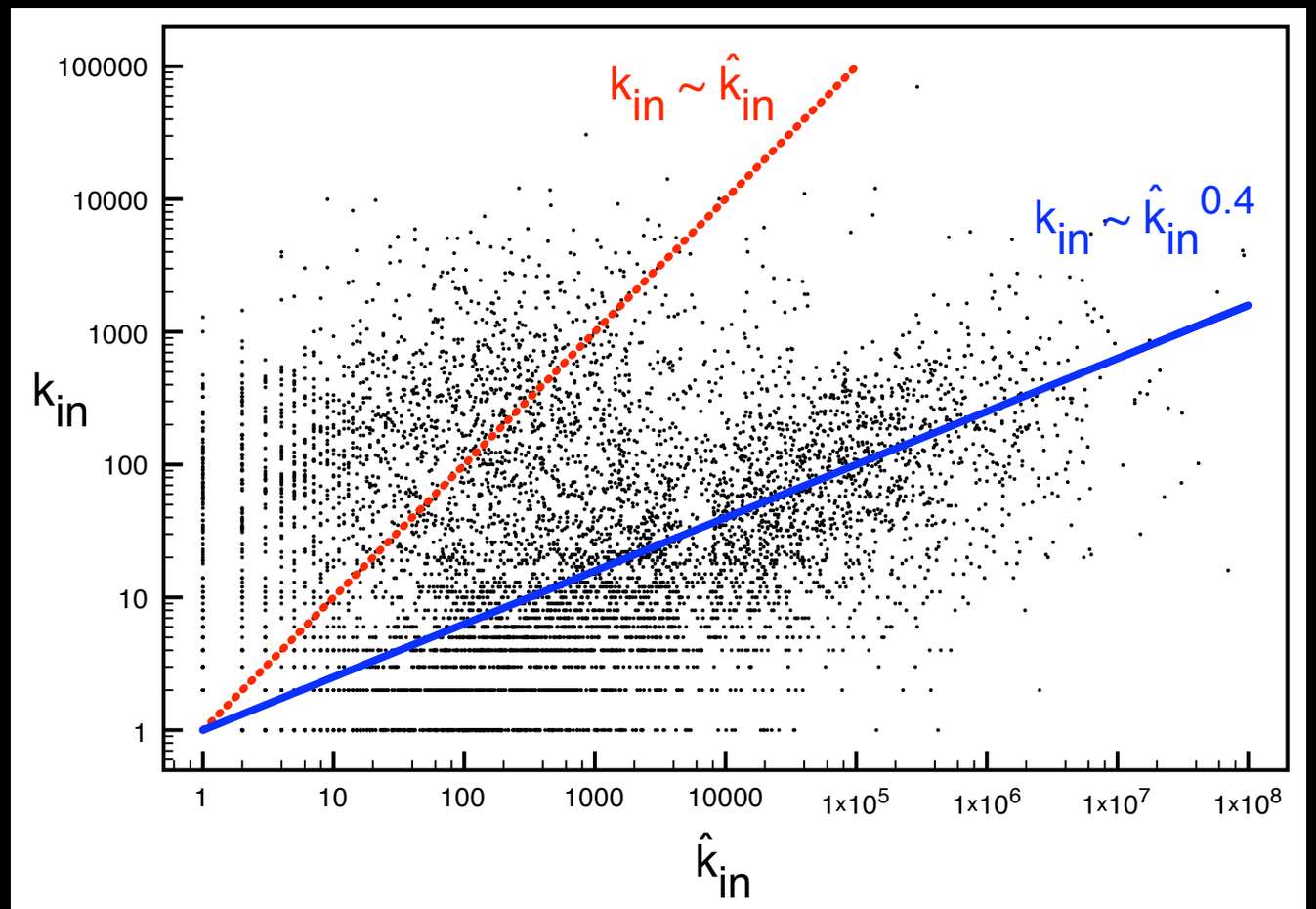








Caveat: sampling bias

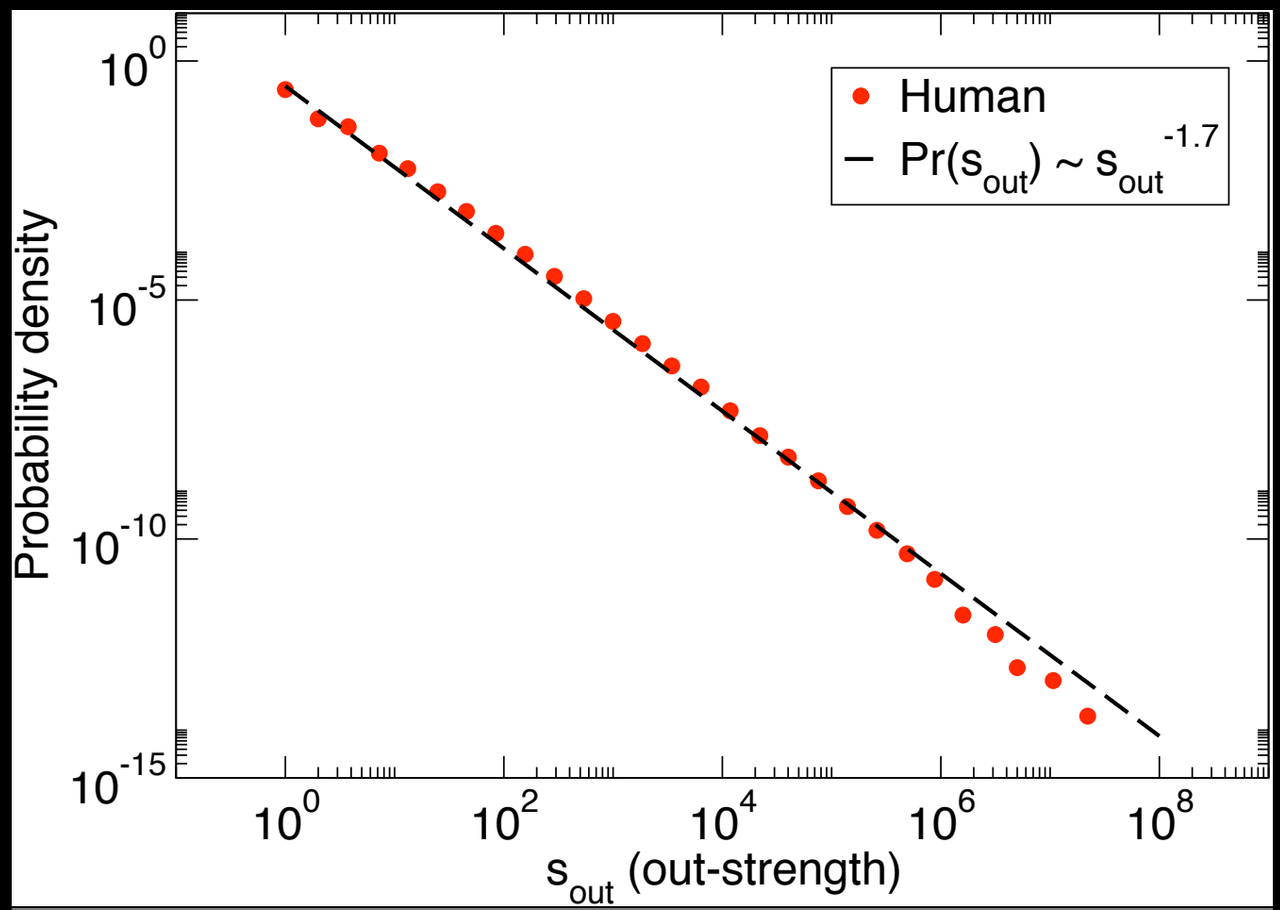


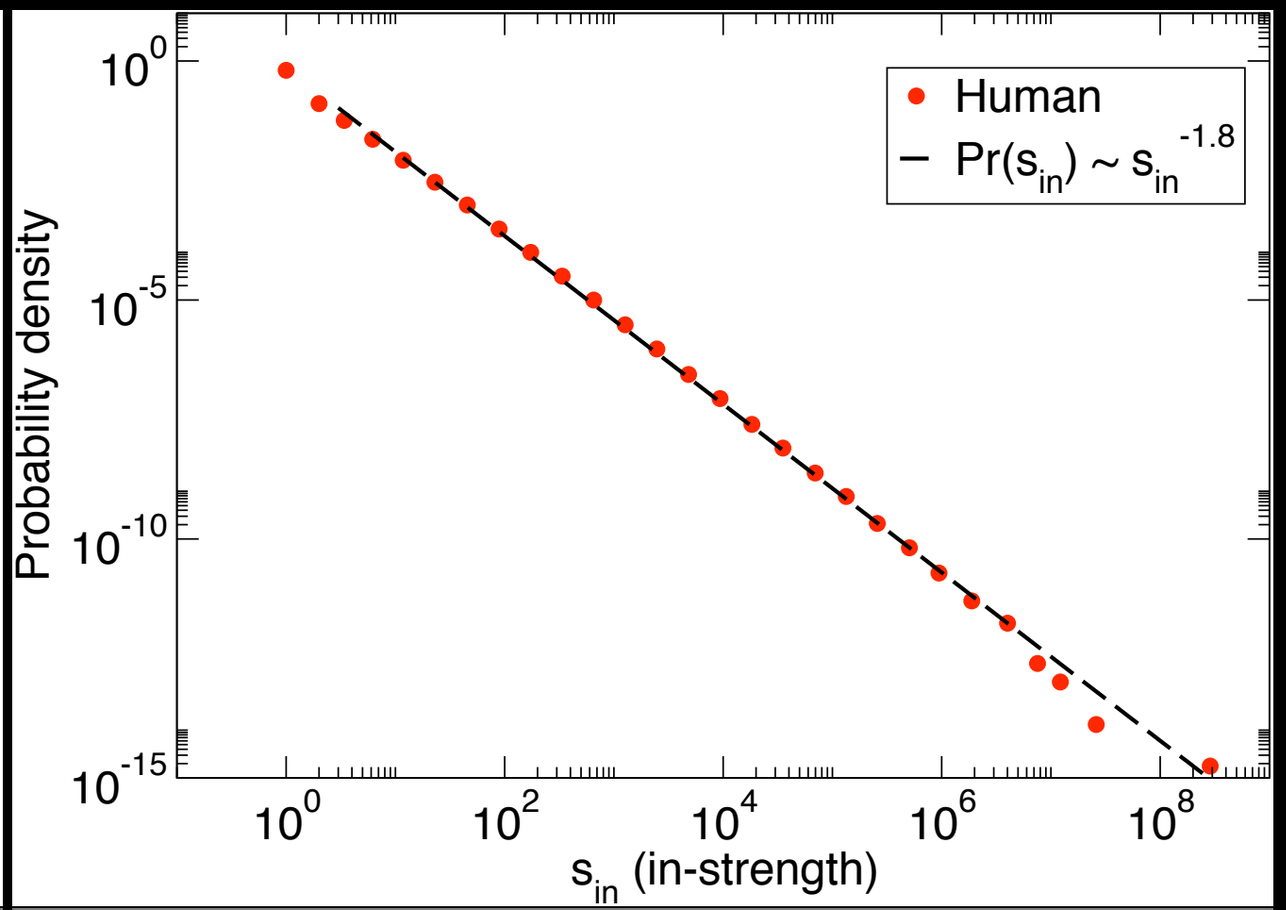
$$k_{in} \sim \hat{k}_{in}^{\eta}$$

$$\begin{aligned} \Pr(k_{in})dk_{in} &\sim k_{in}^{-\gamma} dk_{in} \\ &\sim \hat{k}_{in}^{-\eta\gamma} d(\hat{k}_{in}^{\eta}) \\ &\sim \hat{k}_{in}^{-\eta\gamma+\eta-1} d\hat{k}_{in} \\ &\sim \hat{k}_{in}^{-\hat{\gamma}} d\hat{k}_{in} \end{aligned}$$

$$\begin{aligned} \gamma &= \frac{\hat{\gamma}-1}{\eta}+1 \\ &> \hat{\gamma} \text{ if } \eta < 1 \end{aligned}$$

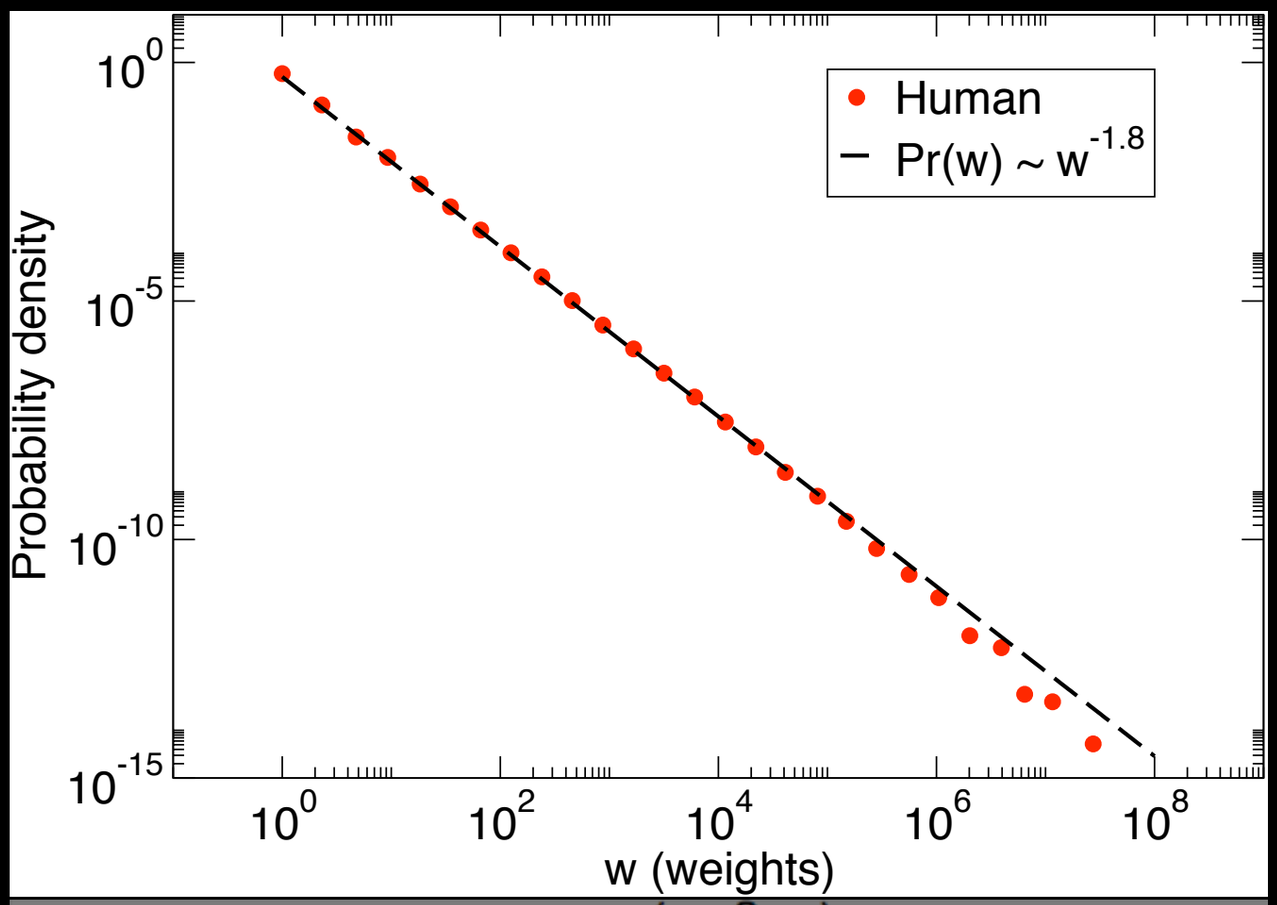
Structural properties: strength (site traffic)



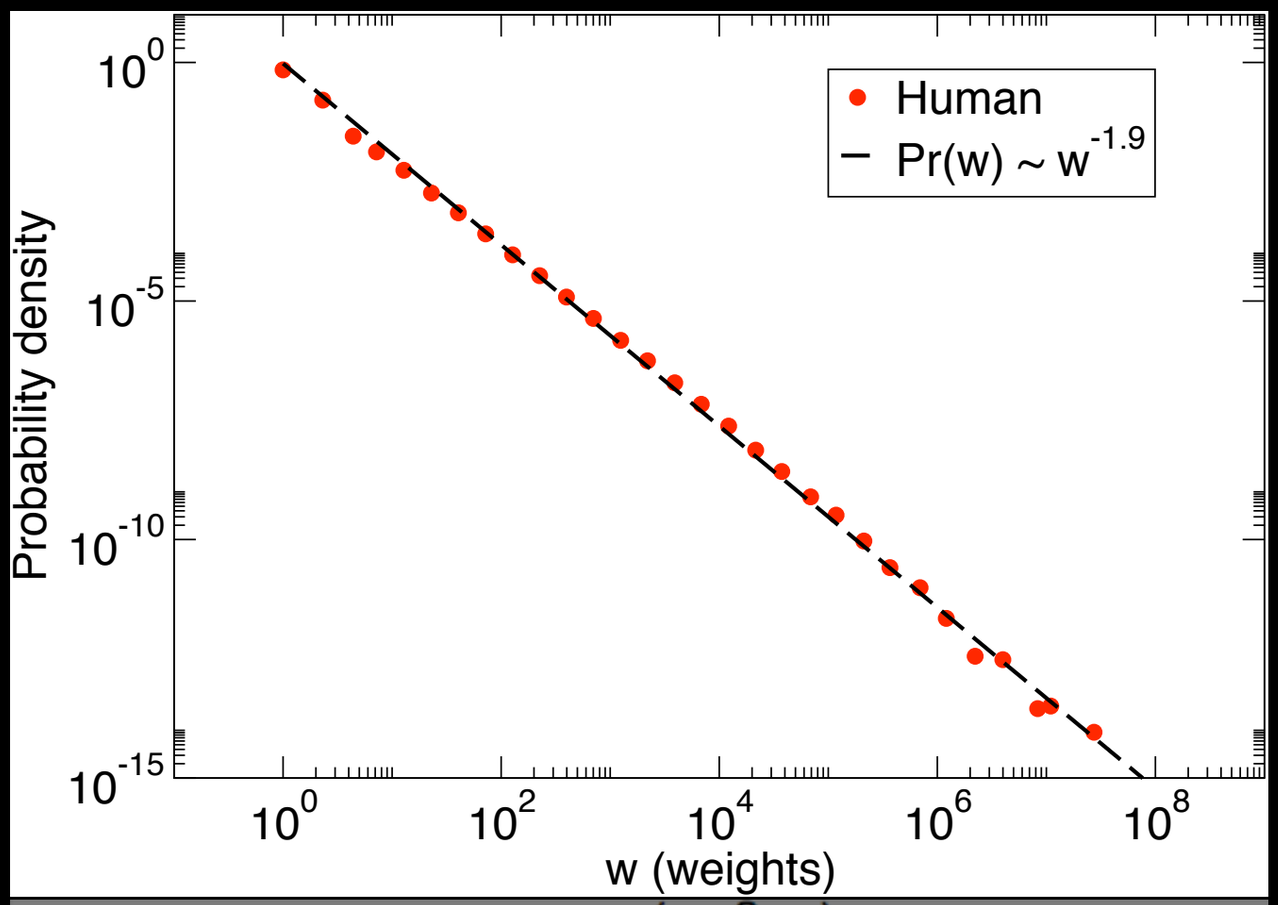


Structural properties: weights (link traffic)

Including empty referrer

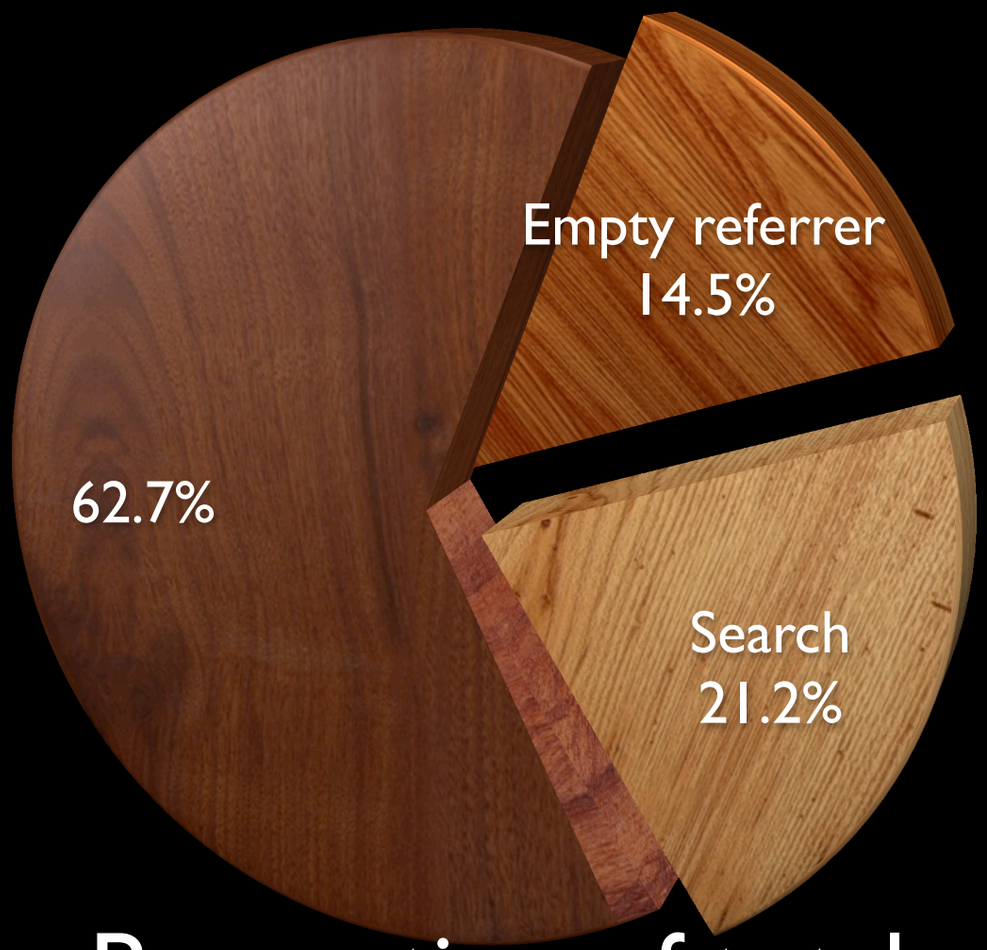


Excluding empty referrer

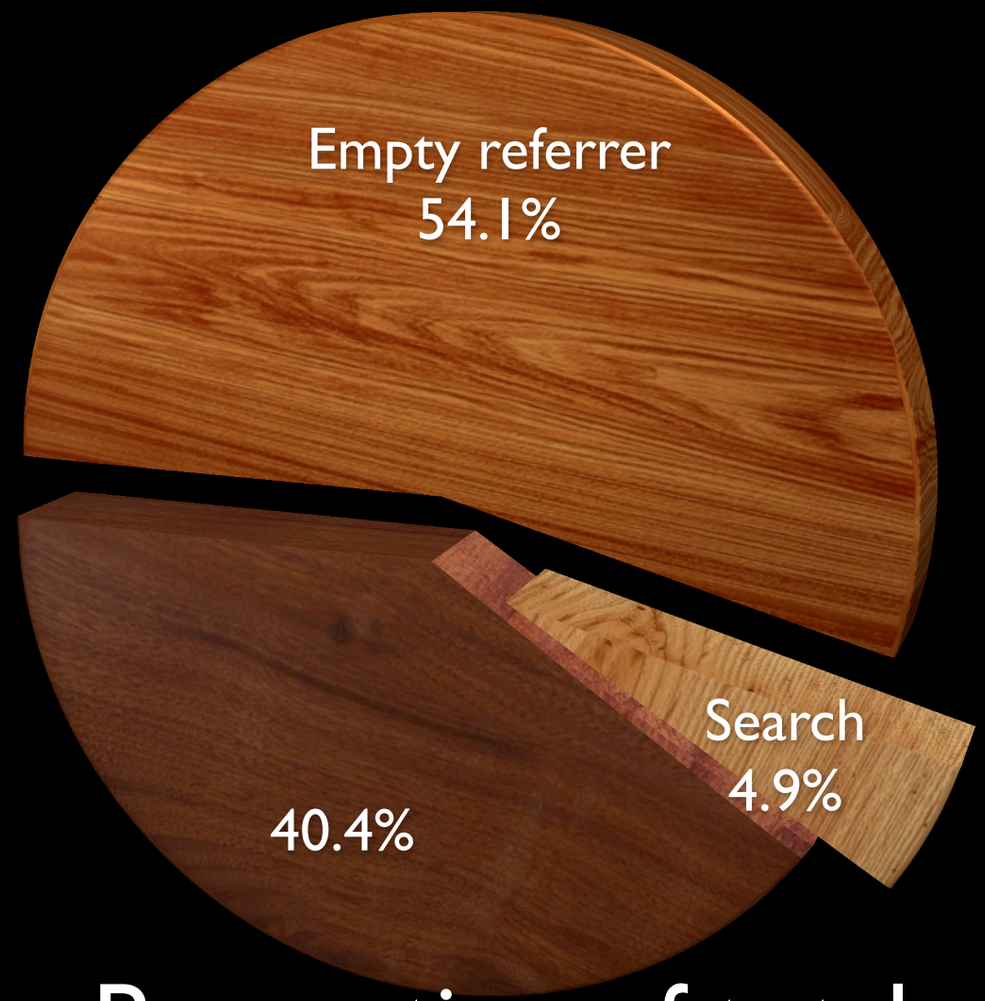


Behavioral patterns (HUMAN)

Empty referrer Search Web mail Other

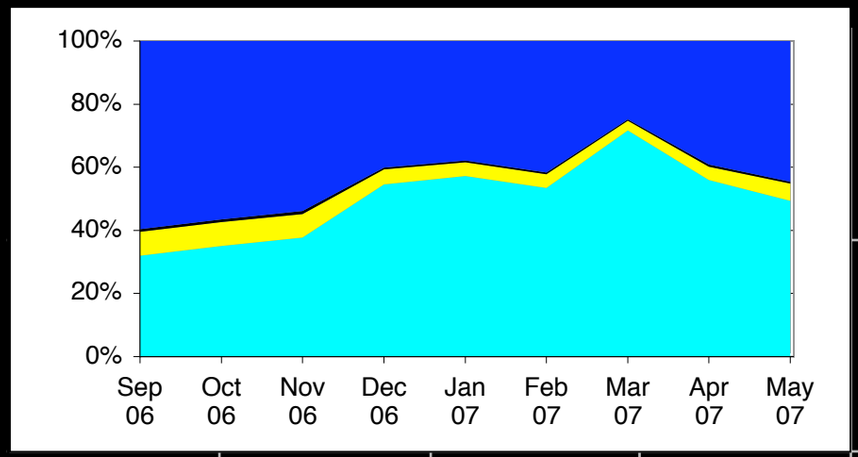


Proportion of total out-degree



Proportion of total out-strength

Ratios are stable



10^8 visits

10^6 visits

10^4 visits

10^2 visits

10^0 visits

Page traffic

100

10,000

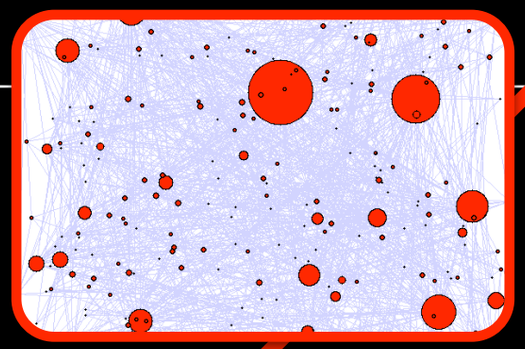
1,000,000

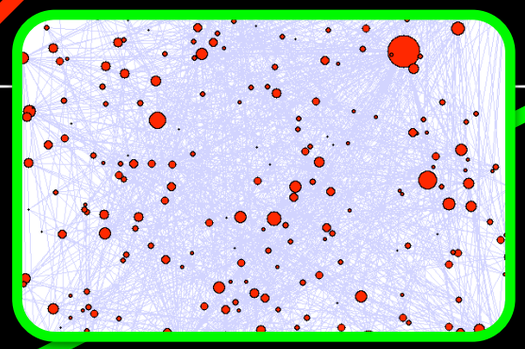
In-degree ~ PageRank

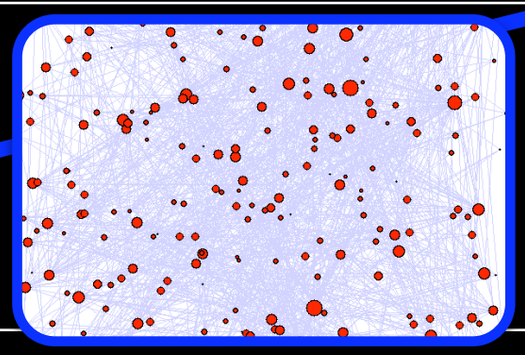
Googlearchy: search engines amplify rich-get-richer bias of the Web

Surfing without search engines: popularity reflects rich-get-richer bias of the Web

Data: search mitigates rich-get-richer bias of the Web

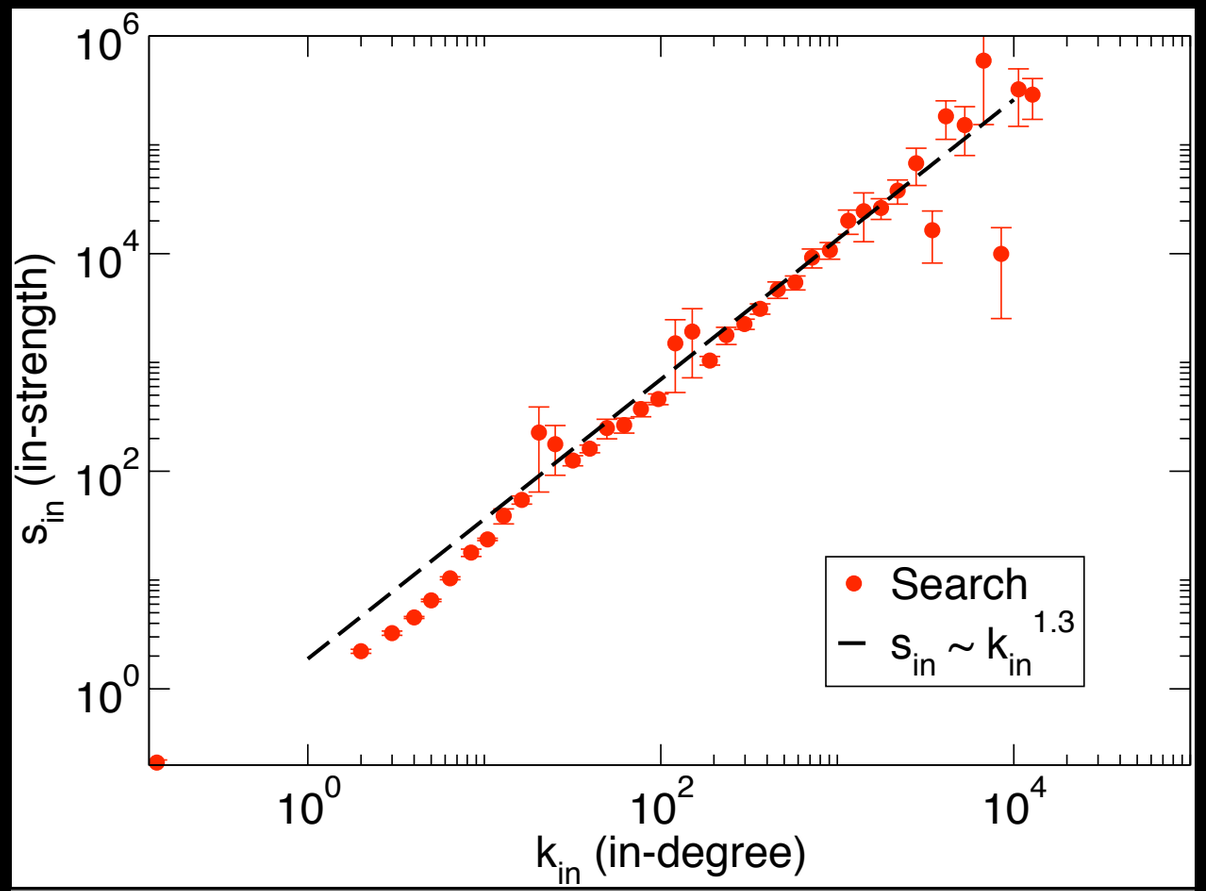






PNAS 2006

Does search mitigate the rich-get-richer dynamics?



$$k_{in} \sim \hat{k}_{in}^{\eta}$$

$$s_{in} \sim k_{in}^{\beta} \sim \hat{k}_{in}^{\eta\beta}$$

$$\eta < 1/\beta \implies \eta\beta < 1$$

Temporal patterns

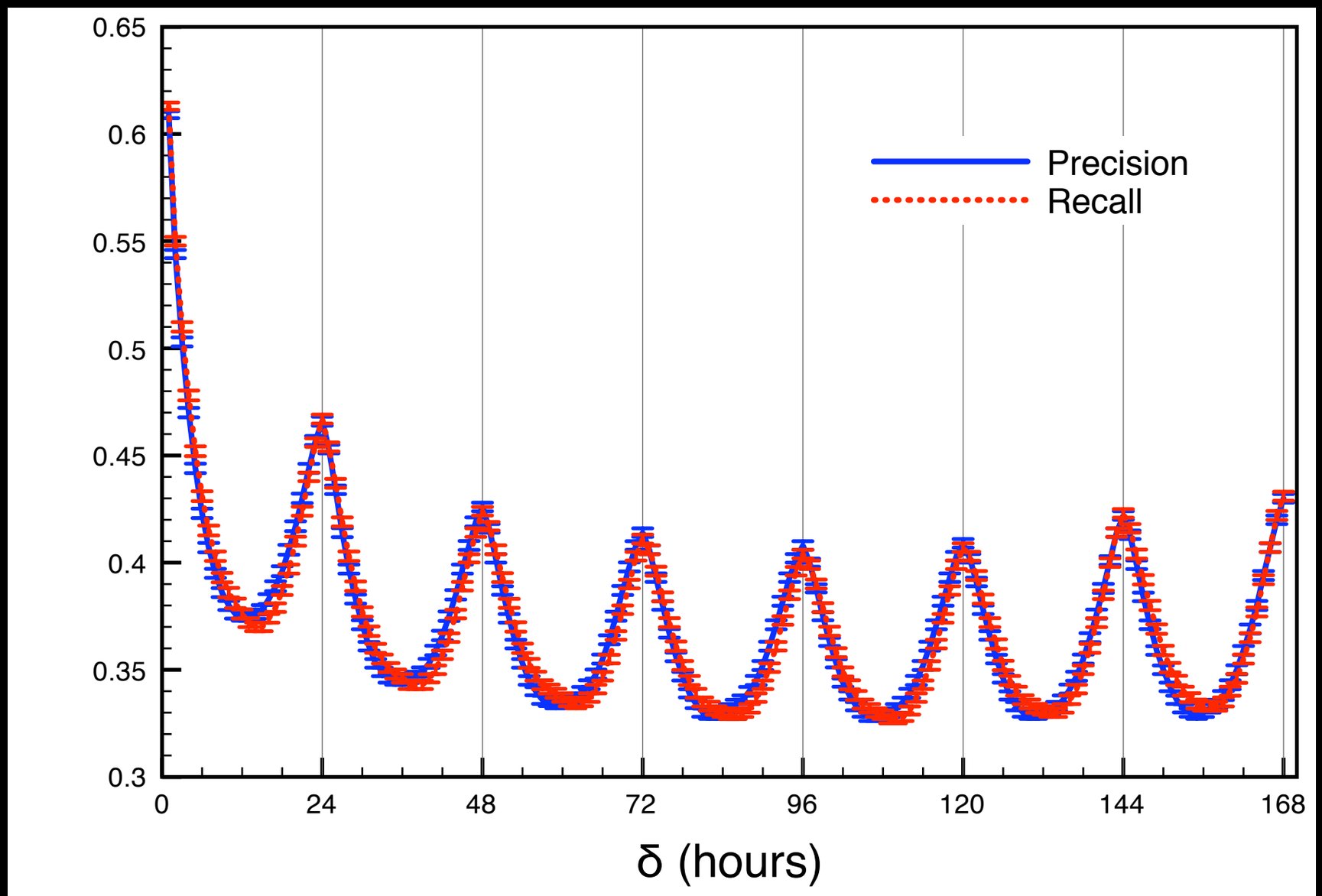
- How predictable are traffic patterns?
  - Cache refreshing (e.g. proxies)
  - Capacity allocation (e.g. provisioning for spikes)
  - Site design (e.g. expose content based on time of day context)

Temporal patterns

- Predict future host graph (clicks) from current one, as a function of delay
- Generalize temporal <u>precision</u> and <u>recall</u>:

$$P(\delta) = \sum_{t=\delta}^{T} \left(\frac{\sum_{ij} \min[w_{ij}(t), w_{ij}(t-\delta)]}{\sum_{ij} w_{ij}(t-\delta)} \right)$$

$$R(\delta) = \sum_{t=\delta}^{T} \left(\frac{\sum_{ij} \min[w_{ij}(t), w_{ij}(t-\delta)]}{\sum_{ij} w_{ij}(t)} \right) .$$

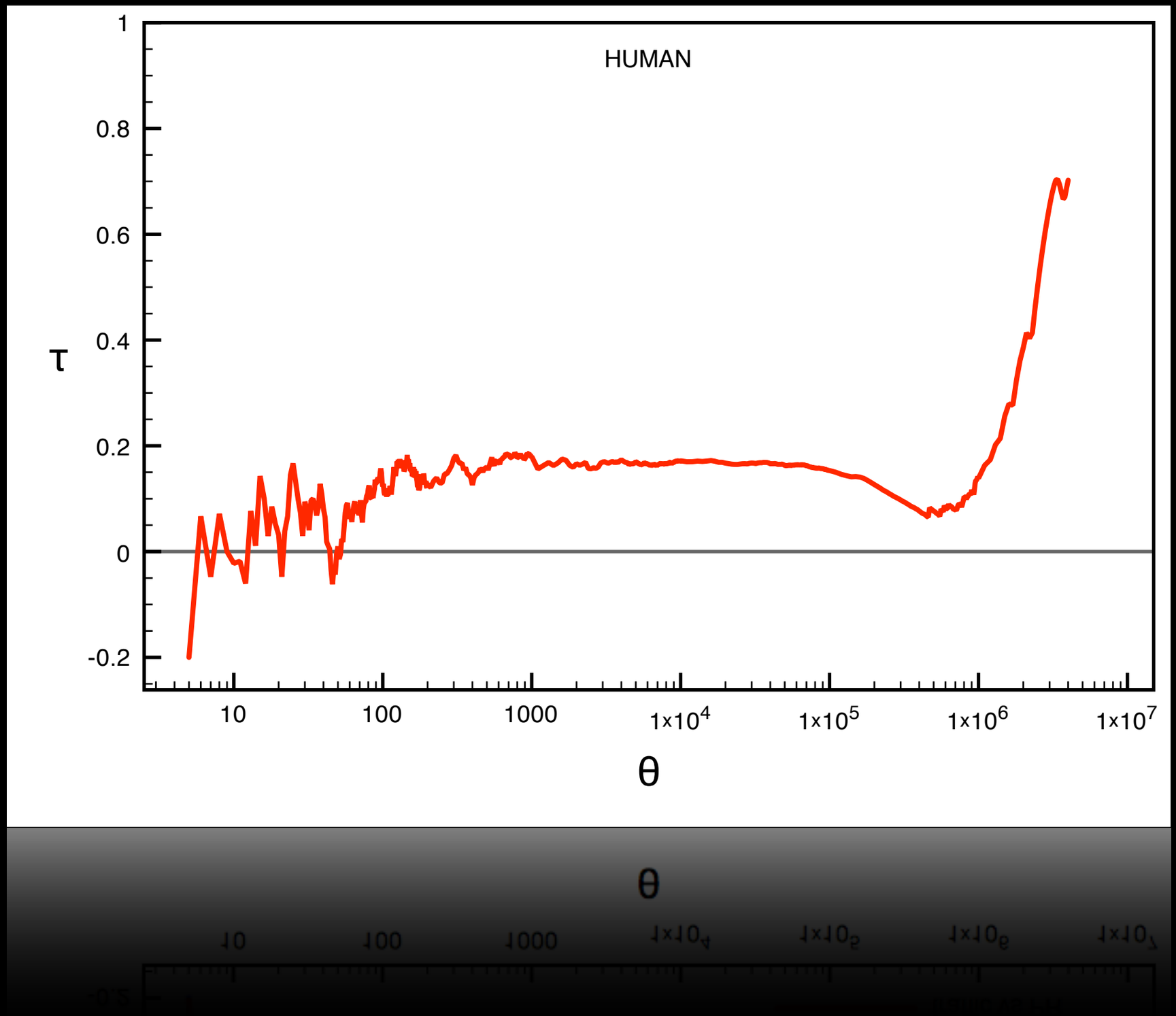


HUMAN host graph
(FULL is about 10% more predictable)

PageRank

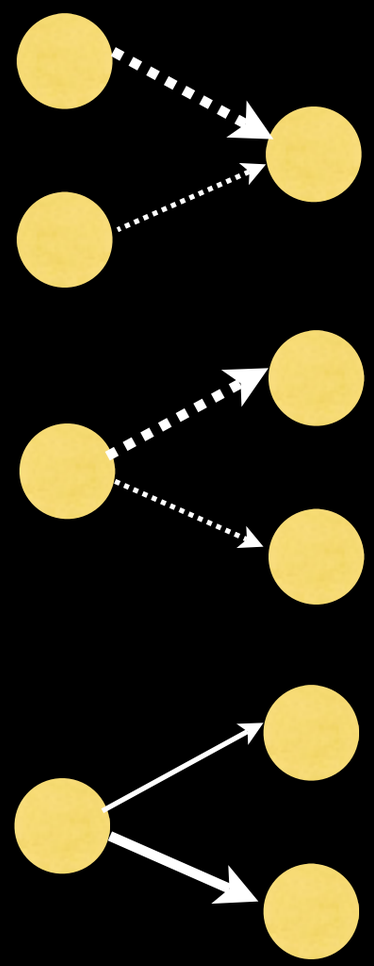
- PR as a model of Web navigation: stationary distribution of visit frequency by a modified random walk (with jumps) on the Web graph
- Compare with actual site traffic (in-strength)
- From an application perspective, we care about the resulting ranking of sites rather than the actual values

Kendall's rank correlation



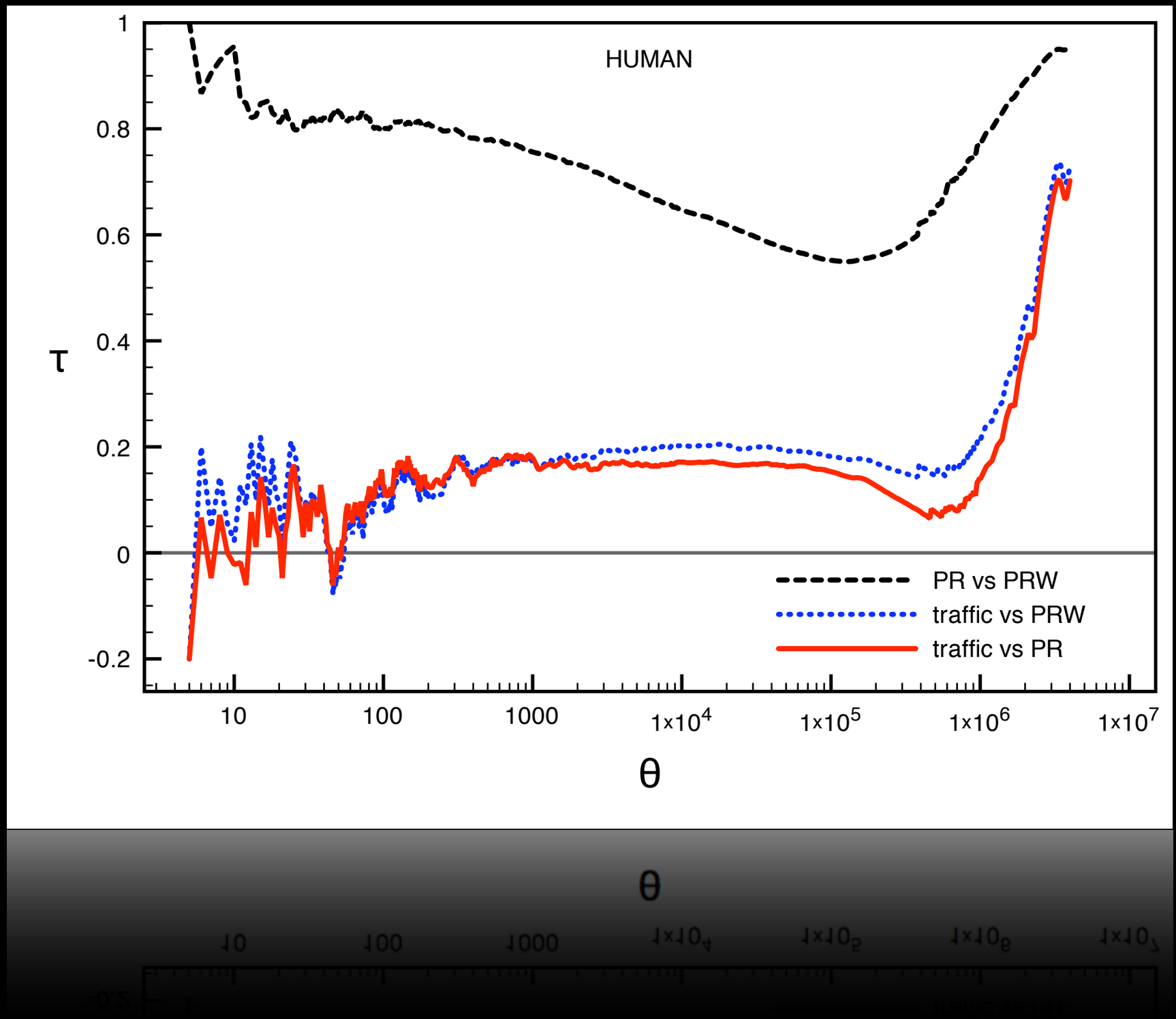
PageRank assumptions

1. Equal probability of teleporting <u>from</u> each of the nodes
2. Equal probability of teleporting <u>to</u> each of the nodes
3. Equal probability of following each link from any given node



$$PRW(j) = \frac{\alpha}{N} + (1-\alpha) \sum_{i:w_{ij} \neq 0} \frac{w_{ij}}{s_{out}(i)} PRW(i)$$

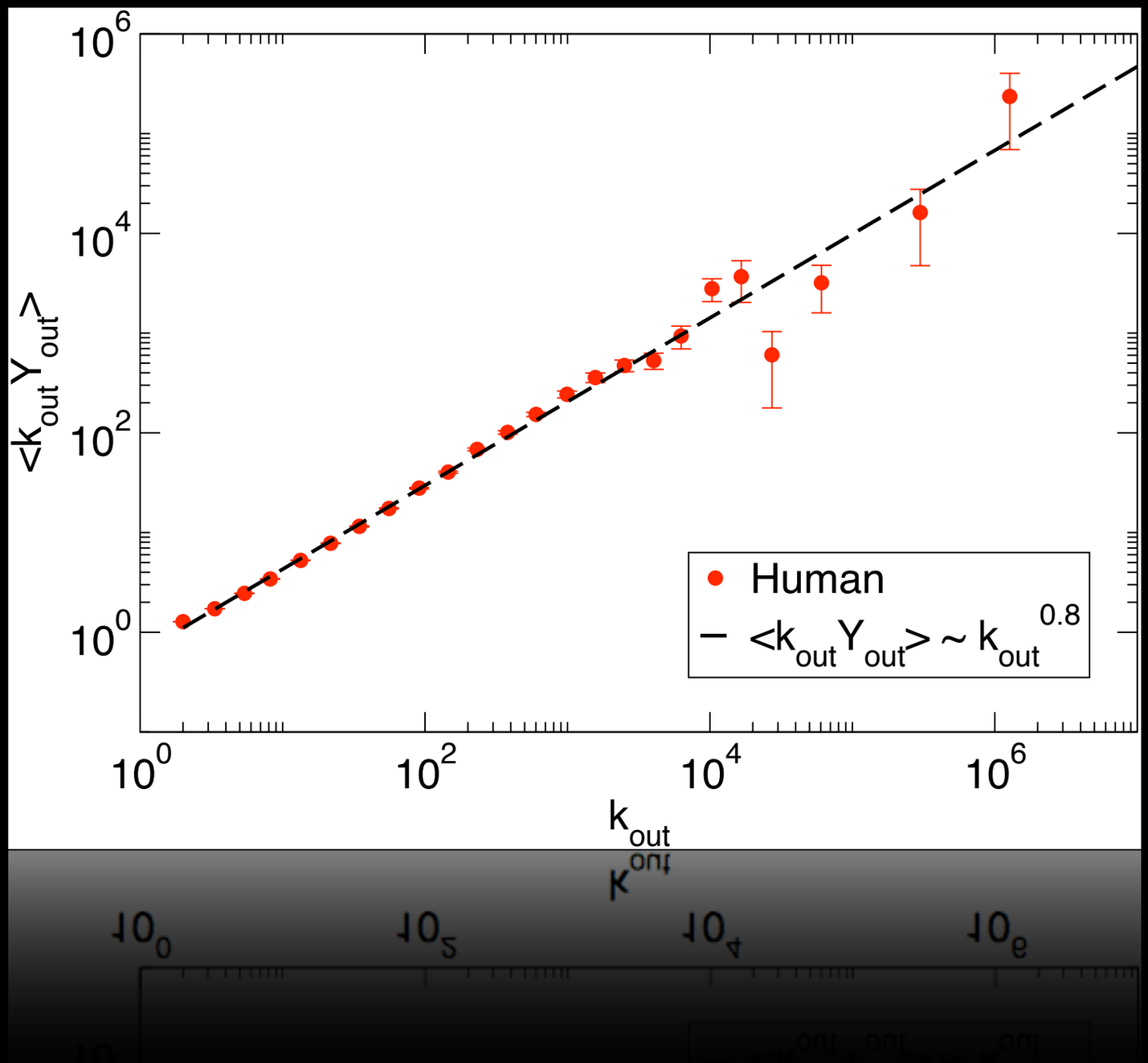
Kendall's rank correlation



Local link heterogeneity

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

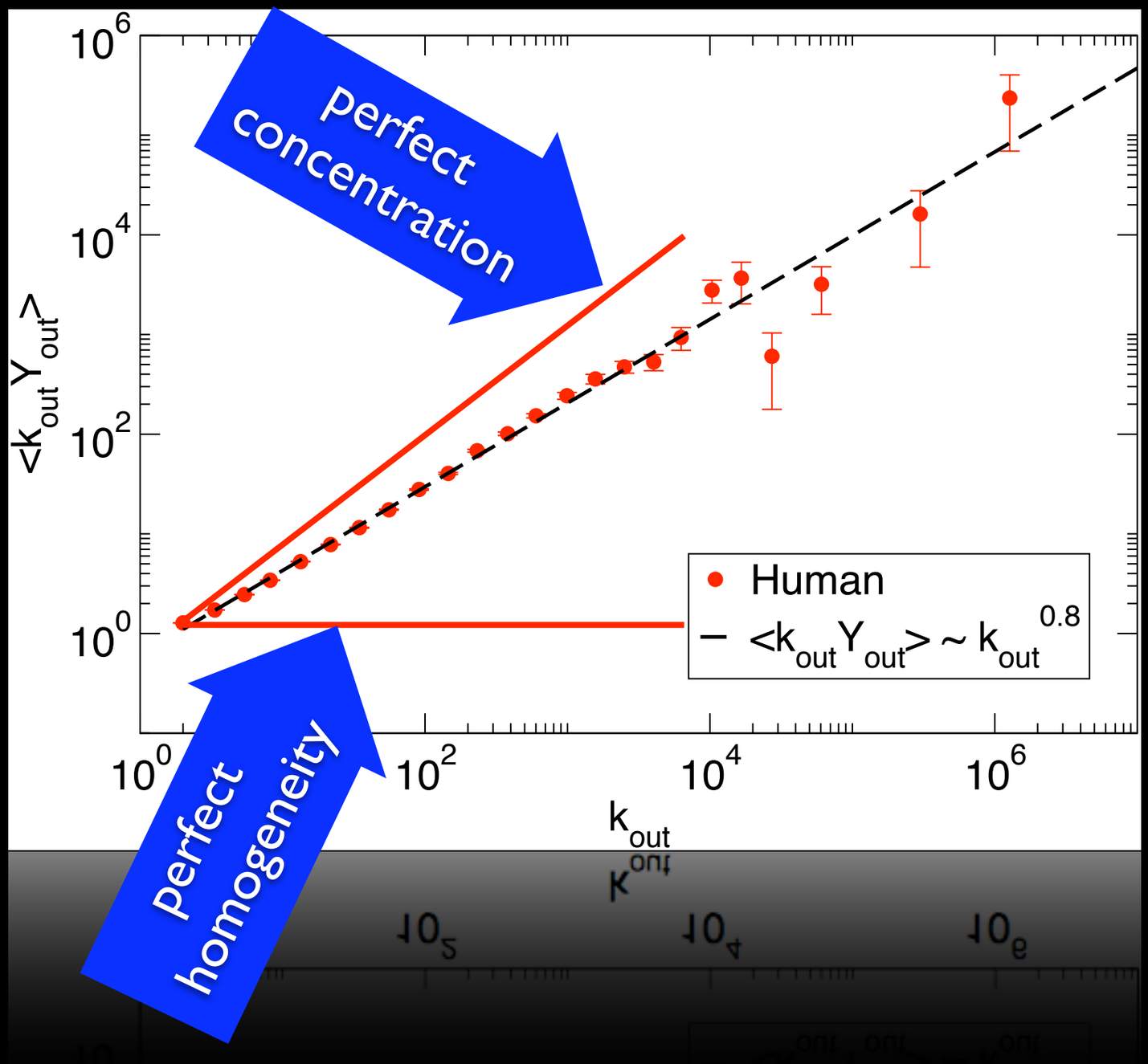
HH Index of concentration or disparity



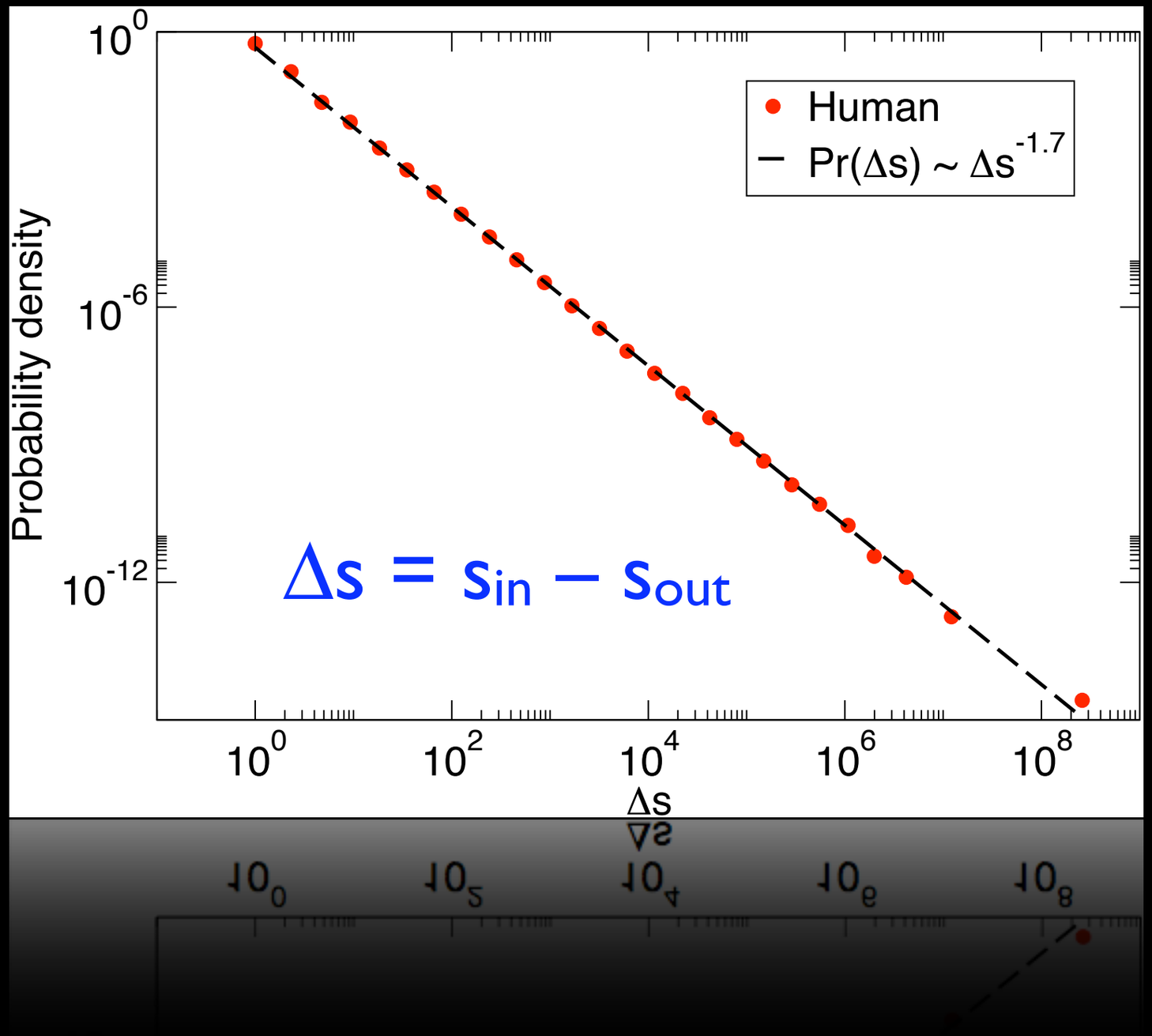
Local link heterogeneity

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

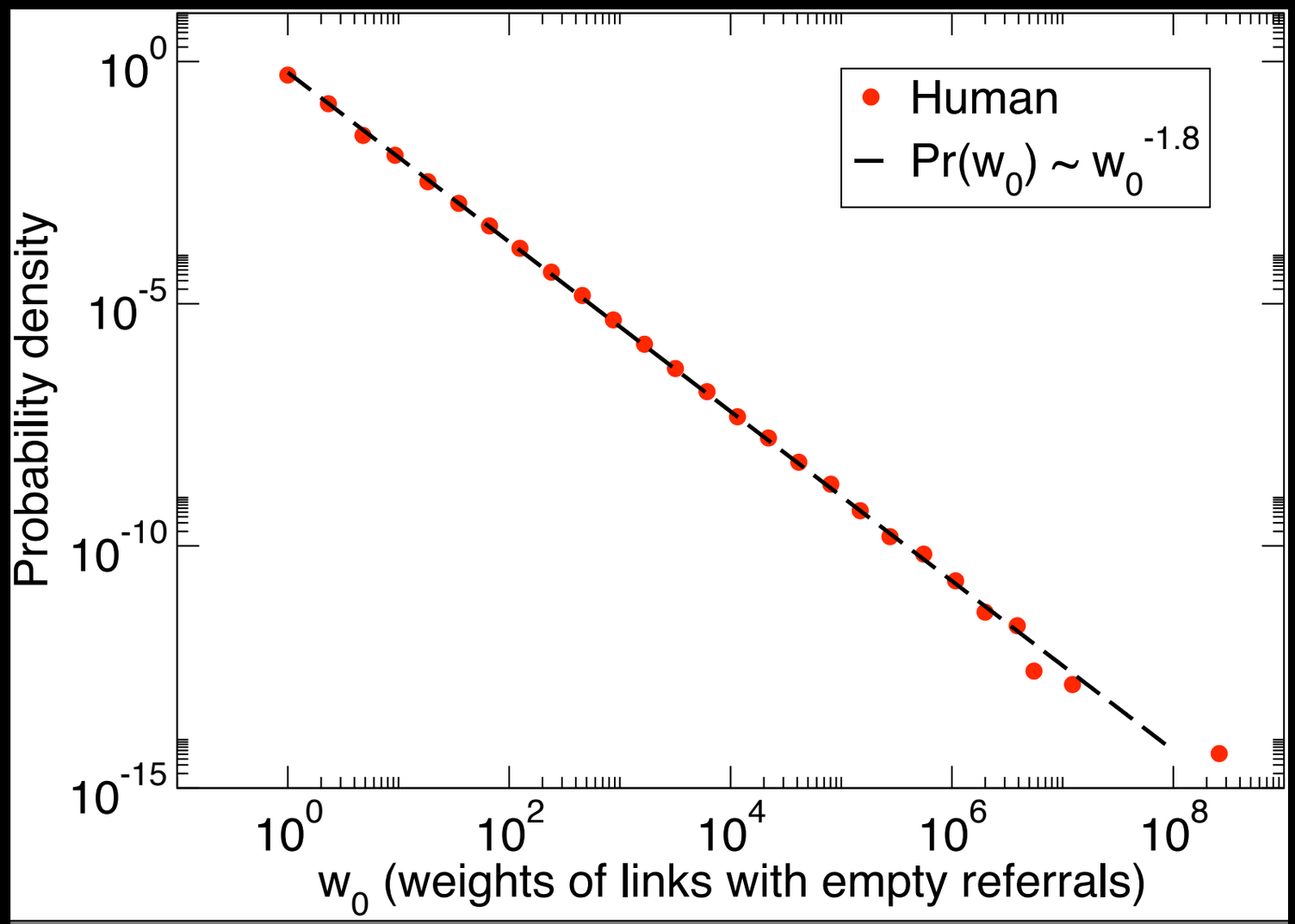
HH Index of concentration or disparity



Teleportation source heterogeneity



Teleportation target heterogeneity



Summary

- Heterogeneity: incoming and outgoing site traffic, link traffic
- Less than half of traffic is from clicks
- Only 5% directly from search engines
- Temporal regularity
- PageRank is a poor predictor of traffic: random walk and random teleportation assumptions violated

Next

- Sampling bias and search bias
- Modeling traffic: Beyond random walk?
- From host graph to page graph
- ClickRank

Thanks

Mark Meiss

Advanced Network Management Lab
Pervasive Technology Labs at Indiana University

Filippo Menczer

Santo Fortunato

Alessandro Flammini

Alessandro Vespignani

Indiana University School of informatics

CNLL
Progetto Lagrange