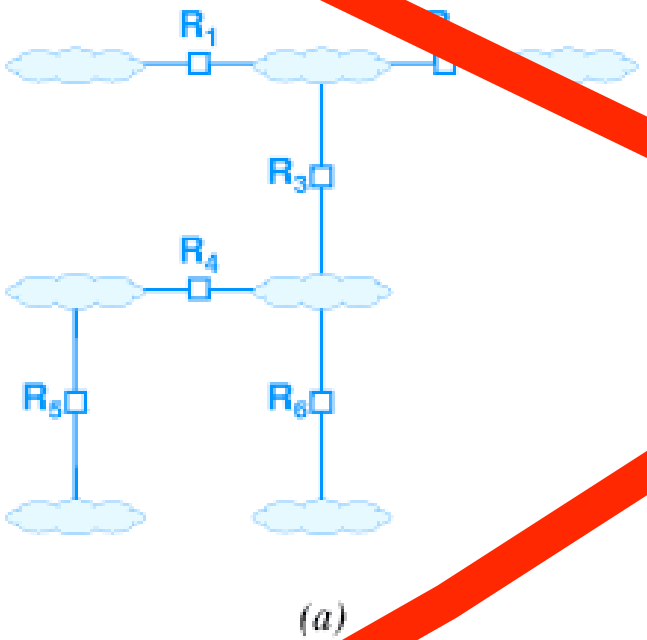
A day of Internet life

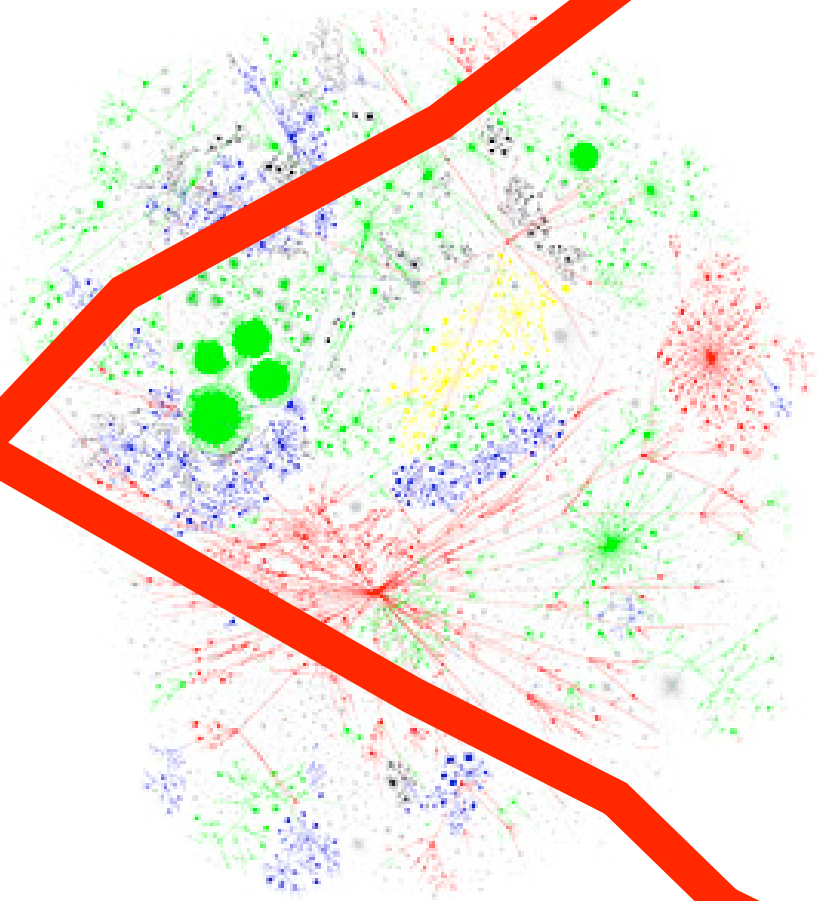
Mark Meiss[1,3], Filippo Menczer[1,2], Alessandro Vespignani[2]

(1) Department of Computer Science
(2) Department of Informatics
(3) Advanced Network Management Lab
Indiana University, Bloomington

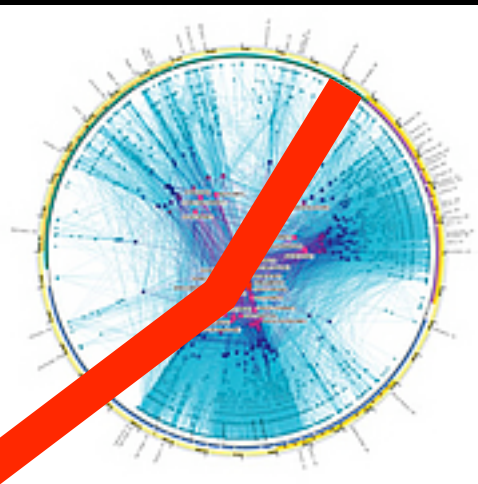
“The Internet”?

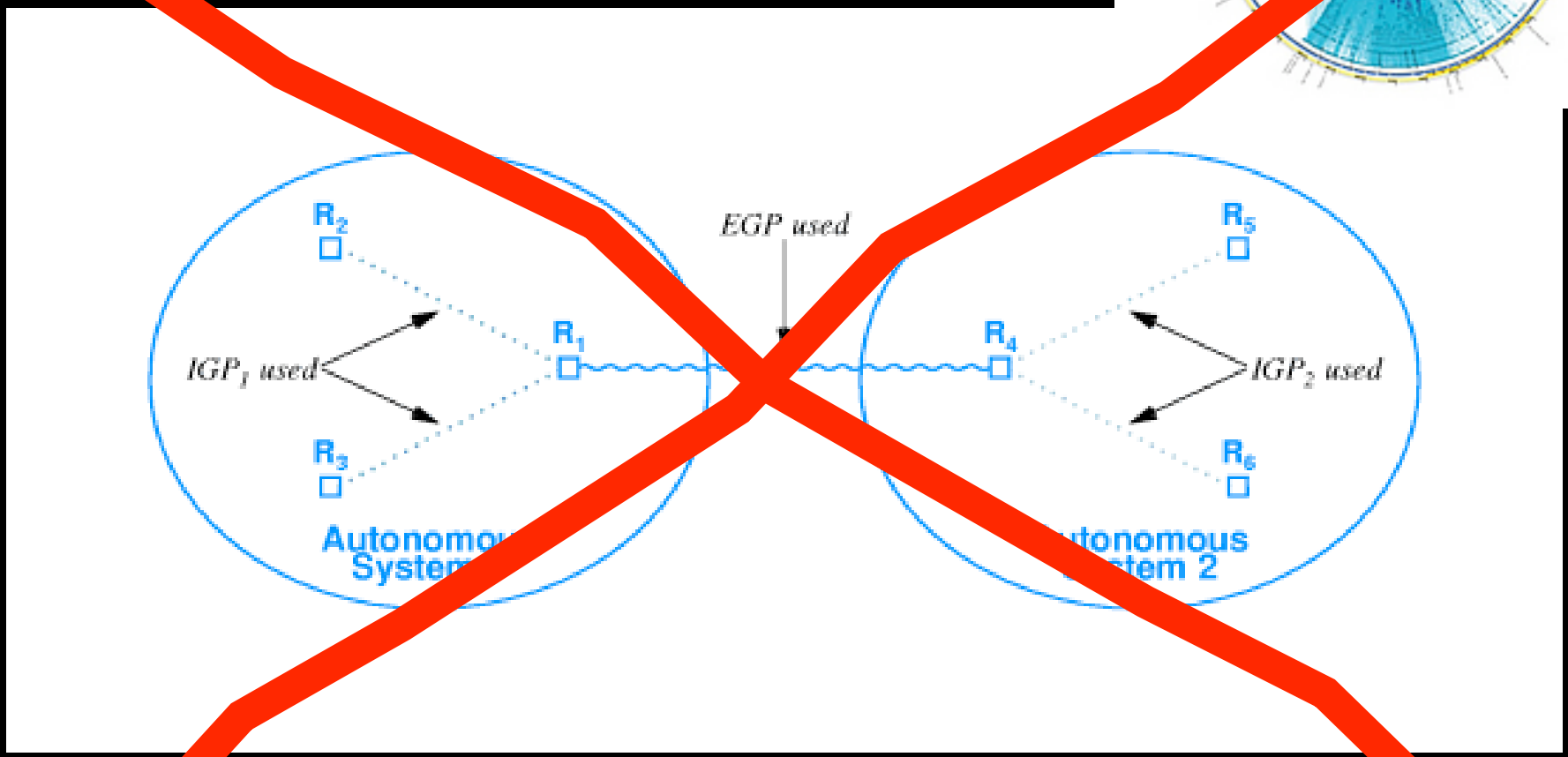


(a)

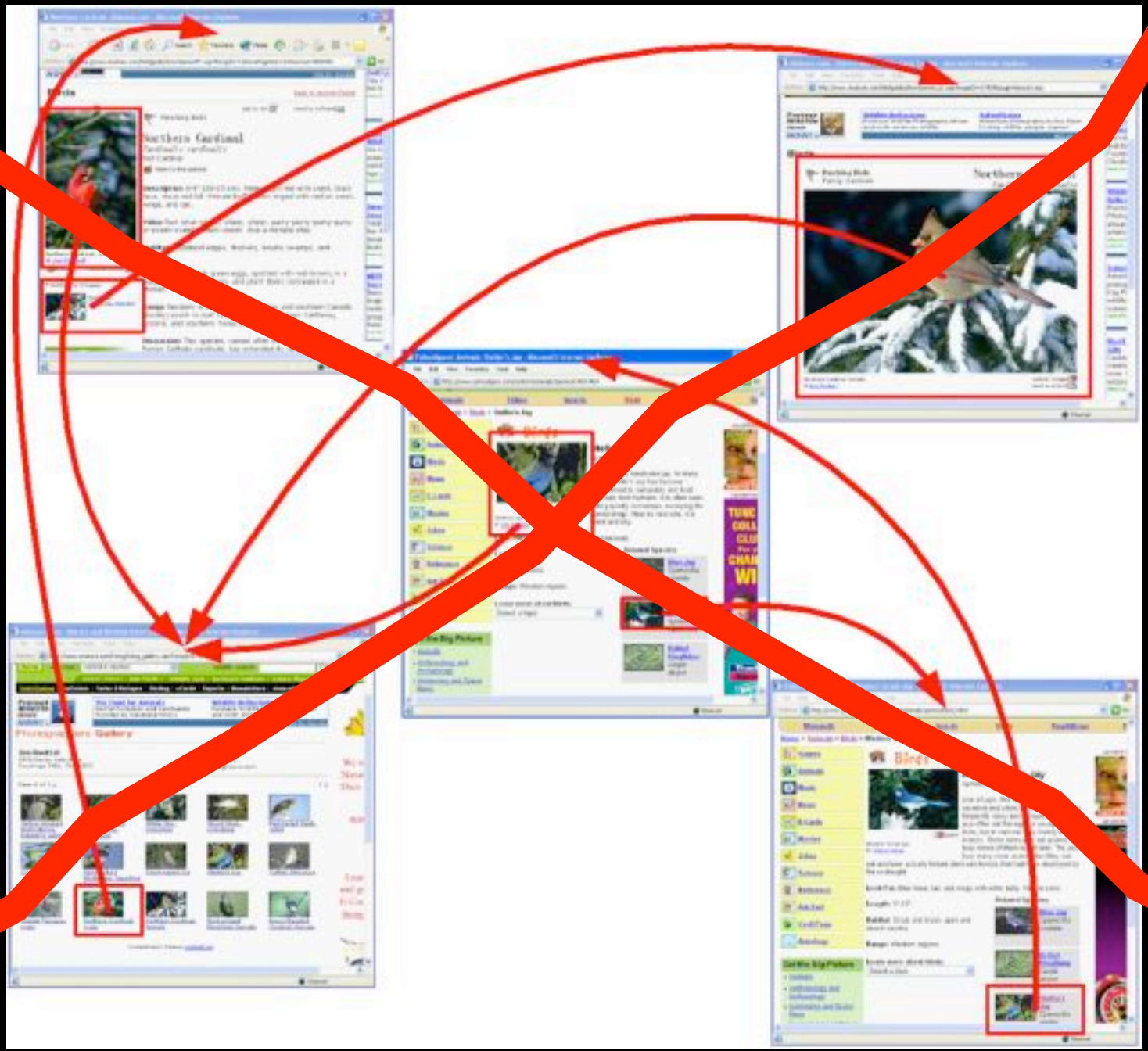


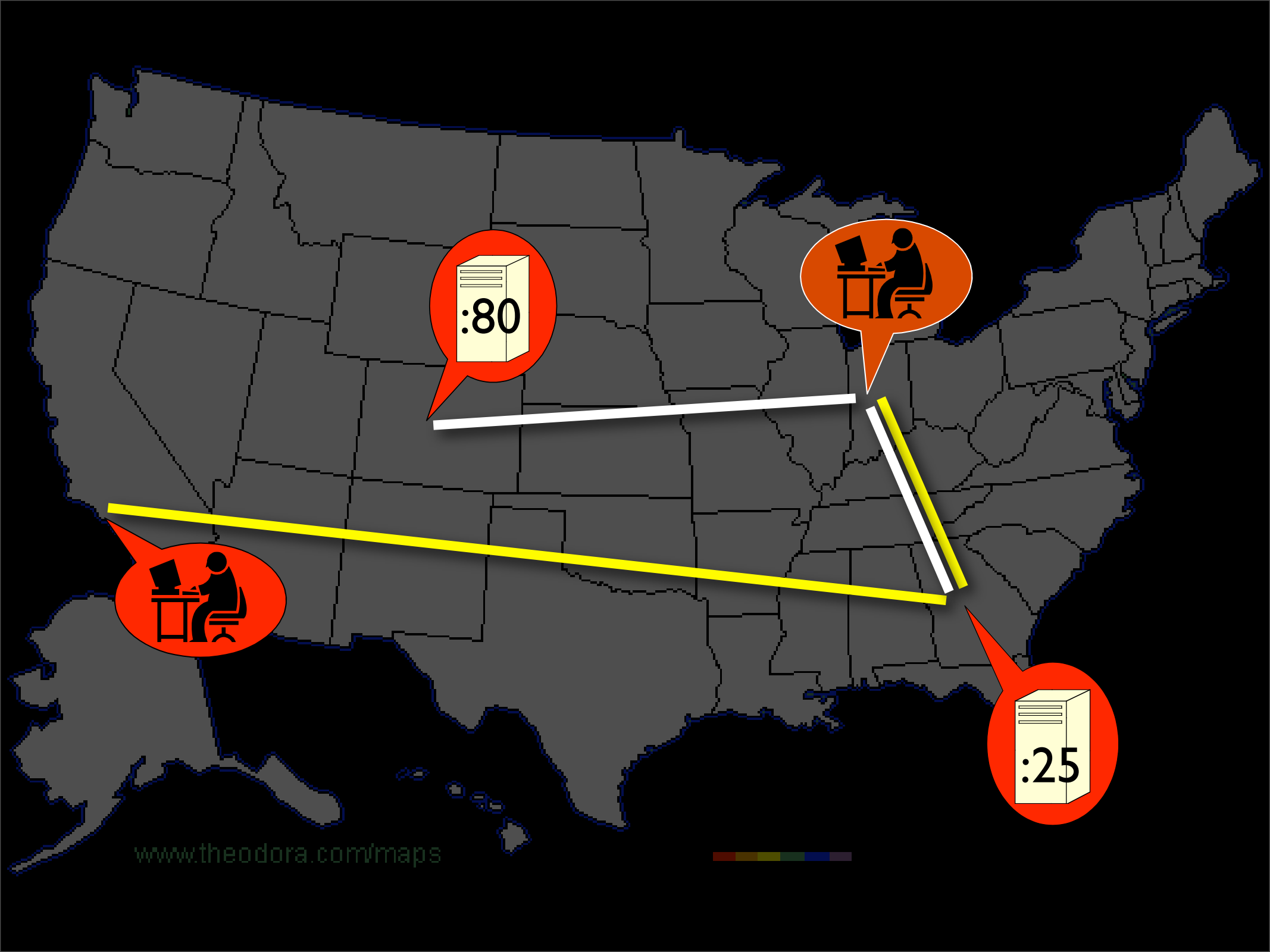
“The Internet”?





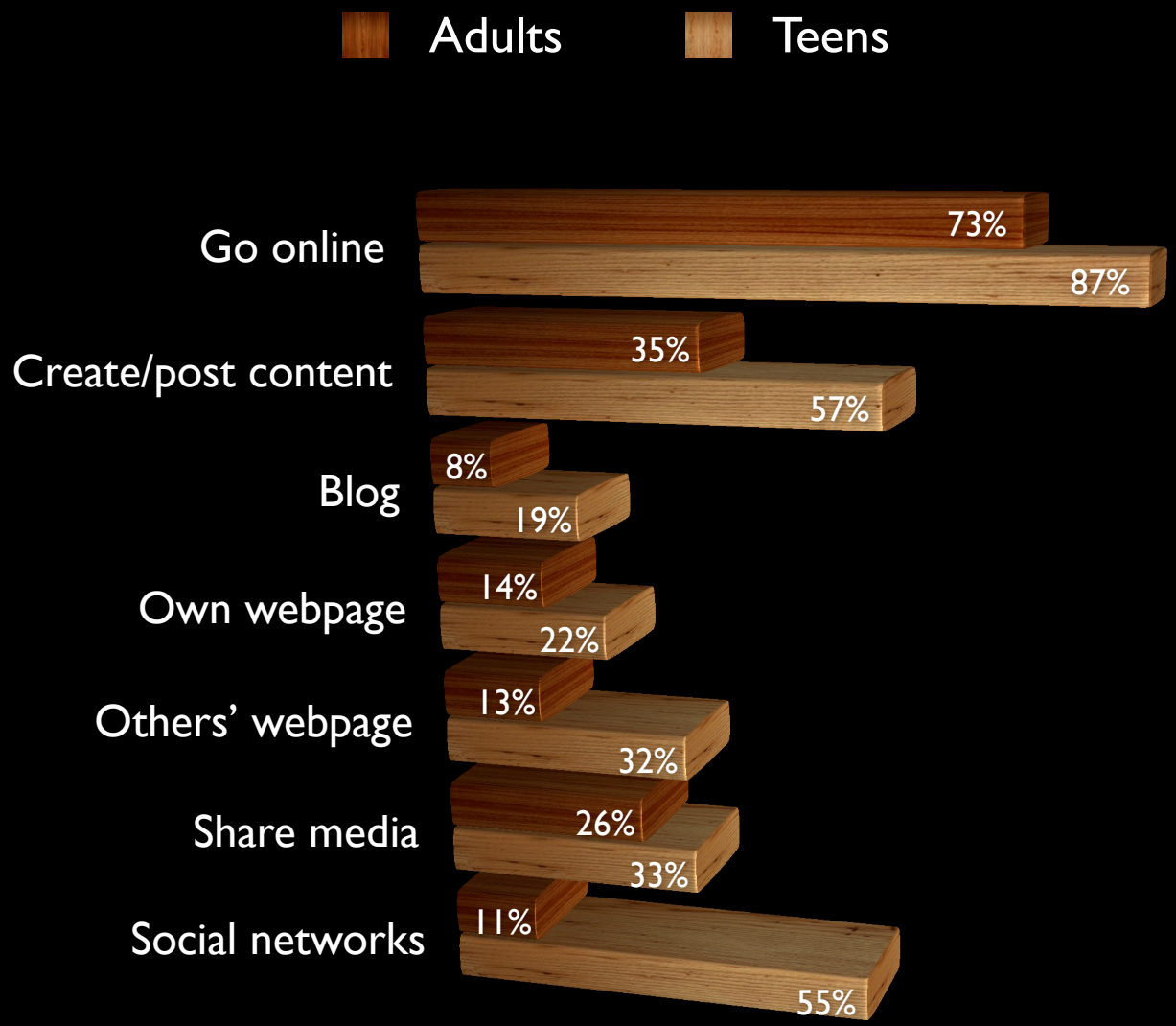
“The Internet”?




:80
:25
www.theodora.com/maps

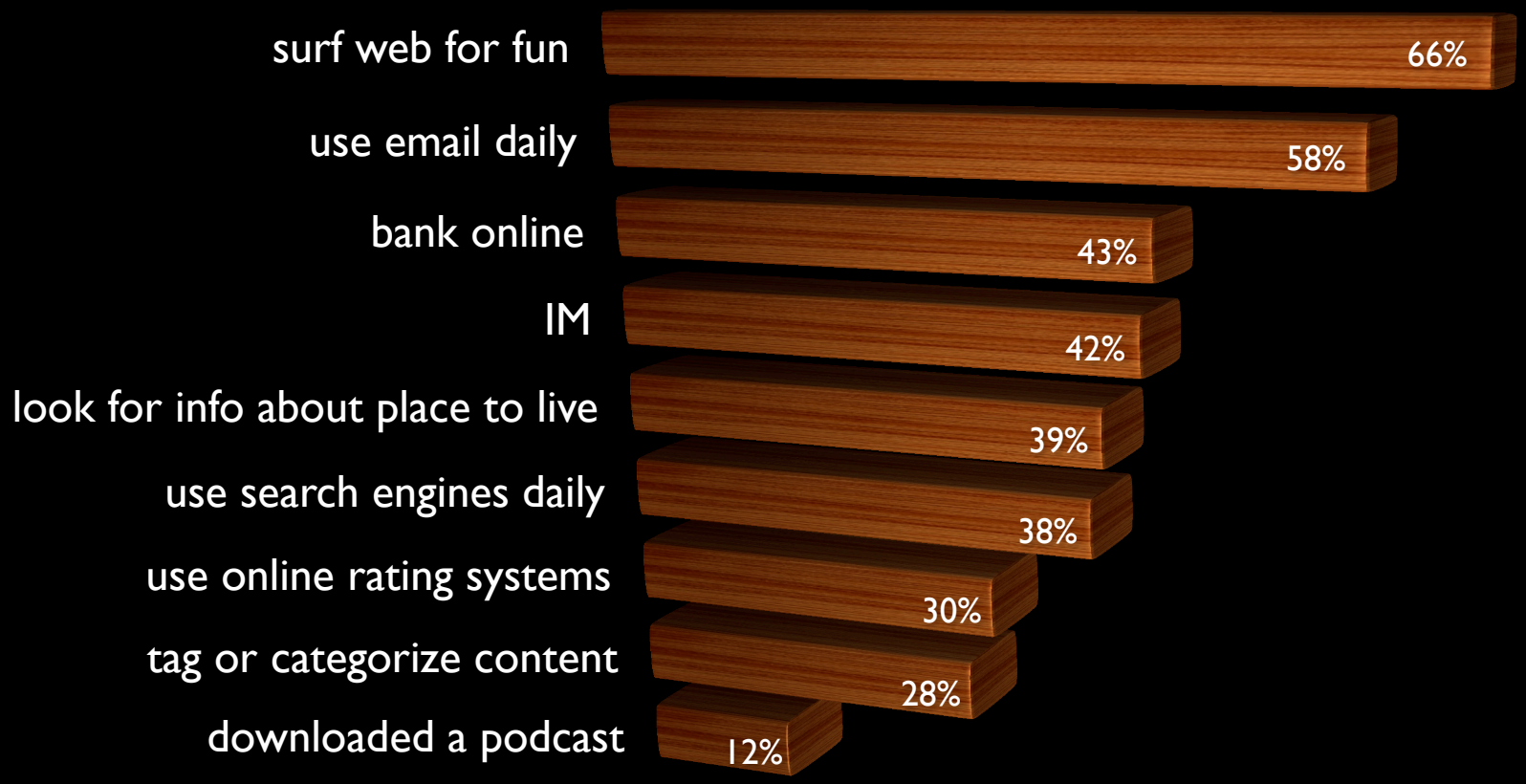
What do we do online?

- 60% of home internet users have broadband access
- 34% of internet users have wireless access



Source: PIP (pewinternet.org)

More stats...



Source: PIP (pewinternet.org)

What's missing

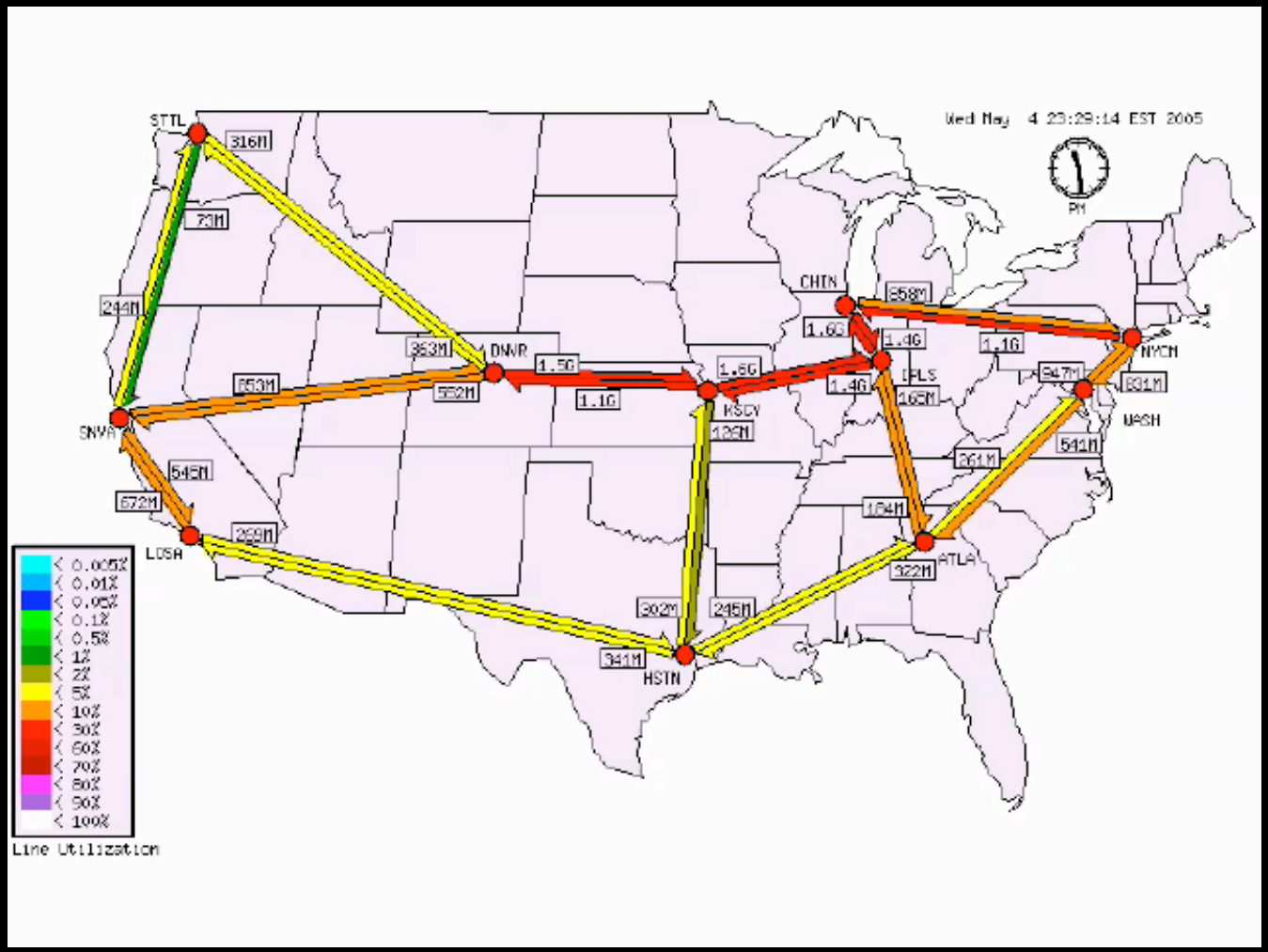
- Distributional information, correlations, dependencies
- Who talks to whom, how, how much...?
- How do these activities impact Internet traffic?
- Can we characterize patterns of traffic?
- Can we deduct what people are doing from looking at the traffic?
  - Covert activities?
  - Without looking at IP addresses, or inspecting payloads, or looking at all packets?

Traffic networks

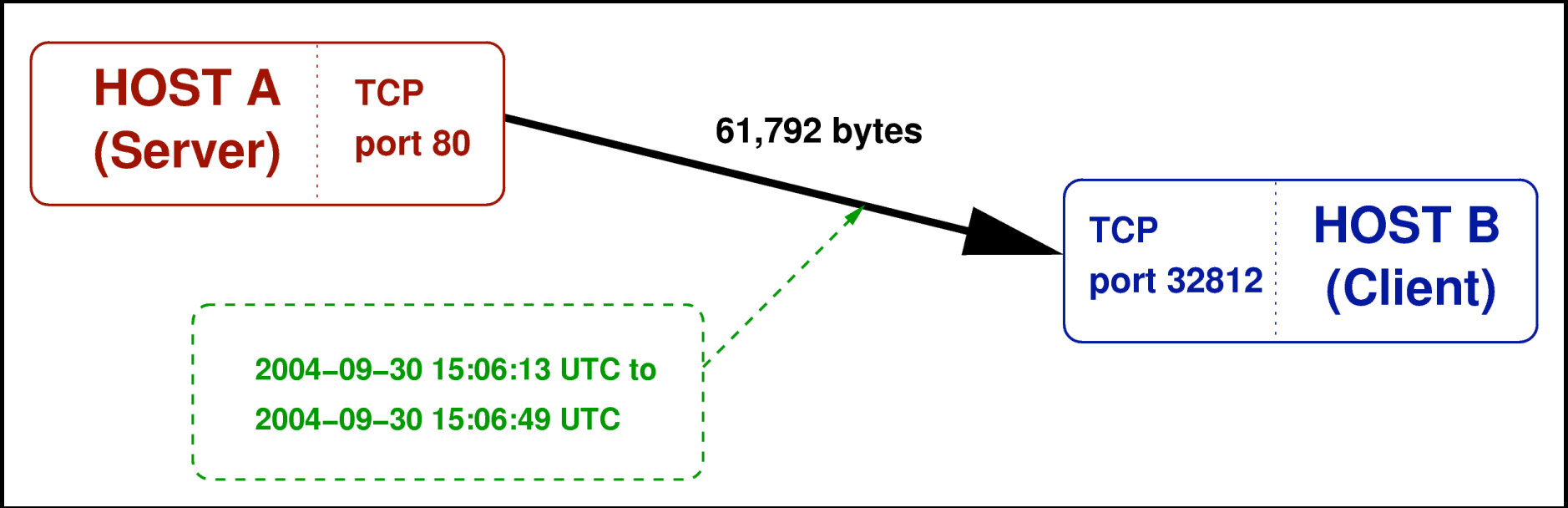
- Build networks from Internet traffic data
- Two proofs of concept:
  - Behavioral networks (host to host)
  - Application networks (app to app)

Internet2/Abilene

- TCP/IP network connecting research and educational institutions in the U.S.
  - Over 200 universities and corporate research labs
  - Hundreds of thousands of undergraduates
- Also provides transit service between Pacific Rim and European networks
- High capacity (never congested)
- Now peered with commodity Internet

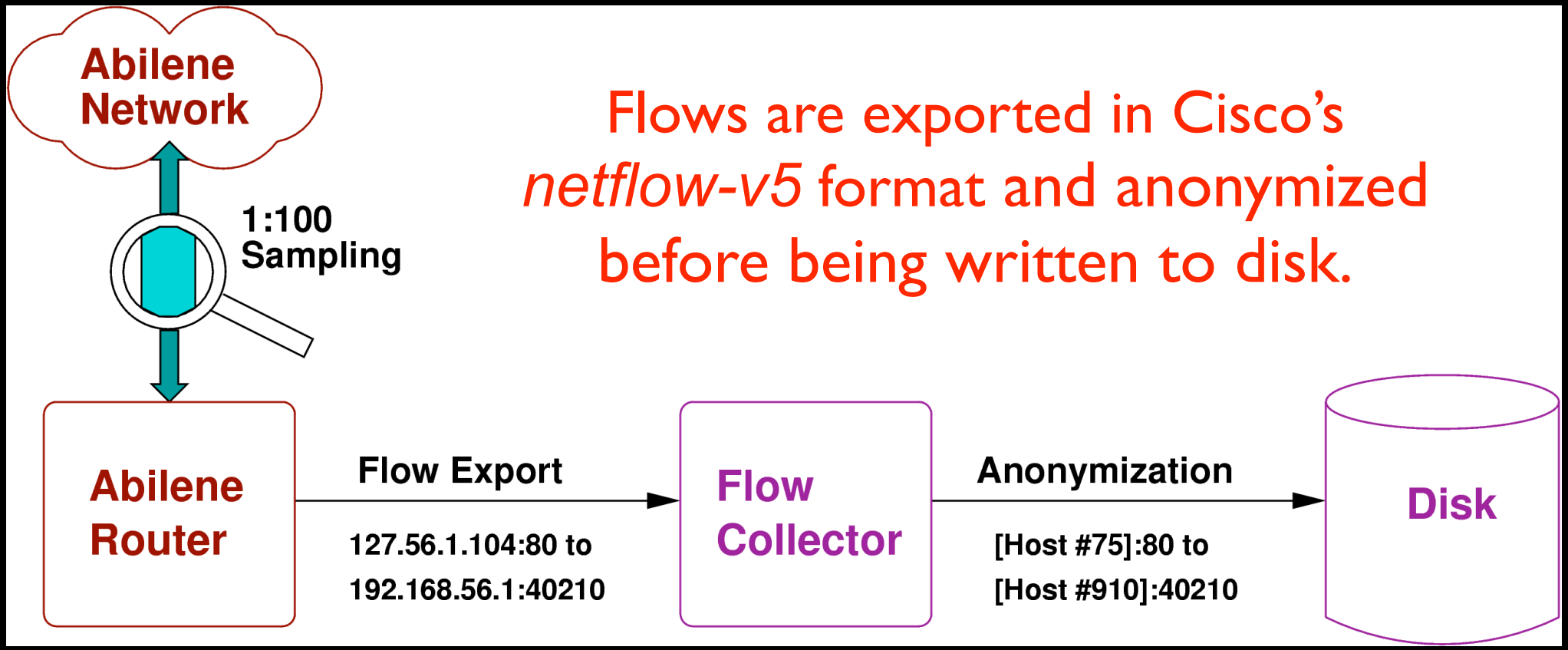


Network flow data



A successful TCP session contains ***two*** flows

Flow collection

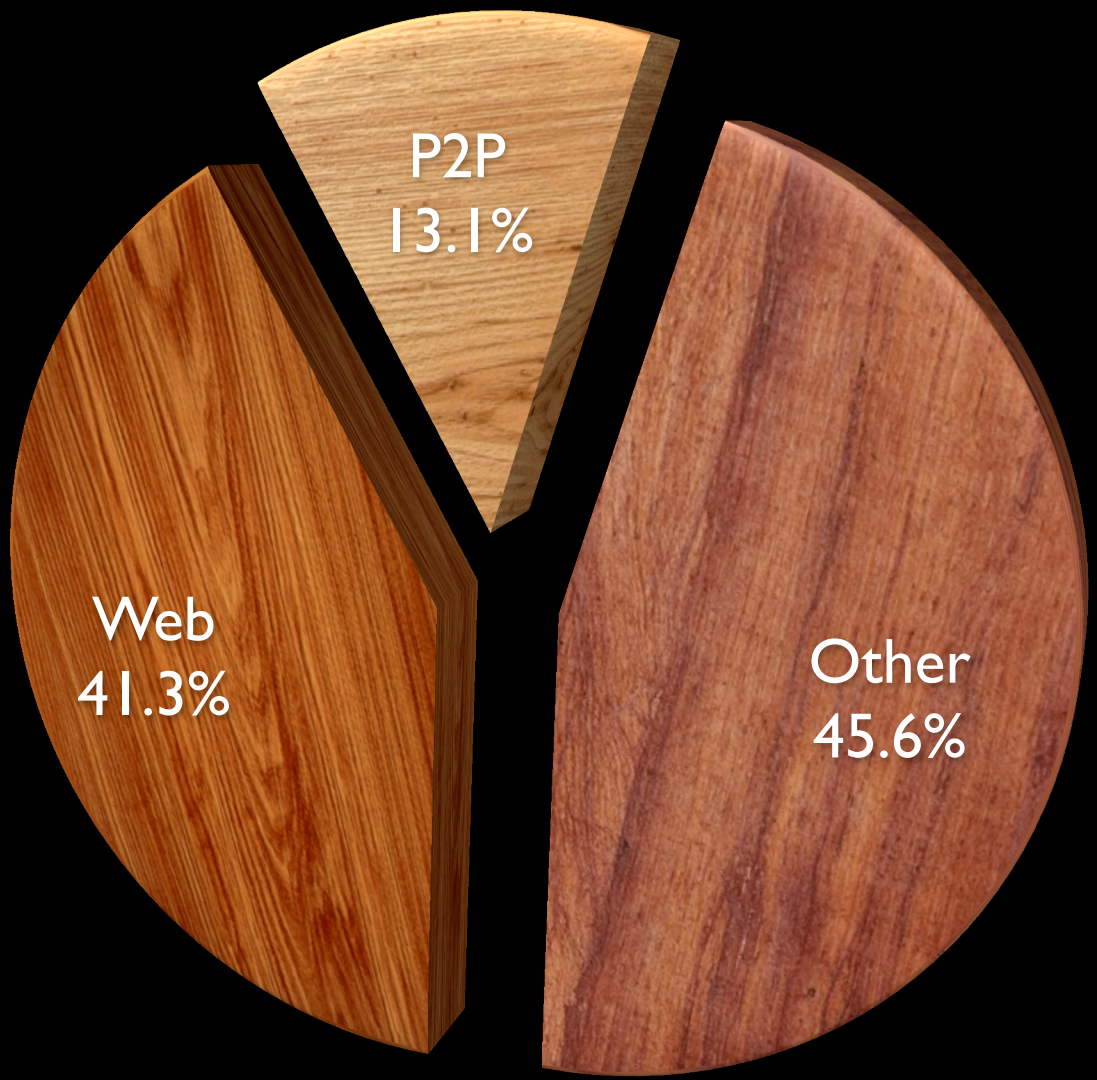


In a typical day: 35GB raw data at 3.1 Mbps

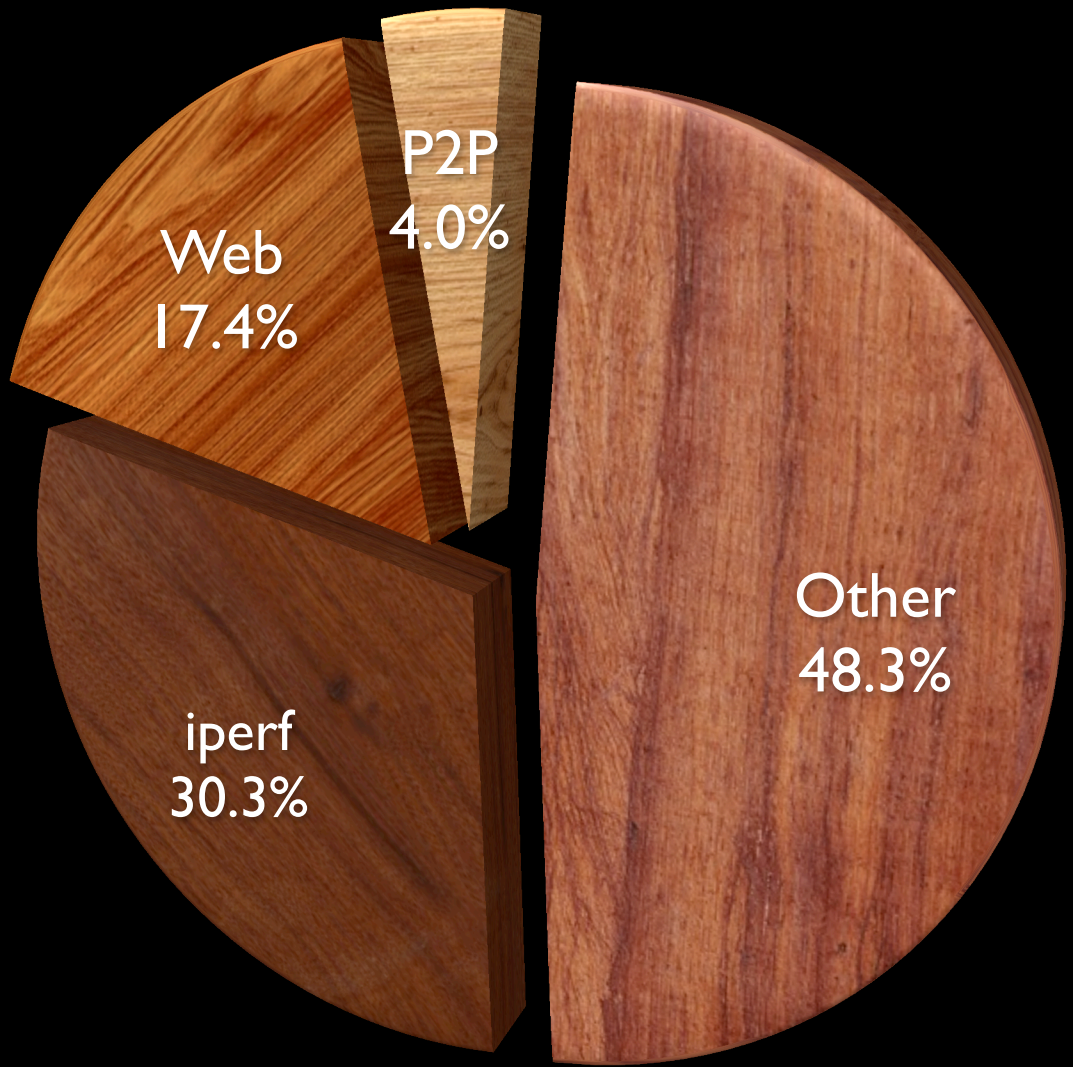
Data set for analysis

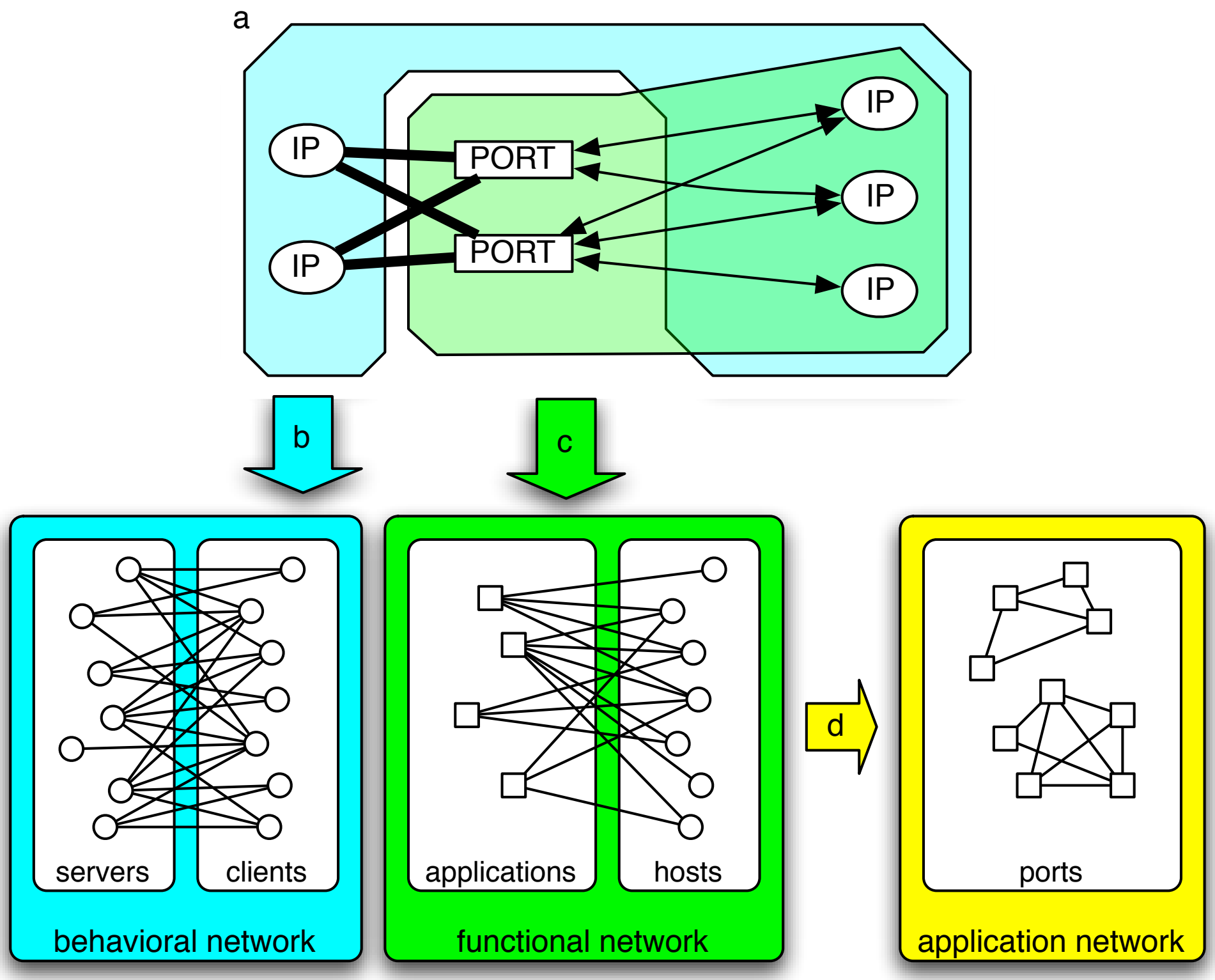
- Full 24-hour day of network flow data starting at 2005-04-14 05:00:00 UTC
  - a typical day
  - 600M flows
  - 15M unique hosts
  - 2TB (1% sampling)

Flows

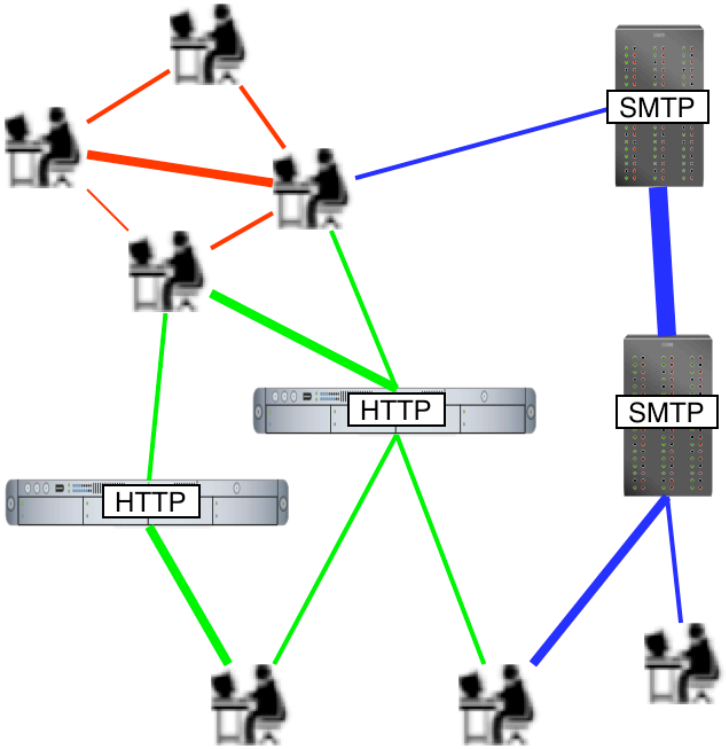


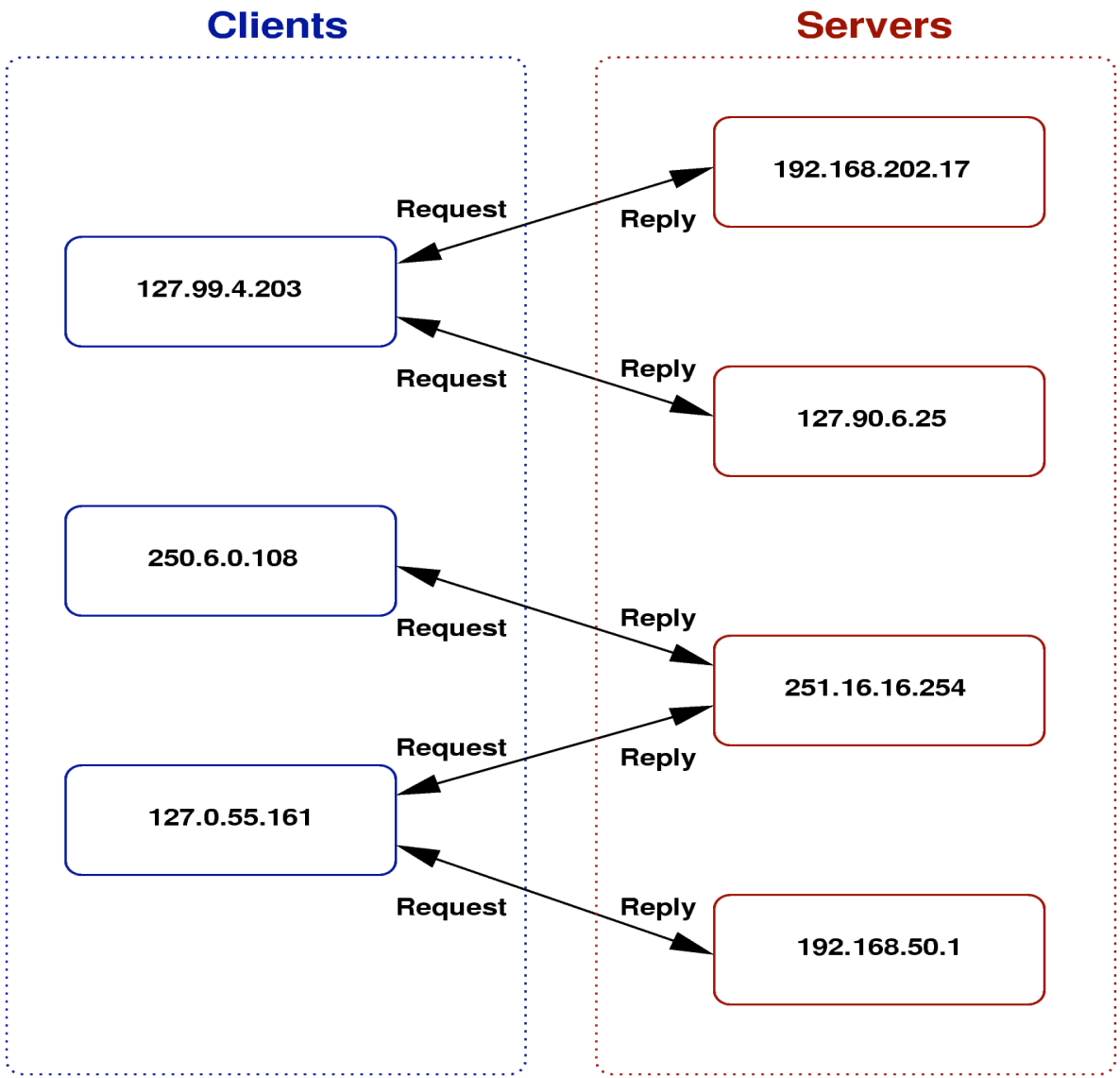
Traffic




a
IP
PORT
PORT
IP
IP
IP
b
c
servers
clients
behavioral network
applications
hosts
functional network
d
ports
application network

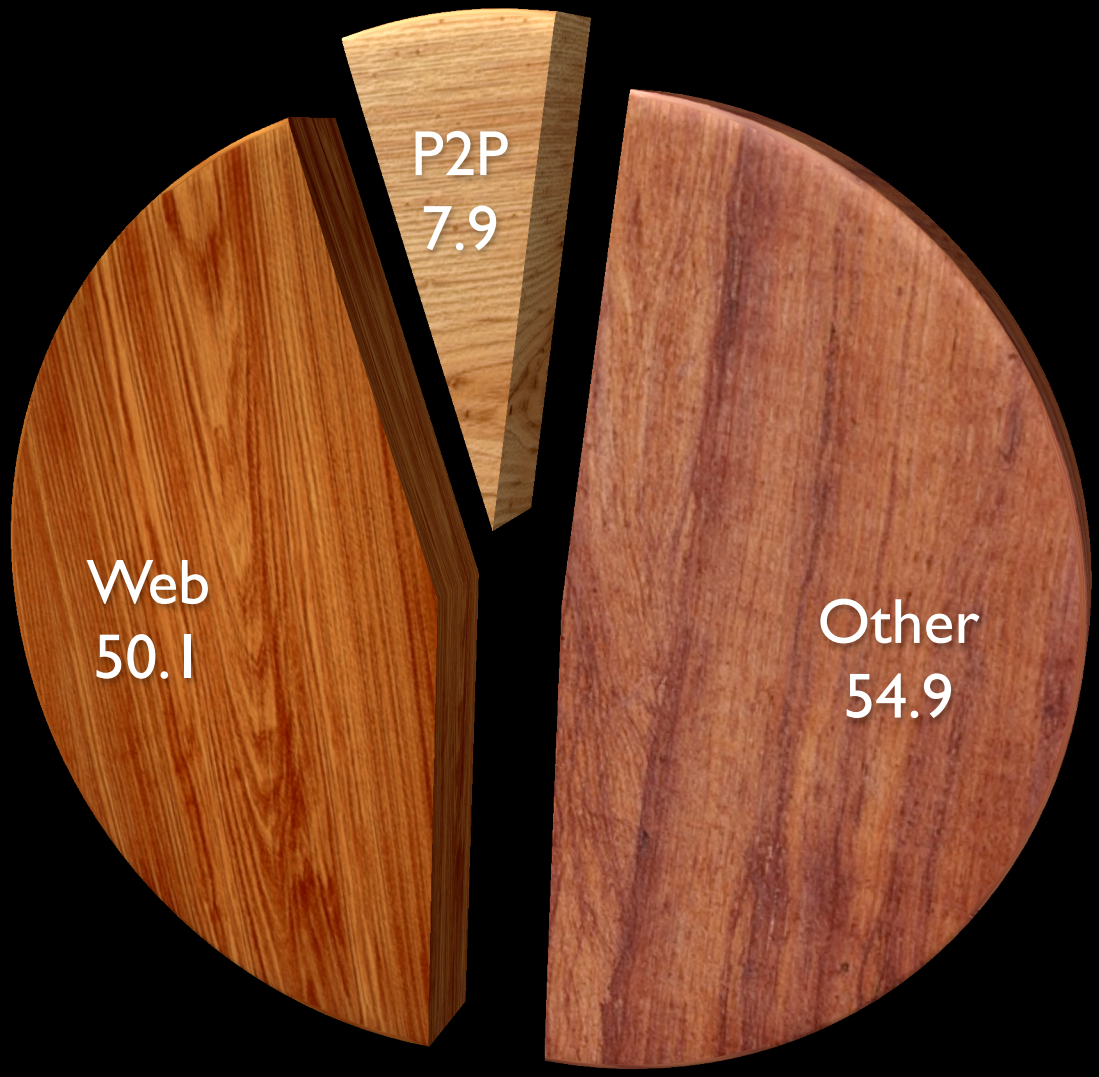
Behavioral networks

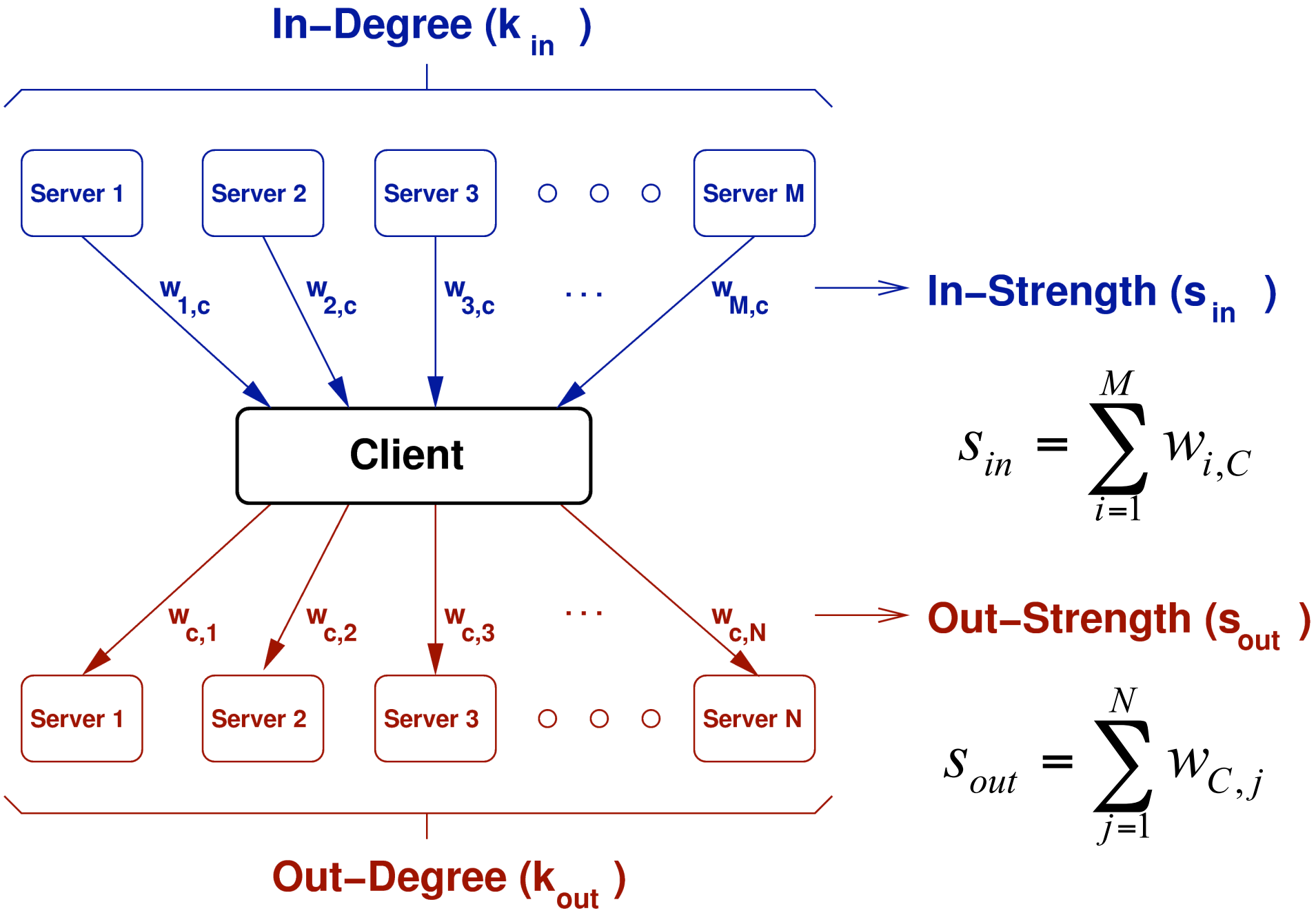




Karagiannis, Papagiannaki & Faloutsos 2005; Meiss, Menczer & Vespignani 2005

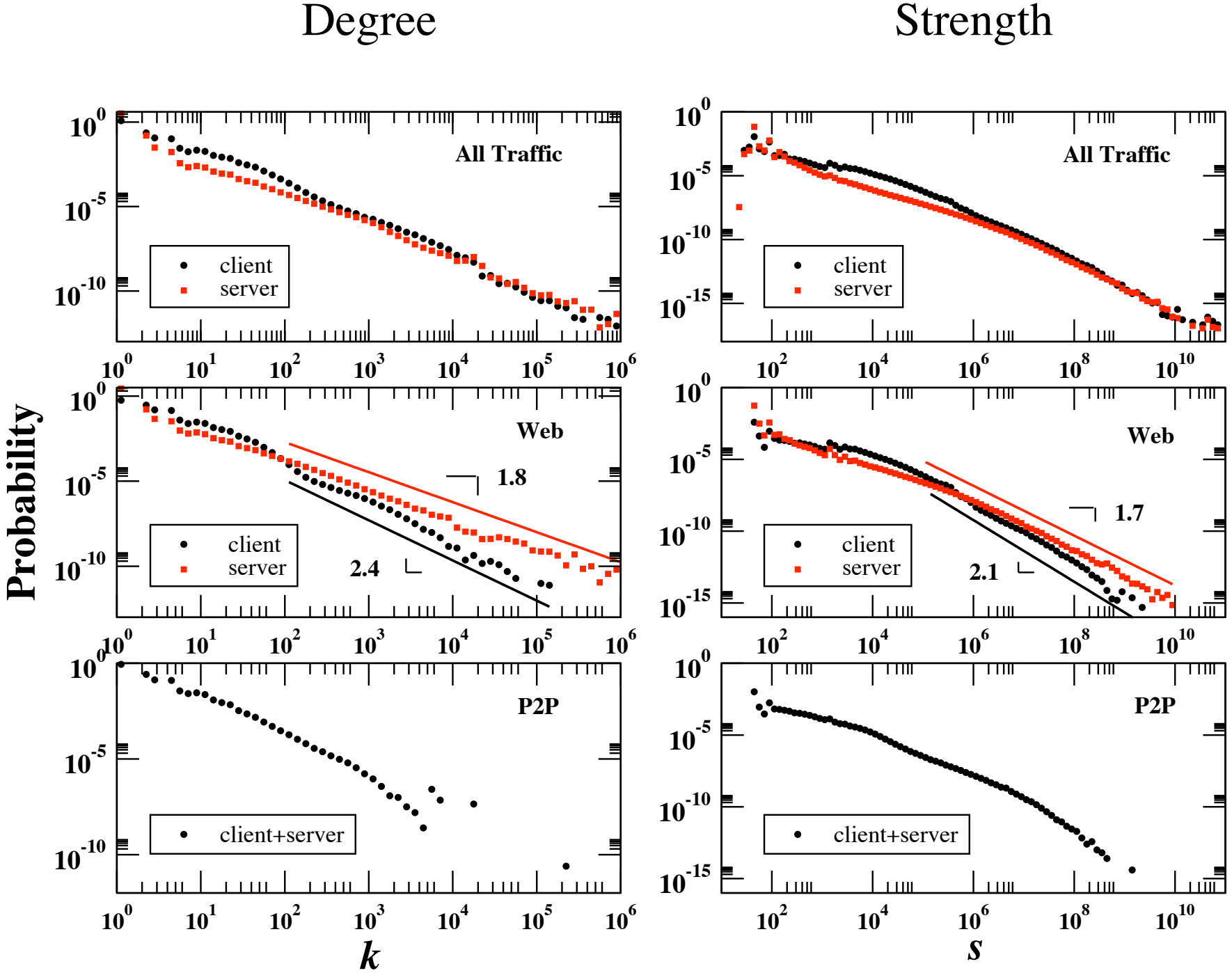
Edges ($\times 10^6$)



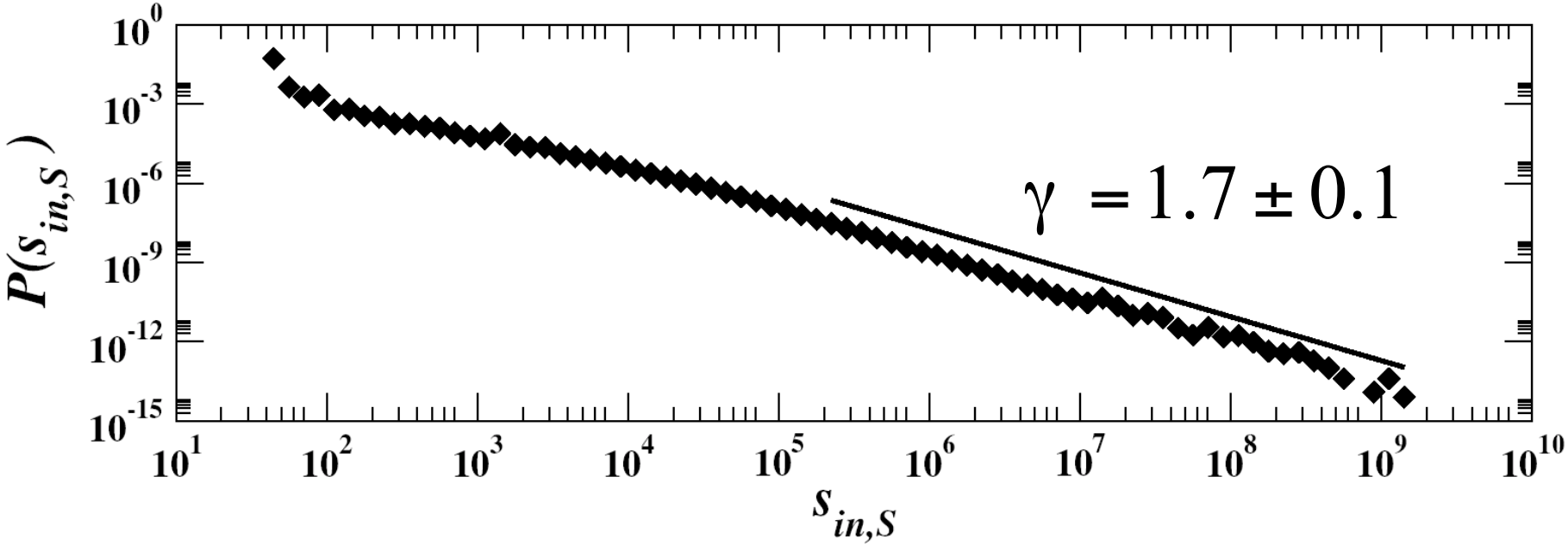


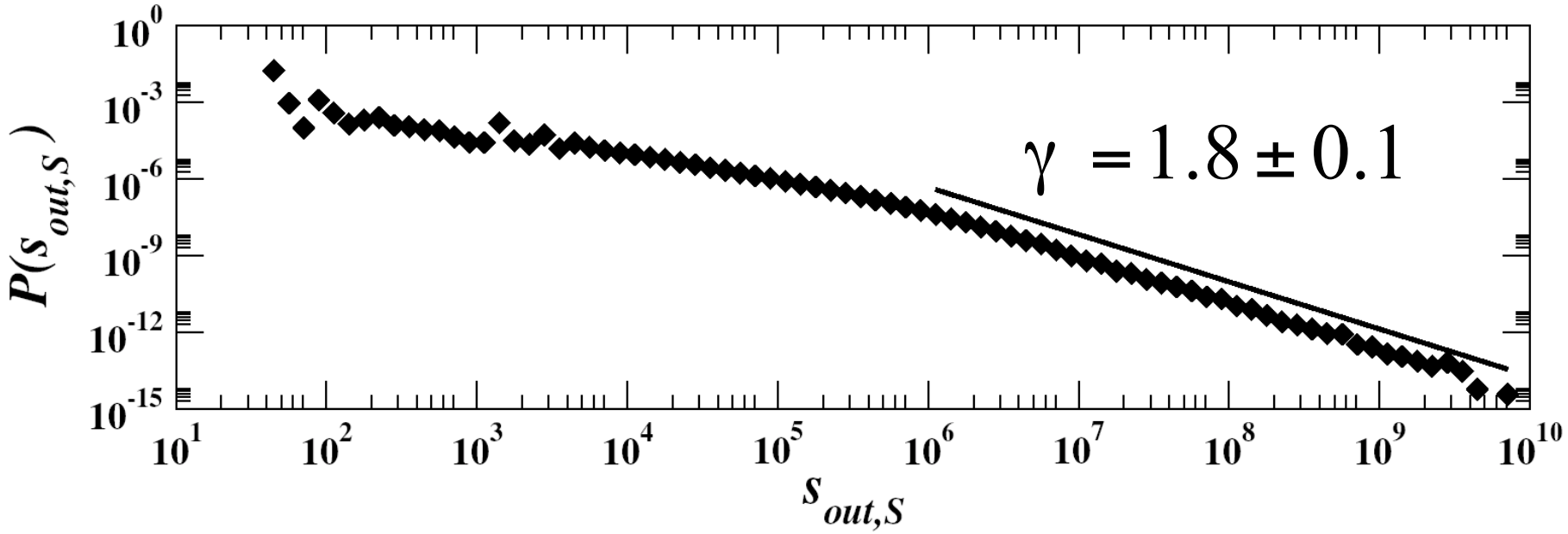
$$s_{in} = \sum_{i=1}^{M} w_{i,C}$$

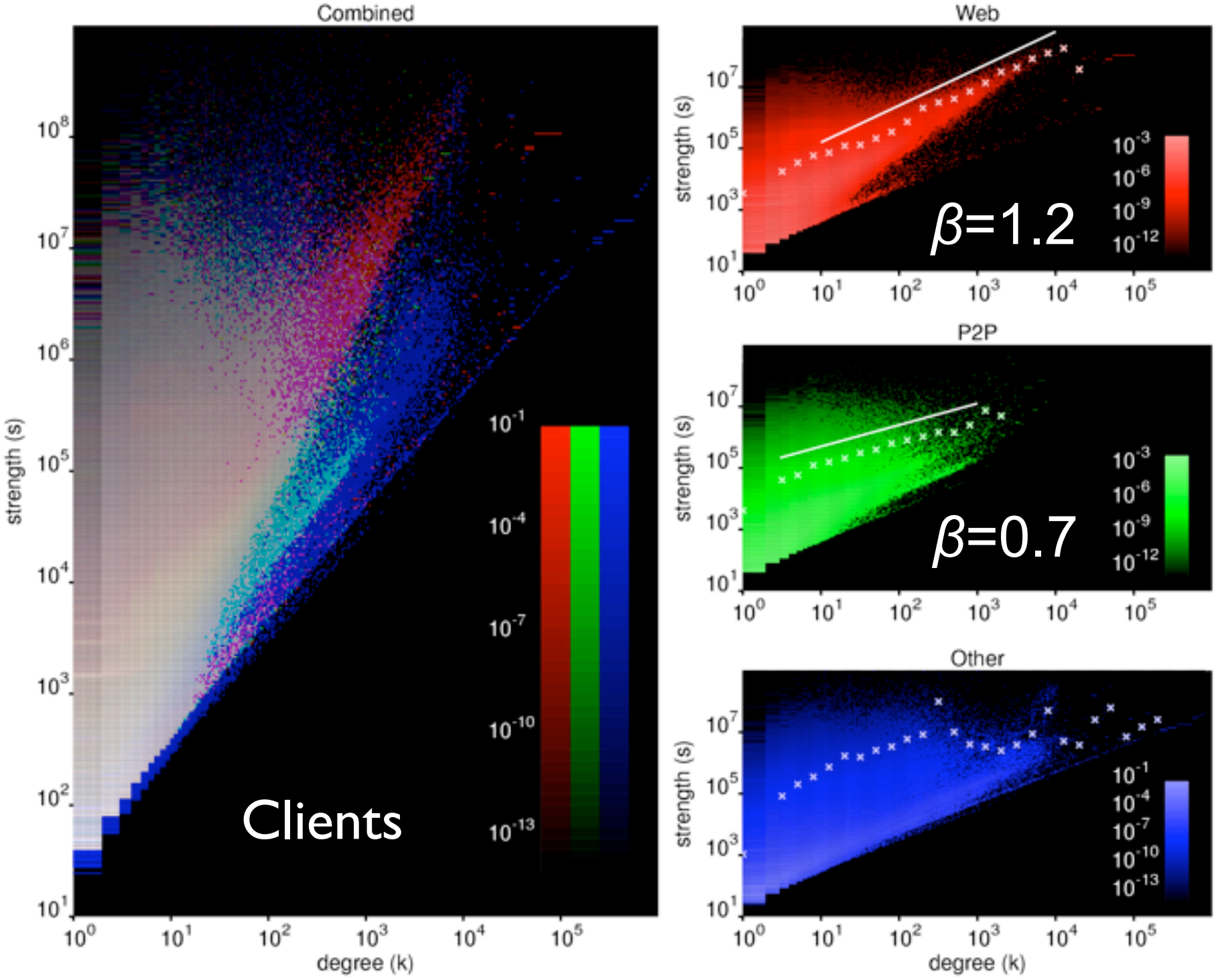
$$s_{out} = \sum_{j=1}^{N} w_{C,j}$$


Degree
Strength
Probability
All Traffic
client
server
Web
1.8
2.4
1.7
2.1
P2P
client+server
k
s

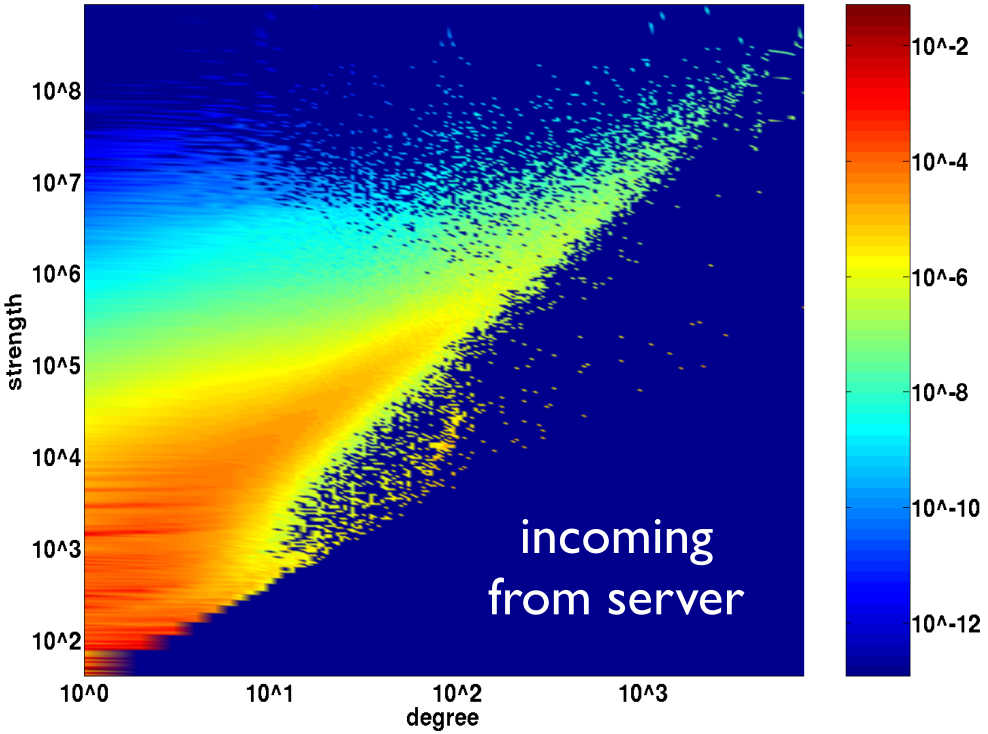
Web servers: Strength distributions

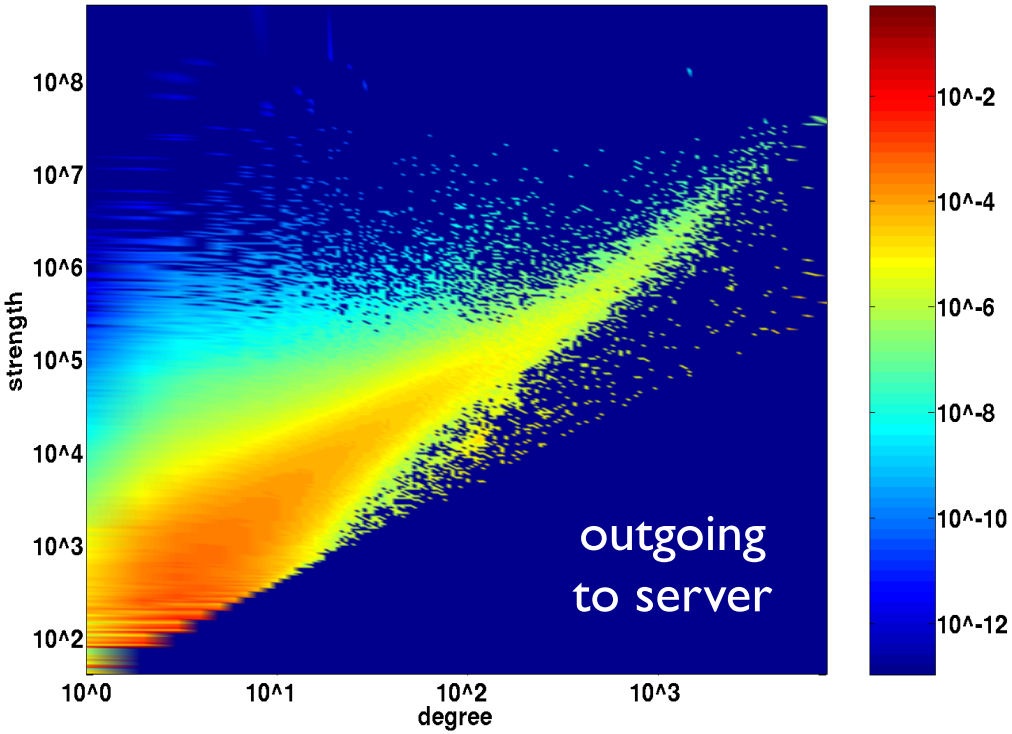


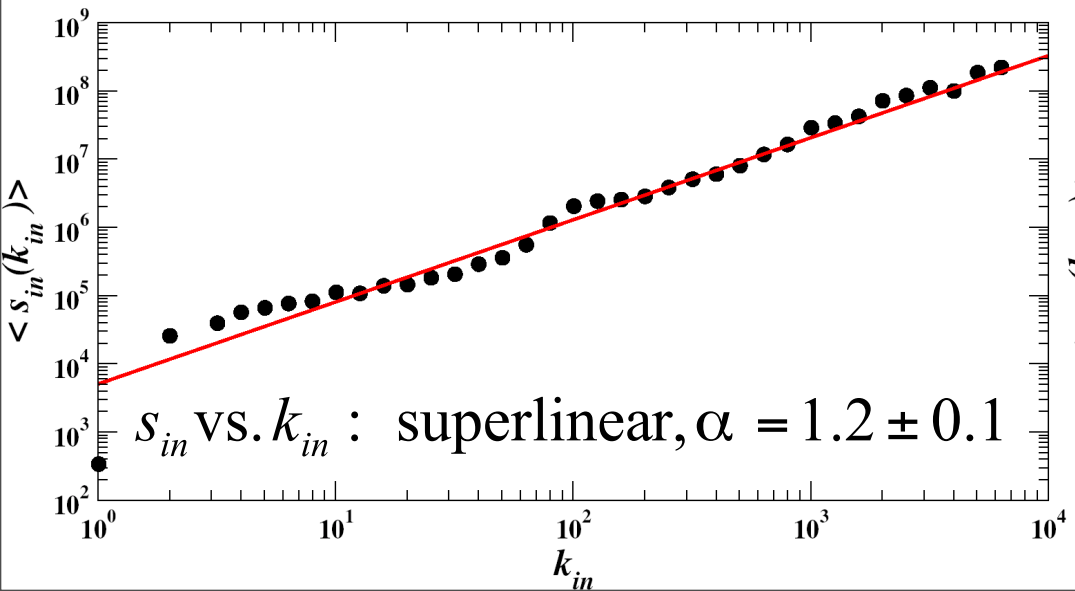


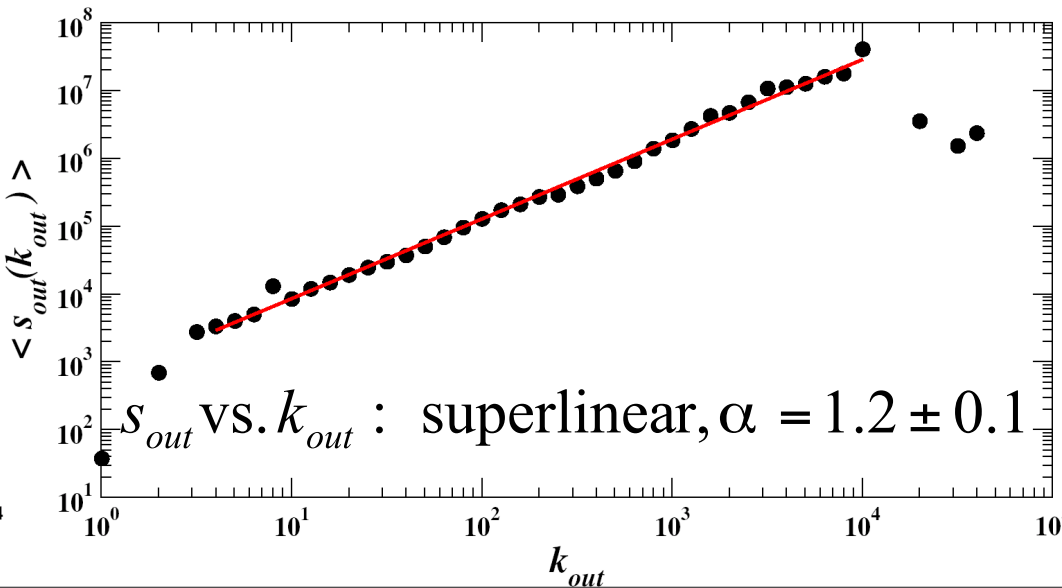

Combined
Web
P2P
Other
Clients
β=1.2
β=0.7
strength (s)
degree (k)

Web scalability problem: The more servers a client contacts, the more data it exchanges with **each** server!



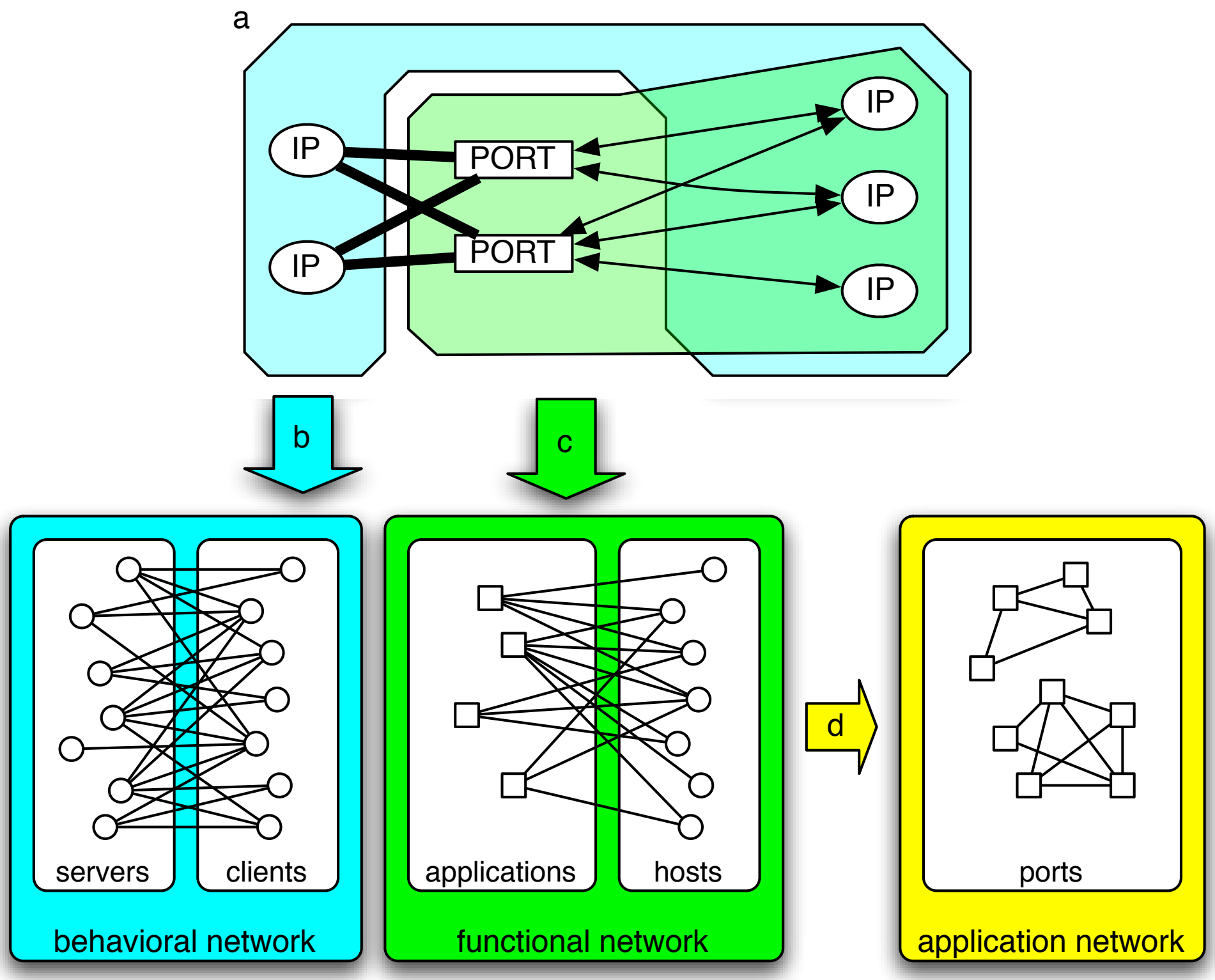




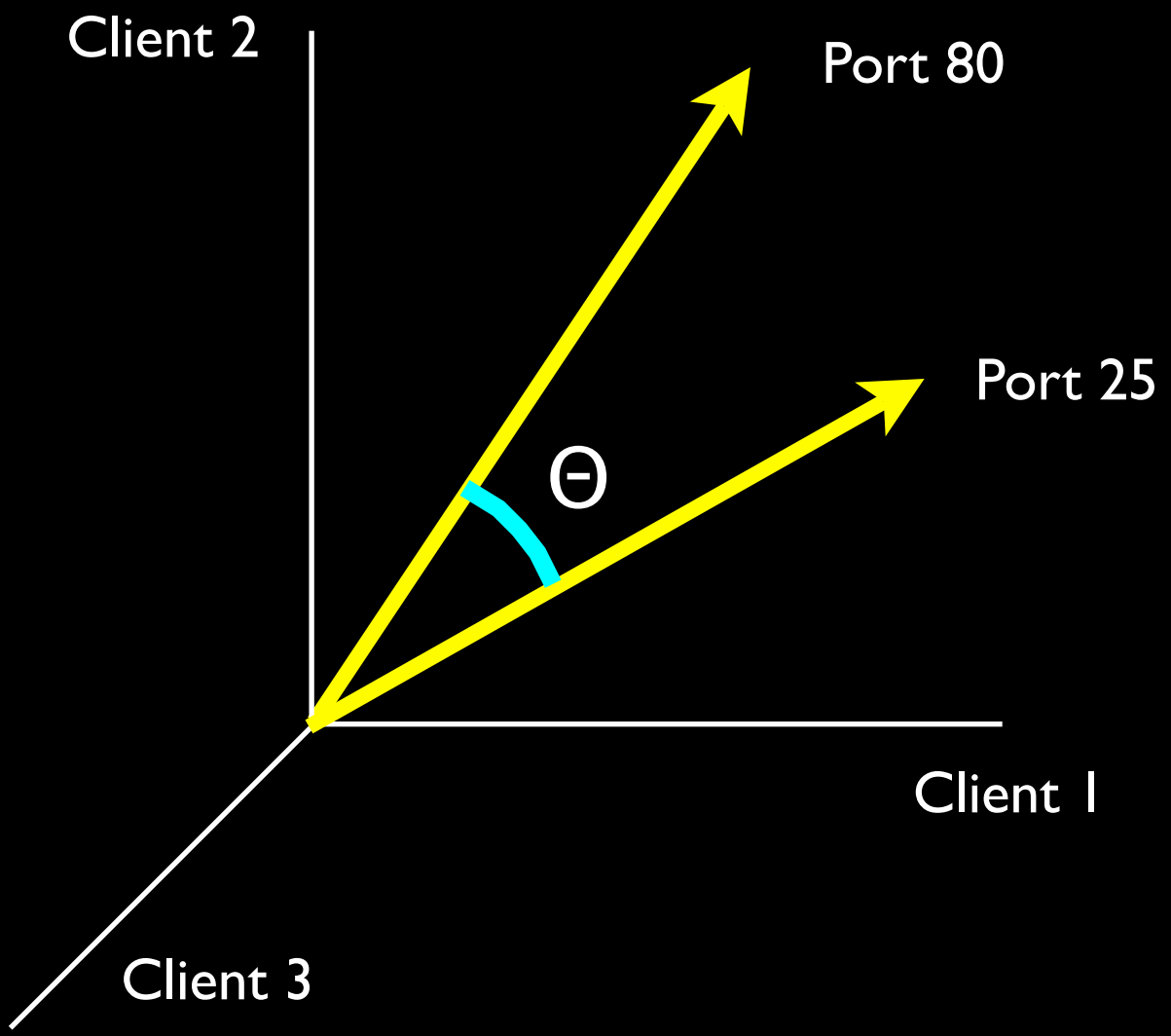


Summary

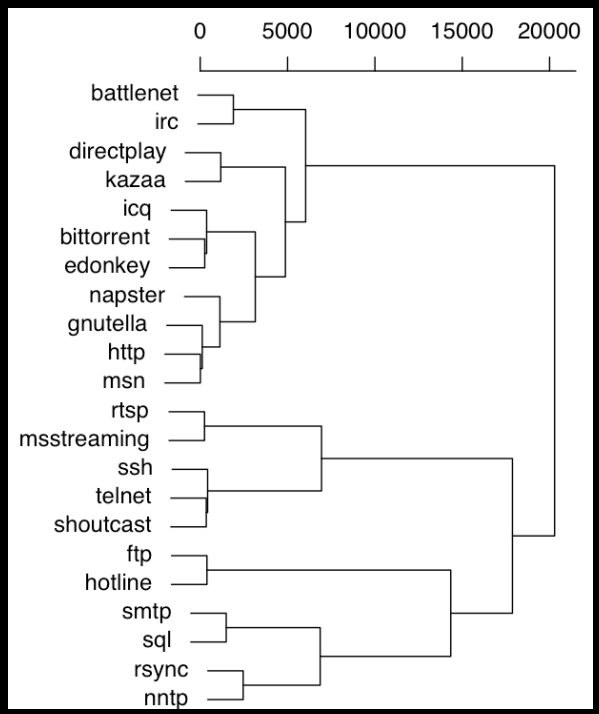
- ***Power-law distributions*** are found in all aspects of traffic networks: degree, strength, and weight distributions for both clients and servers
  - The ***strength*** distribution for Web ***servers*** lacks any ***mean value***
  - The relationship between ***degree*** and ***strength*** for Web clients is ***super-linear***
  - ***Models*** must be able to account for these heavy-tailed distributions and non-linear coupling
  - Classes of traffic can be characterized by their strength-degree ***signatures***

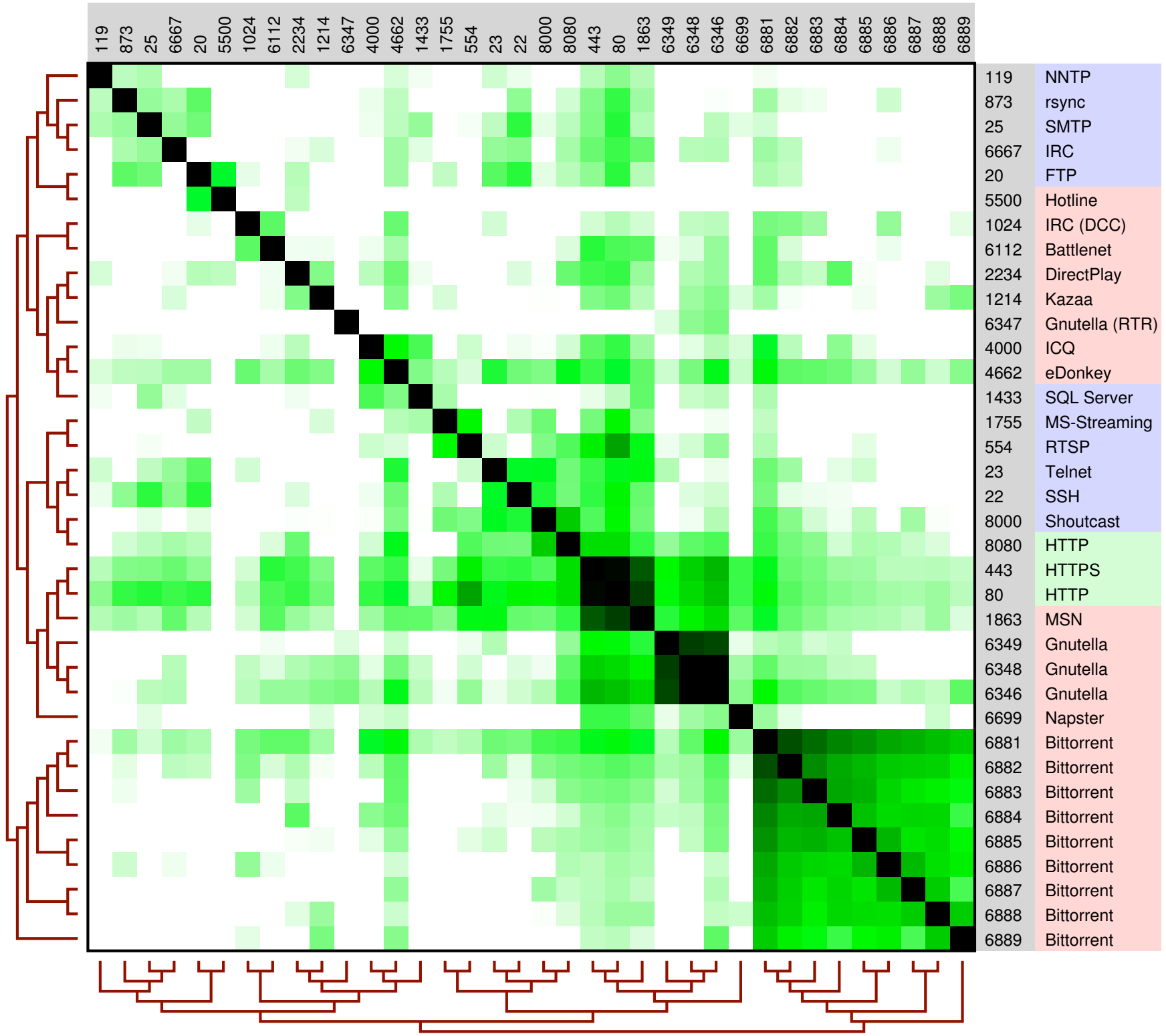

a
IP
PORT
PORT
IP
IP
IP
b
c
servers
clients
behavioral network
applications
hosts
functional network
d
ports
application network

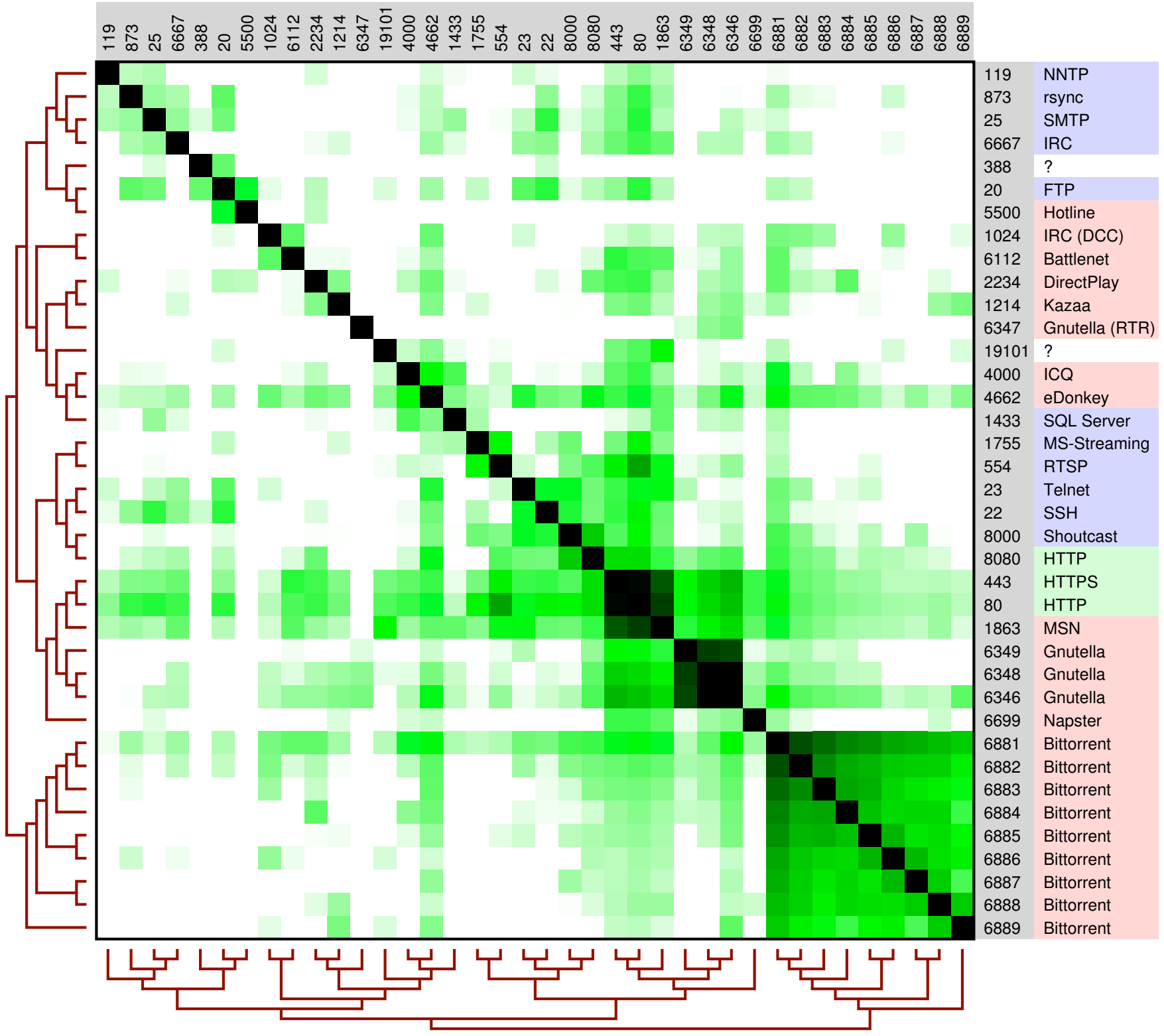
From functional to application clusters



$$d = \frac{1}{\cos\theta} - 1$$




119 873 25 6667 20 5500 1024 6112 2234 1214 6347 4000 4662 1433 1755 554 23 22 8000 8080 443 80 1863 6349 6348 6346 6699 6881 6882 6883 6884 6885 6886 6887 6888 6889
119 NNTP
873 rsync
25 SMTP
6667 IRC
20 FTP
5500 Hotline
1024 IRC (DCC)
6112 Battlenet
2234 DirectPlay
1214 Kazaa
6347 Gnutella (RTR)
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


119 873 25 6667 388 20 5500 1024 6112 2234 1214 6347 19101 4000 4662 1433 1755 554 23 22 8000 8080 443 80 1863 6349 6348 6346 6699 6881 6882 6883 6884 6885 6886 6887 6888 6889
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

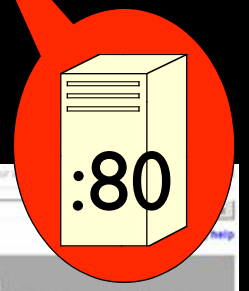
Unknown port	Application
388	Unidata/LDM
19101	Clubbox

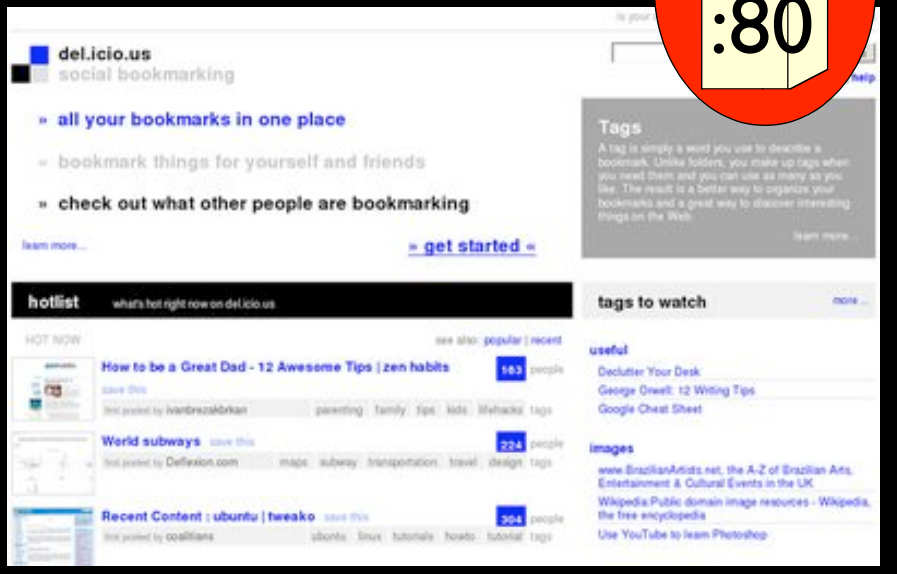
Port	App	Match?
388	weather data transfer	yes
19101	individual file shares	yes
9080	team collaboration	yes
8090	Weblog server	yes
5020	BBFTP file transfer	partial
42899	*unknown*	?
8301	several trojans	partial
1025	many different trojans	yes
20000	BitTorrent	yes
59174	*unknown*	?
20001	several trojans	partial
15002	biology collaboration tool	partial
16881	BitTorrent	yes
9000	several trojans	partial
3124	Web proxy (Windows)	yes
39281	grid-based computing	partial

Summary

- Use application networks to guess the ***function*** of ***unknown or covert*** ports from topological information alone
  - No need to look at payload — preserve ***privacy***
  - No need to inspect all packets — ***efficient*** (possibly real-time)
  - Detect potentially ***malicious*** traffic

Clicks?





Not all dynamics are captured

Future applications

- Anomaly detection
- Application fingerprinting
- Classification of Web clients according to their purpose: *browsers, crawlers, scanners, proxies,* etc.
  - This may provide insight into scalable design
- Identify unknown or covert activities in real time
- Using traffic data to improve the performance of search engines

Thanks

Abilene

INTERNET2

Qwest
Spirit of Service

NORTEL NETWORKS

Juniper NETWORKS

Indiana University School of informatics