The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

SMR.1656 - 4

School and Workshop on
Structure and Function of Complex Networks

16 - 28 May 2005

On the Lack of Typical Behavior in the Global Web Traffic Network

Filippo MENCZER
Indiana University
Eigenmann 909
1900 East Tenth Str.
Bloomington, IN 47406
U.S.A.

On the lack of typical behavior in the global Web traffic network

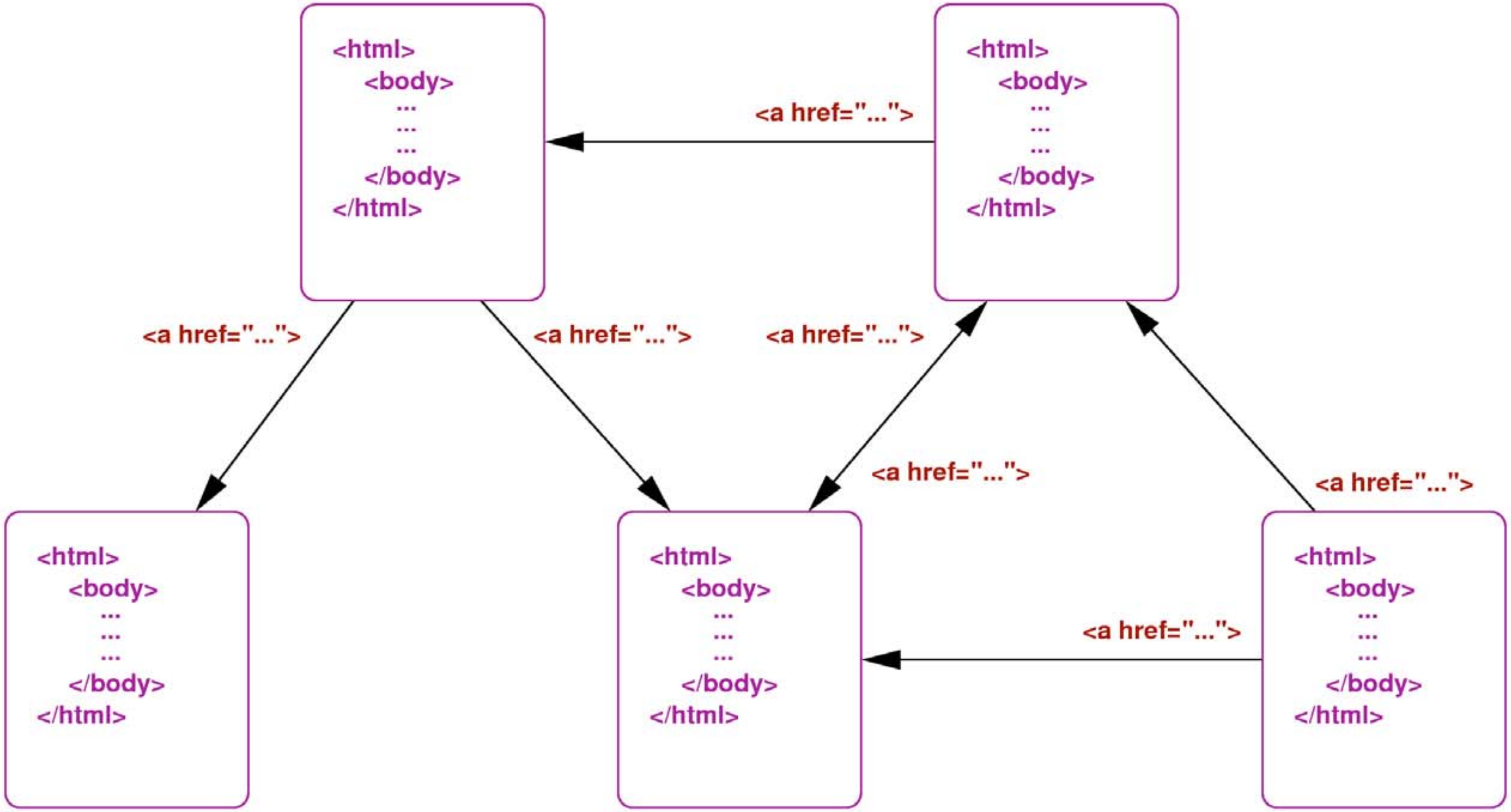
Mark Meiss *Indiana University Department of Computer Science and Advanced Network Management Laboratory*

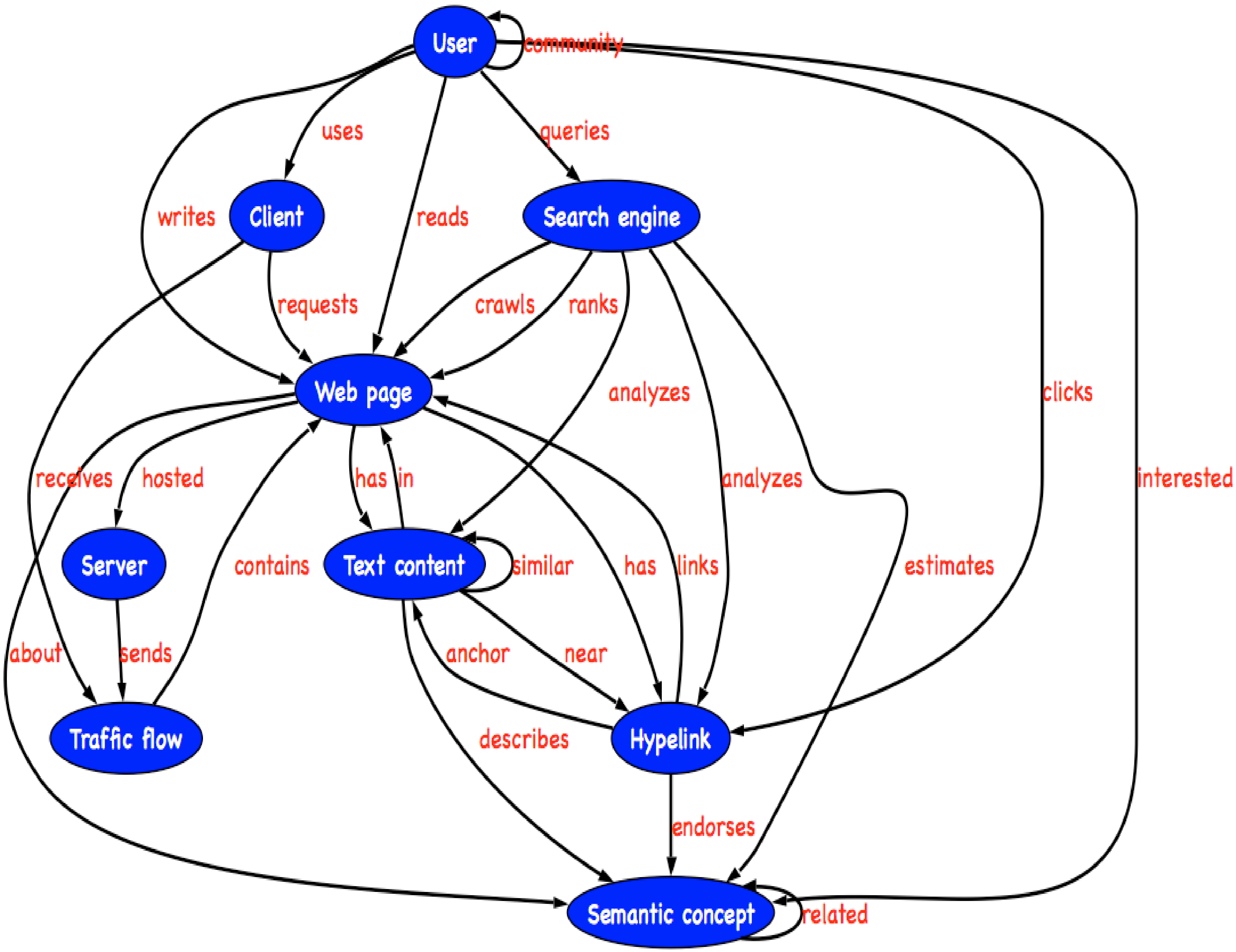
Filippo Menczer *Indiana University School of Informatics and Department of Computer Science*

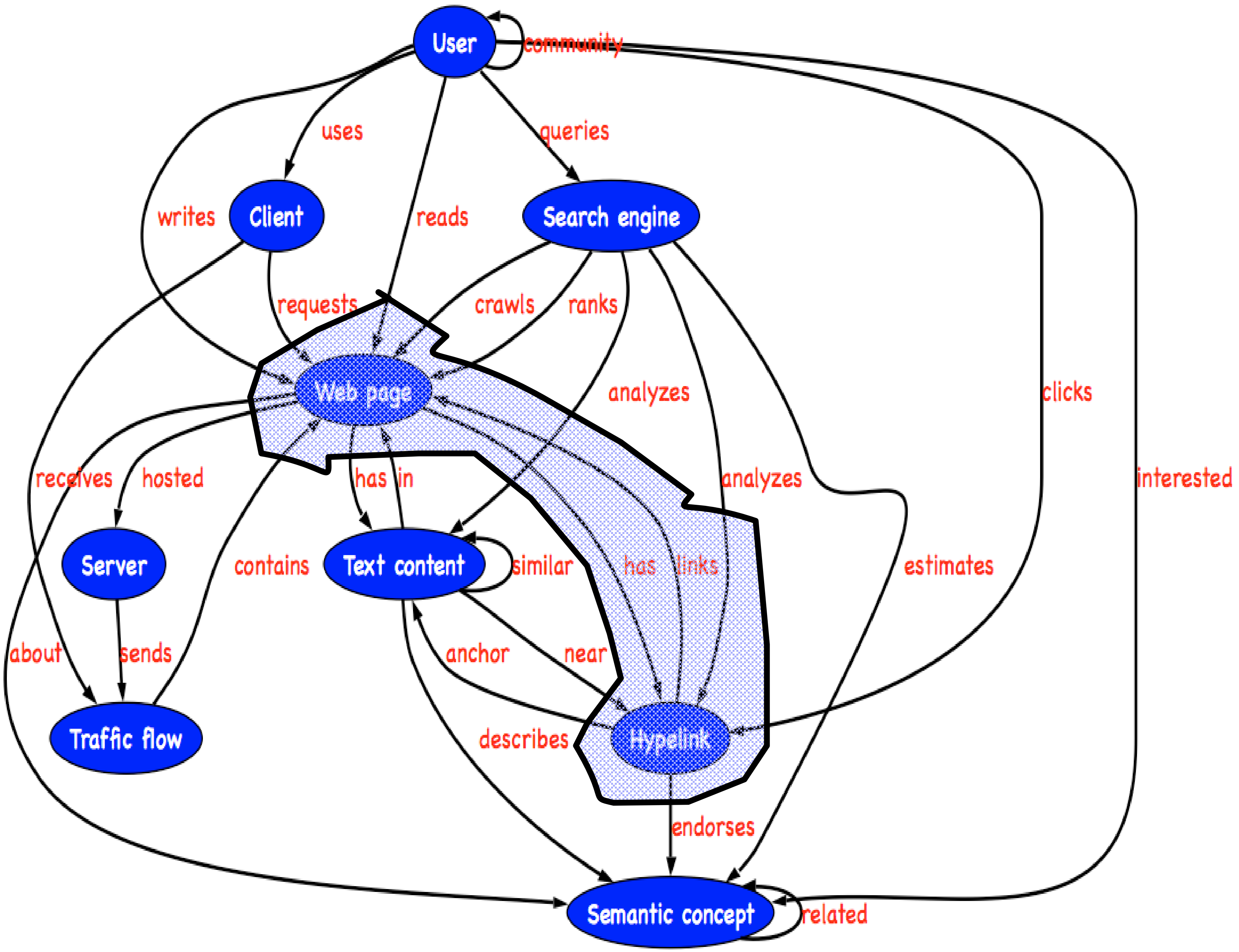
Alessandro Vespignani *Indiana University School of Informatics*

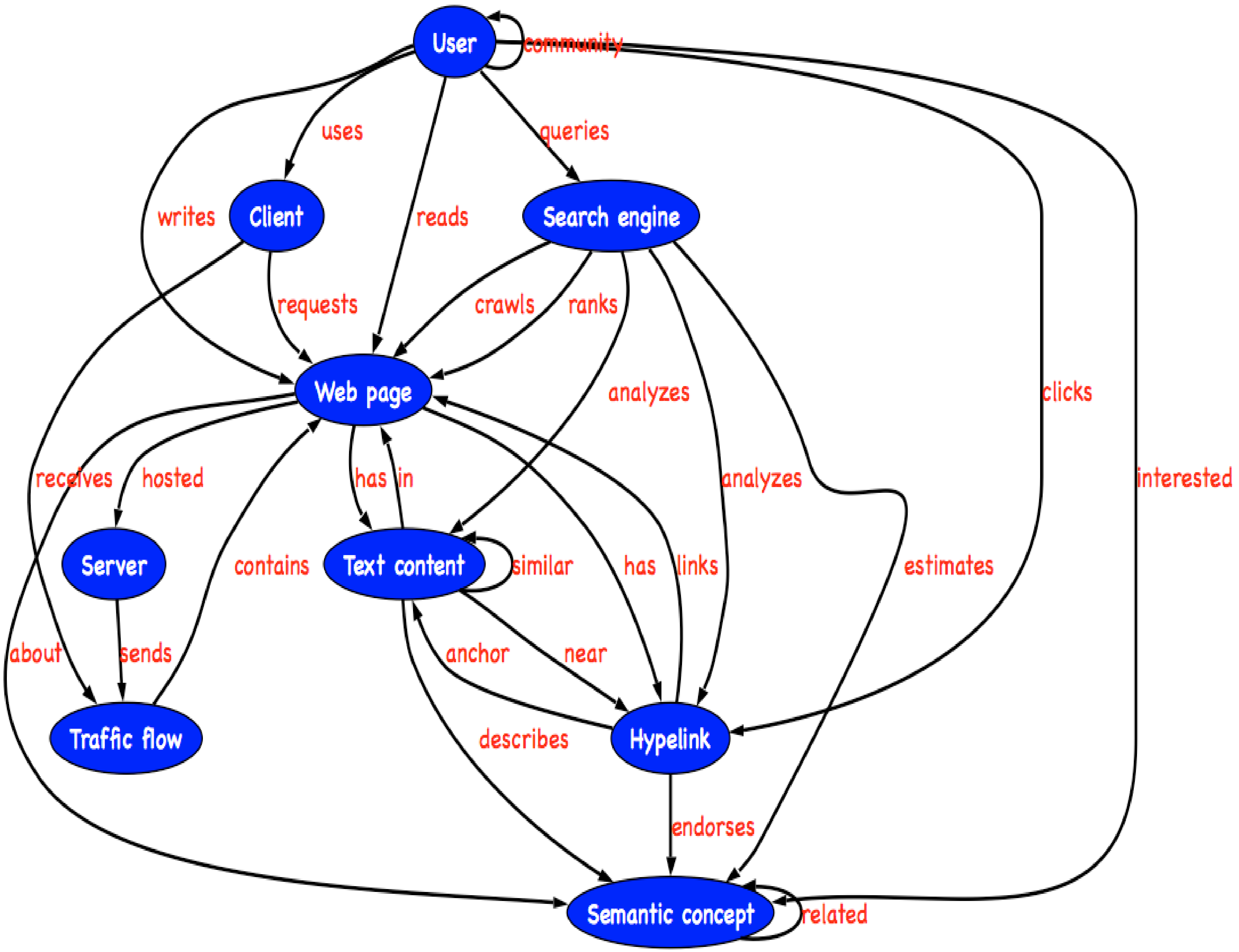
Various complex networks coevolve in the Web

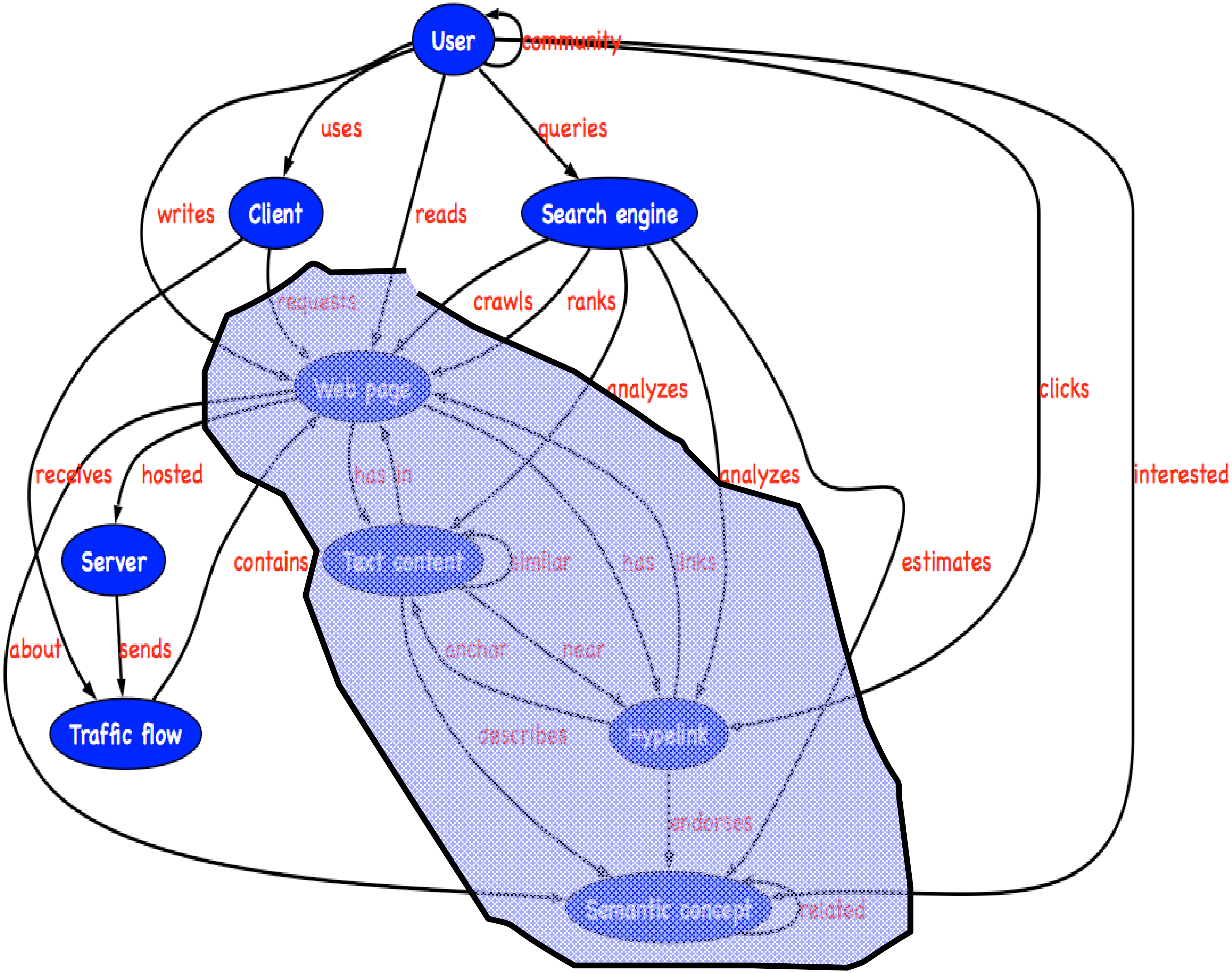
- The *Link Graph*

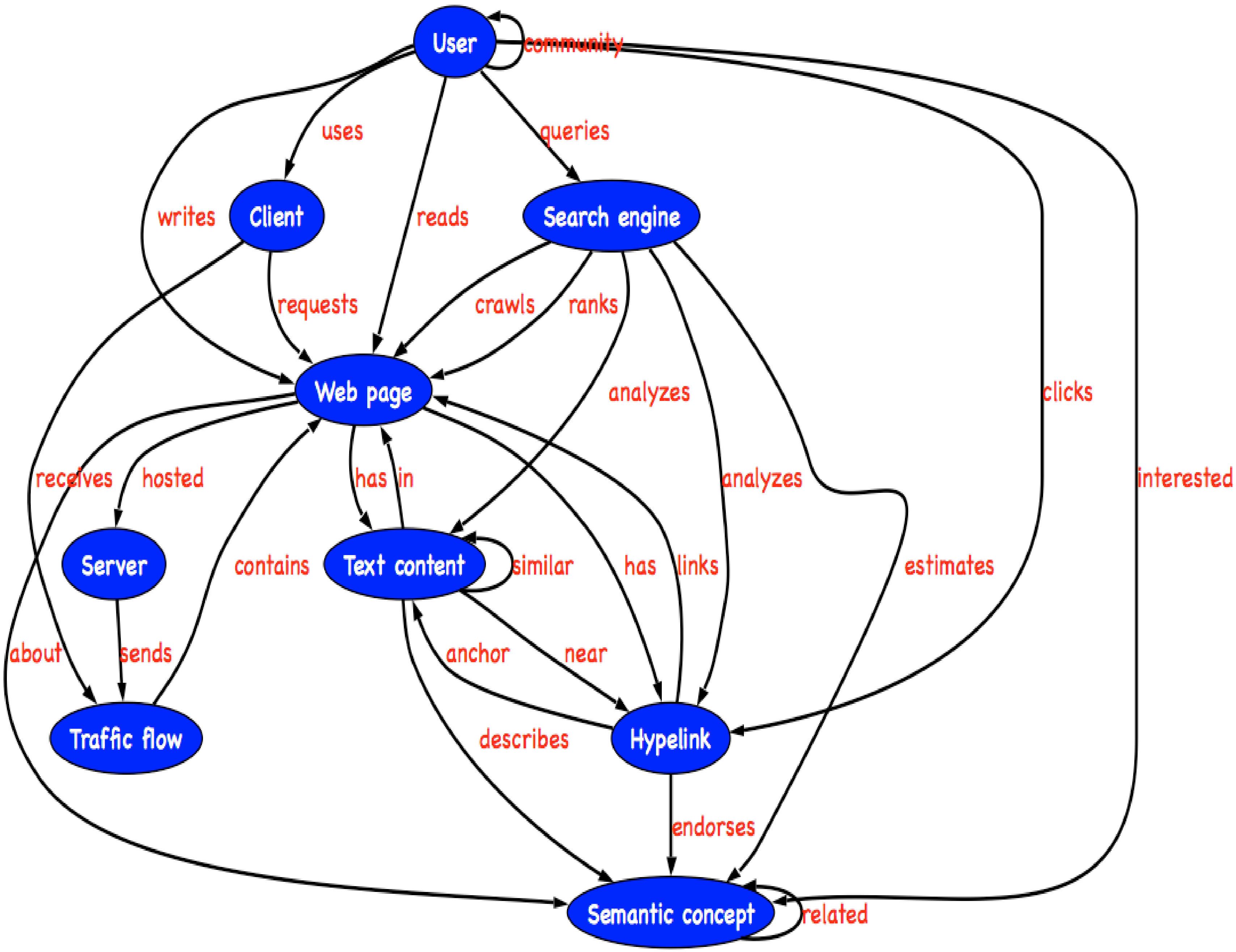


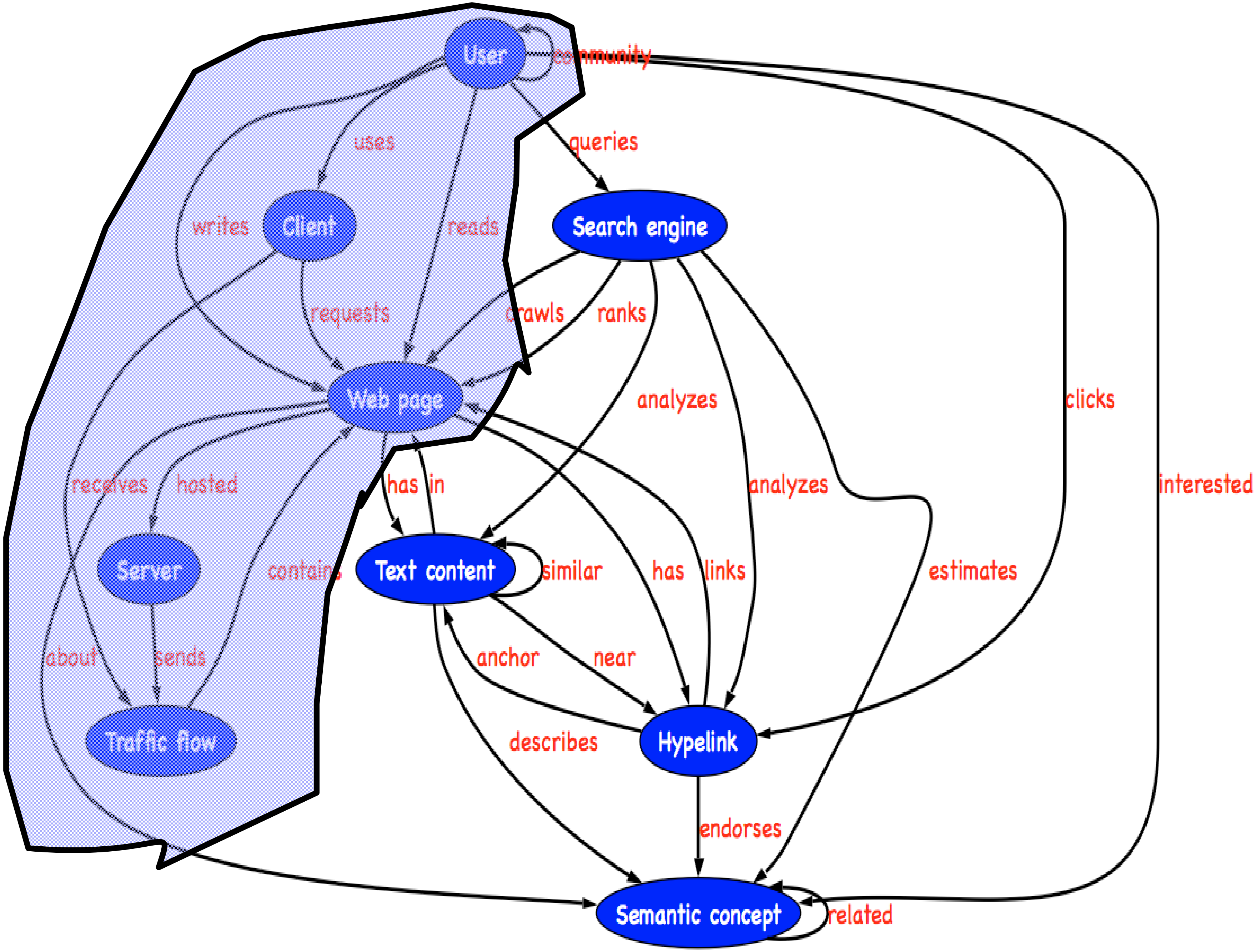

User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
clicks
receives
hosted
has
in
analyzes
interested
Server
contains
Text content
similar
has
links
estimates
about
sends
anchor
near
Traffic flow
describes
Hypelink
endorses
Semantic concept
related


User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
clicks
receives
hosted
has
in
analyzes
interested
Server
contains
Text content
similar
has
links
estimates
about
sends
anchor
near
Traffic flow
describes
Hypelink
endorses
Semantic concept
related


User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
clicks
receives
hosted
has
in
analyzes
interested
Server
contains
Text content
similar
has
links
estimates
about
sends
anchor
near
Traffic flow
describes
Hypelink
endorses
Semantic concept
related

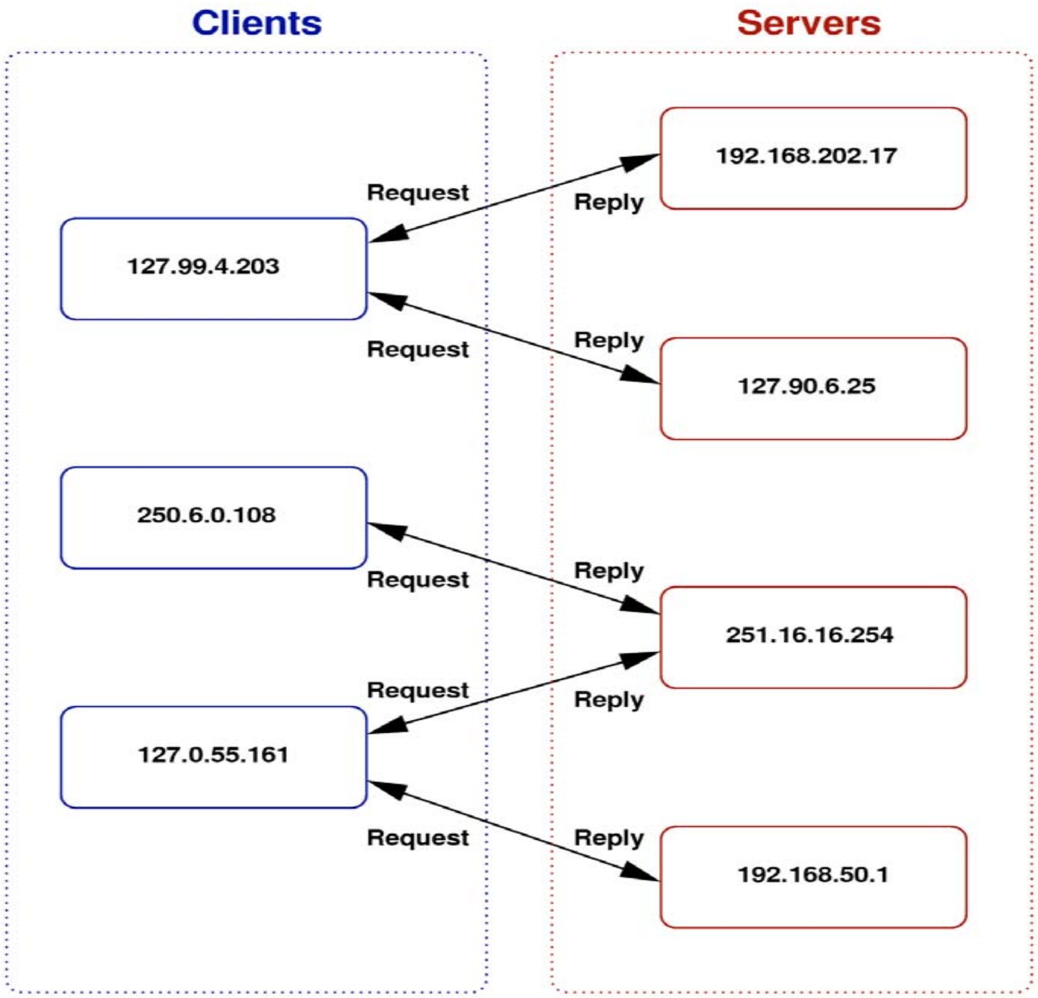

User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
receives
hosted
has
in
analyzes
Server
contains
Text content
similar
has
links
estimates
clicks
interested
about
sends
anchor
near
Traffic flow
describes
Hyperlink
endorses
Semantic concept
related


User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
clicks
receives
hosted
has
in
analyzes
interested
Server
contains
Text content
similar
has
links
estimates
about
sends
anchor
near
Traffic flow
describes
Hypelink
endorses
Semantic concept
related


User
community
uses
queries
writes
Client
reads
Search engine
requests
crawls
ranks
Web page
analyzes
clicks
receives
hosted
has
in
analyzes
interested
Server
contains
Text content
similar
has
links
estimates
about
sends
anchor
near
Traffic flow
describes
Hypelink
endorses
Semantic concept
related

Another way of studying the Web

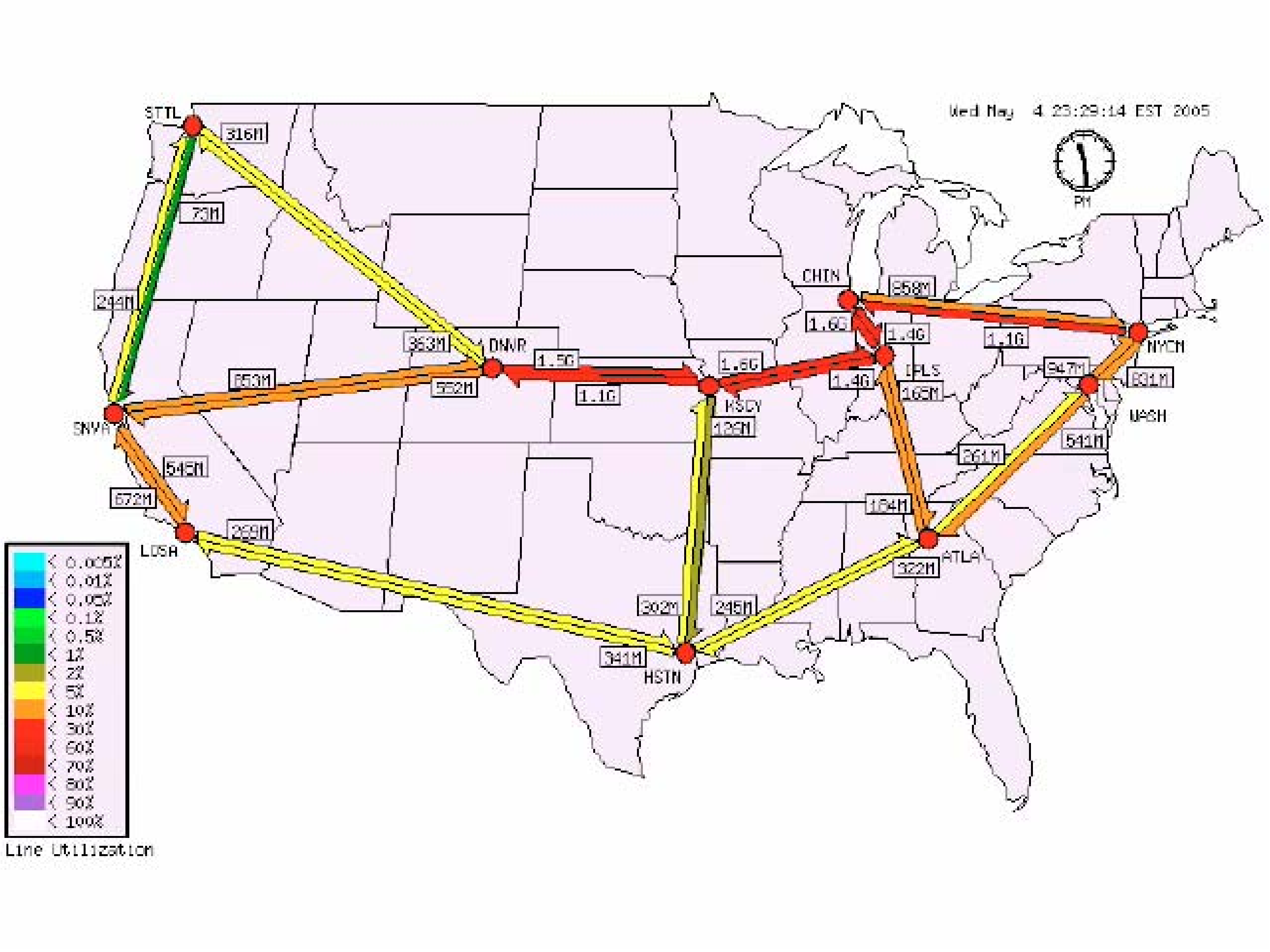
- The *Behavioral Network(s)*



Overview

1. Collection of ***network flow data*** from Internet2 core routers
2. ***Weighted bipartite digraph*** representation of Web traffic
3. Analysis of ***Web client*** behavior
4. Analysis of ***Web server*** behavior
5. ***Summary*** and future work

The Internet2/Abilene network

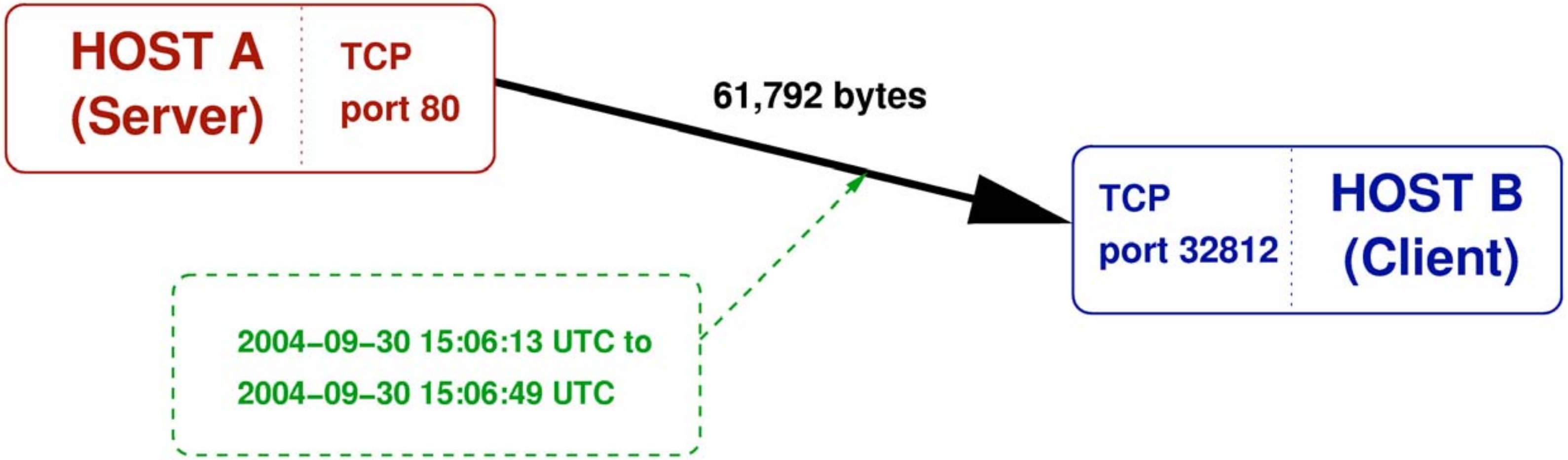


- TCP/IP network connecting ***research and educational*** institutions in the U.S.
  - Over 200 universities and corporate research labs
- Also provides ***transit service*** between Pacific Rim and European networks

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

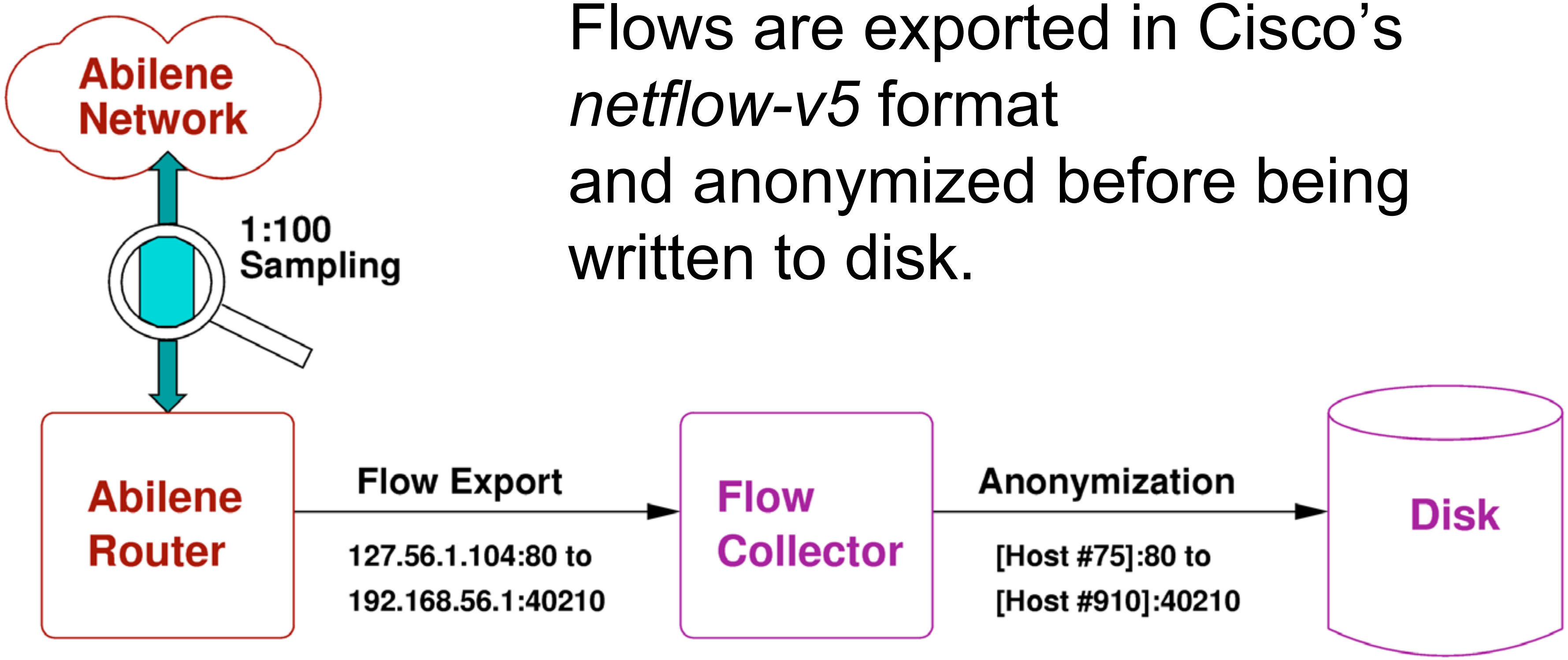
Network flow data



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

Flow collection

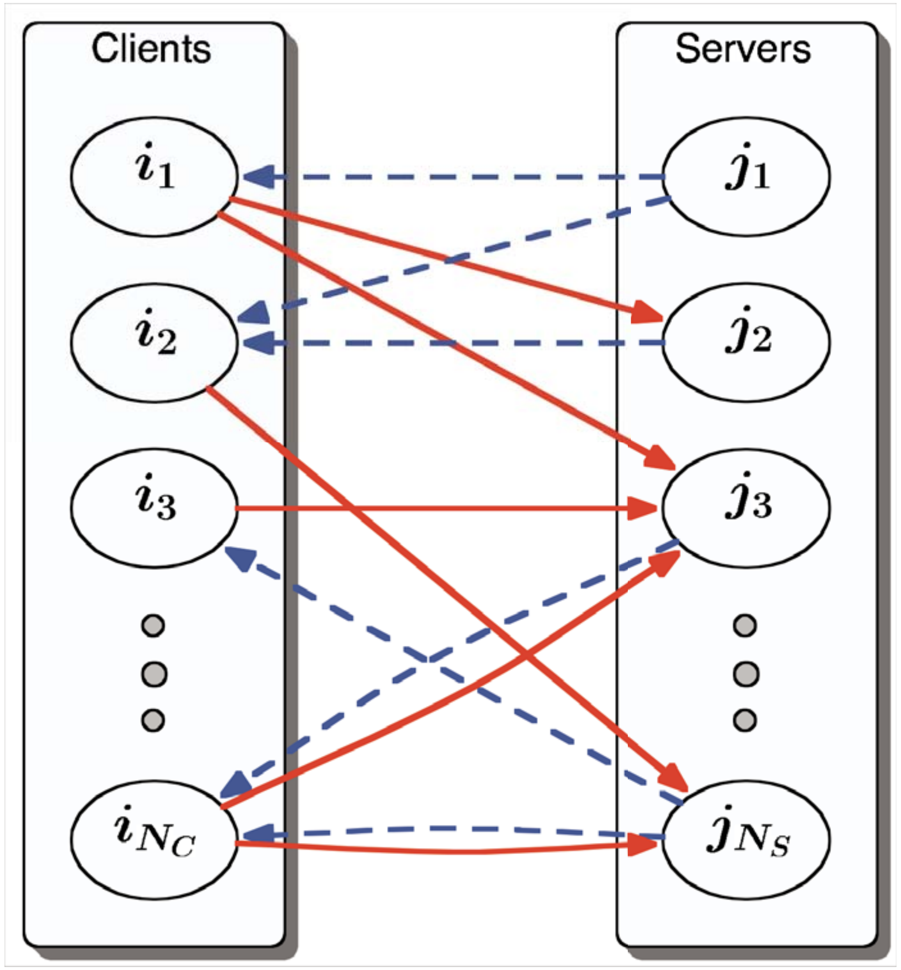
Flows are exported in Cisco's *netflow-v5* format
and anonymized before being written to disk.

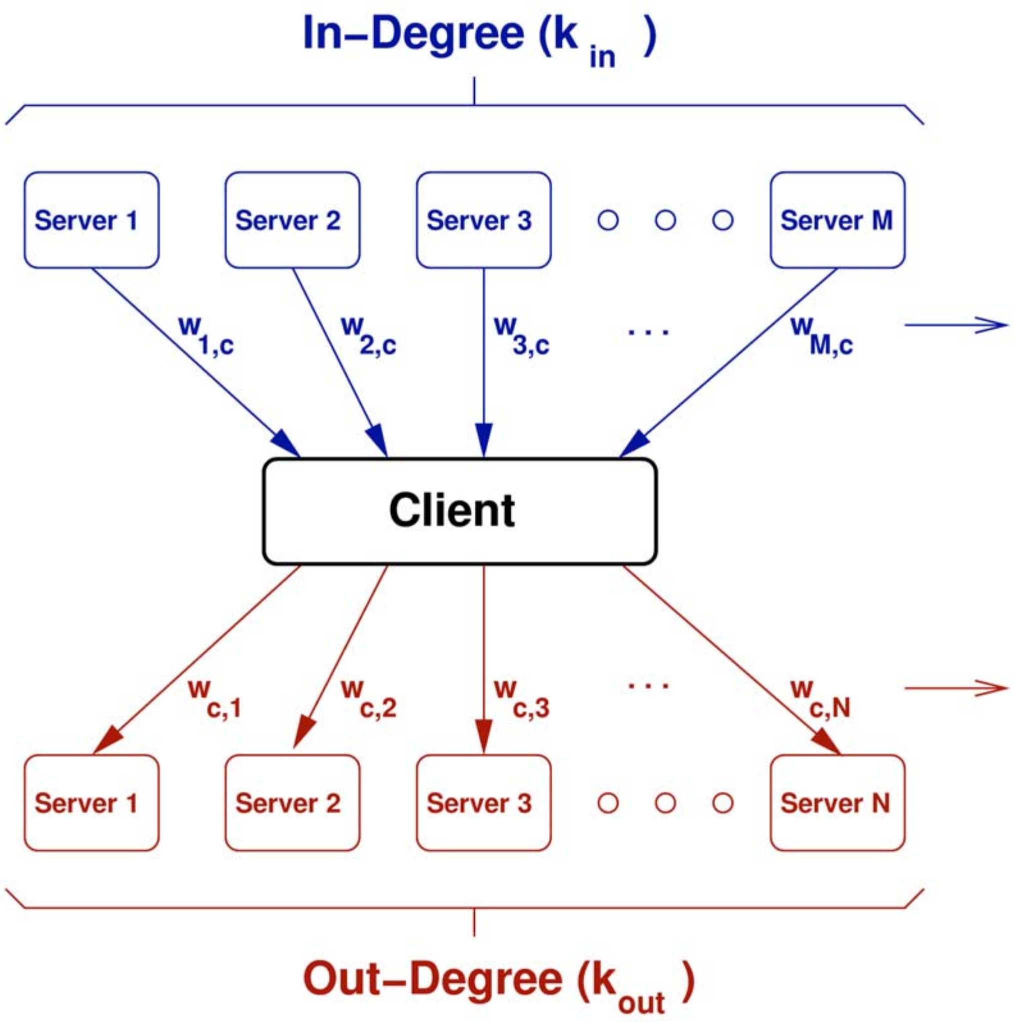


Data set for analysis

- Full 24-hour day of network flow data starting at 2004-09-30 05:00:00 UTC
  - 742,000,000 flows
  - 30,000,000 unique hosts
  - 319,000,000 flows involving port 80

Weighted bipartite digraph





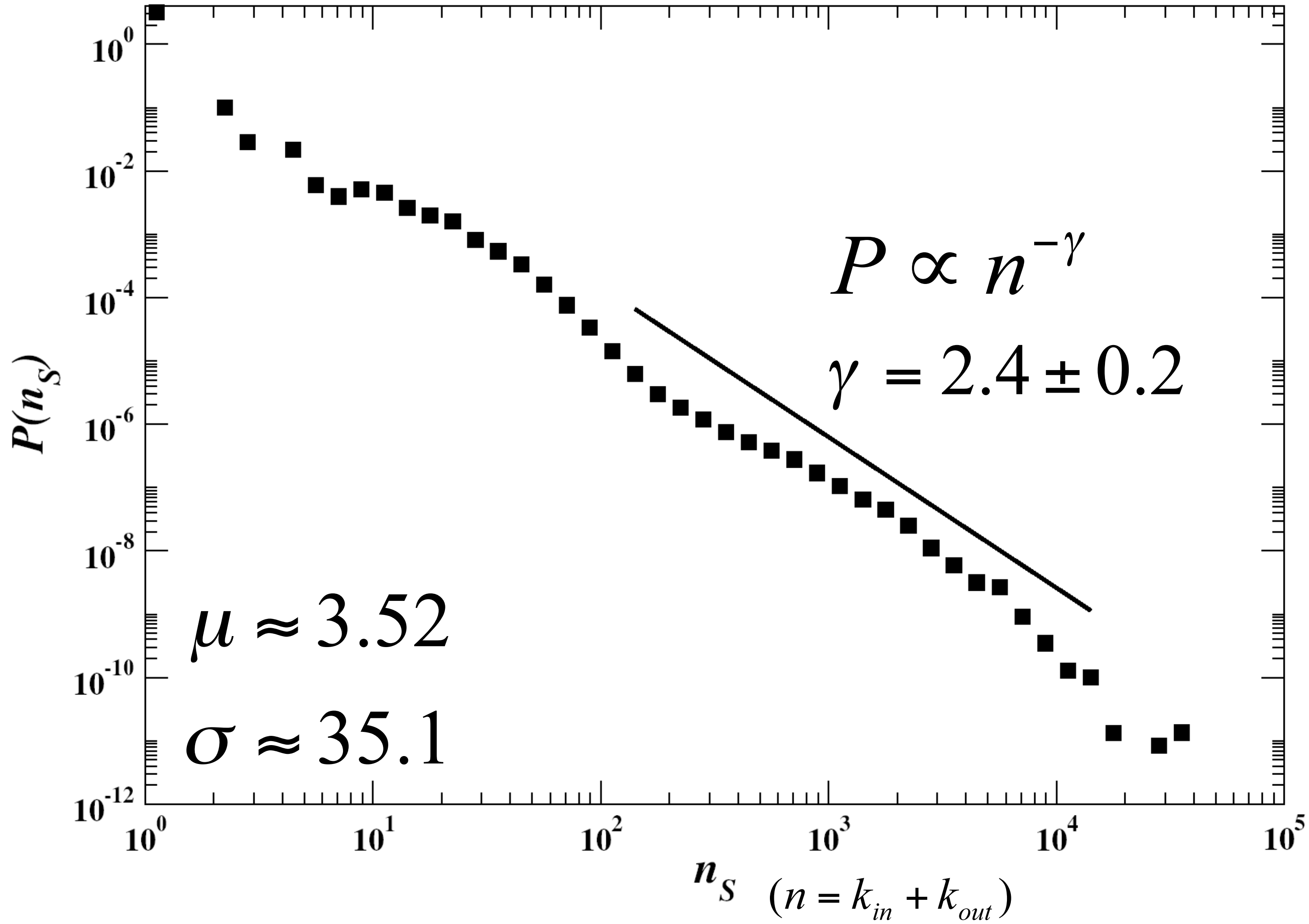
In–Strength (s_{in})

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

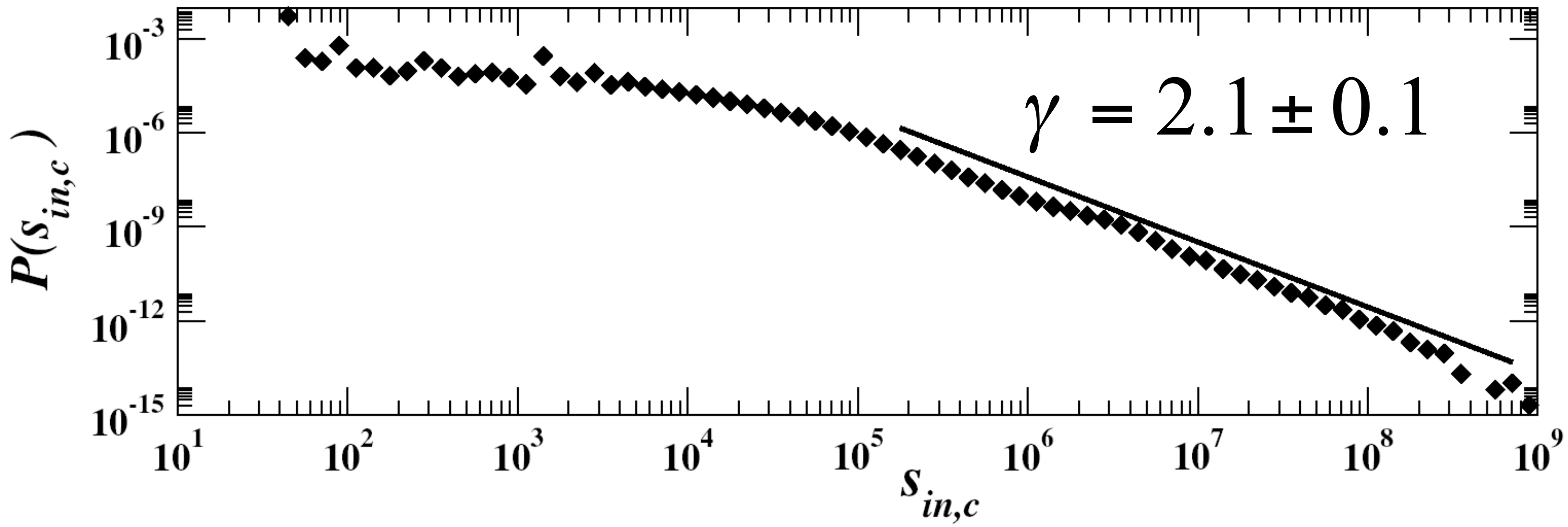
Out–Strength (s_{out})

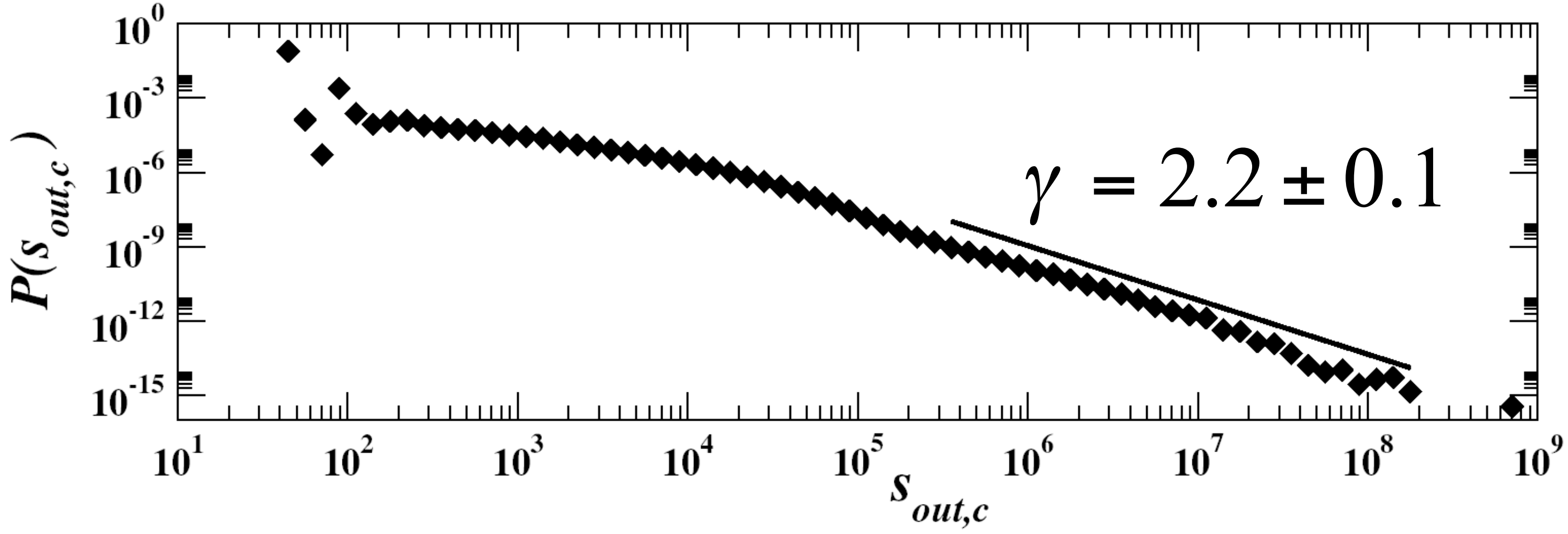
$$s_{out} = \sum_{j=1}^{N} w_{C,j}$$

Clients: Degree distribution

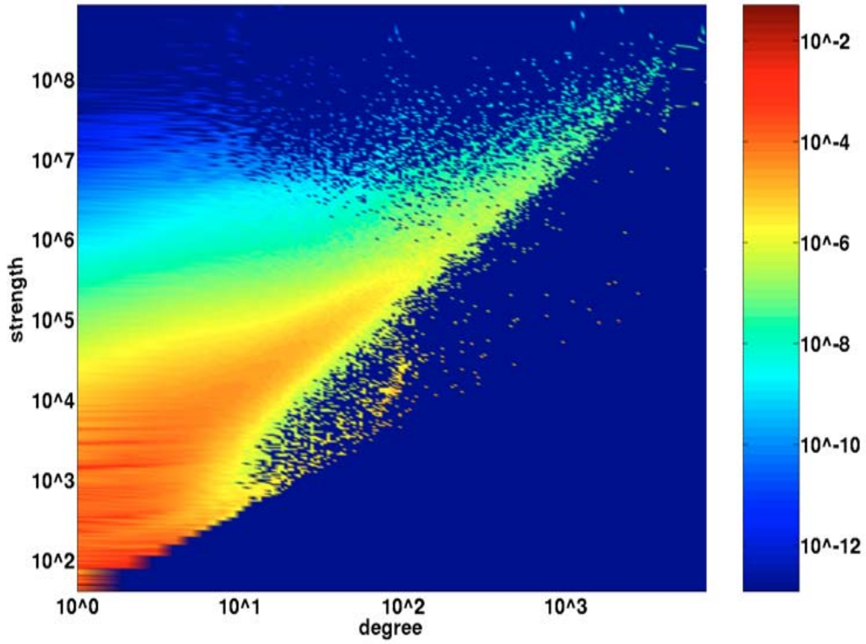


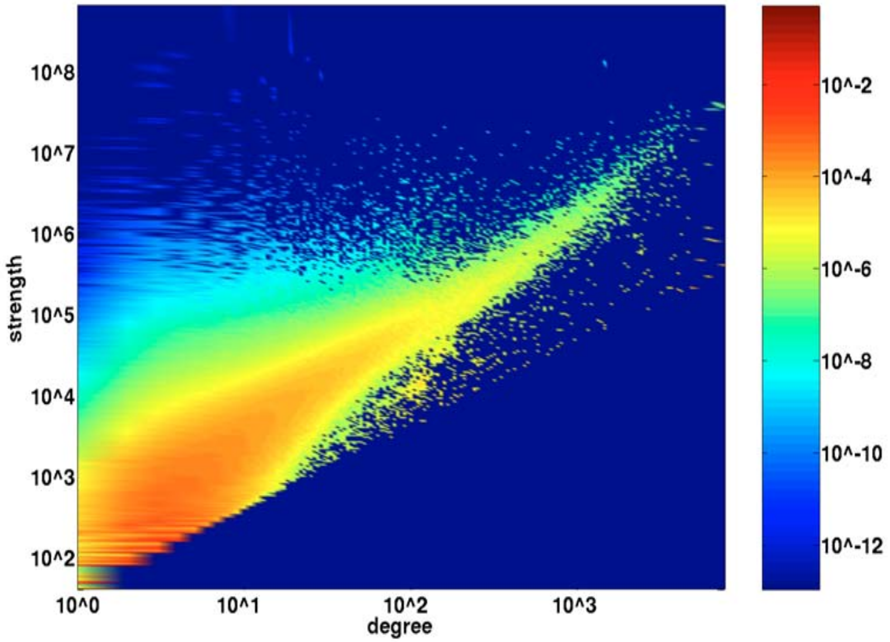
Clients: Strength distributions



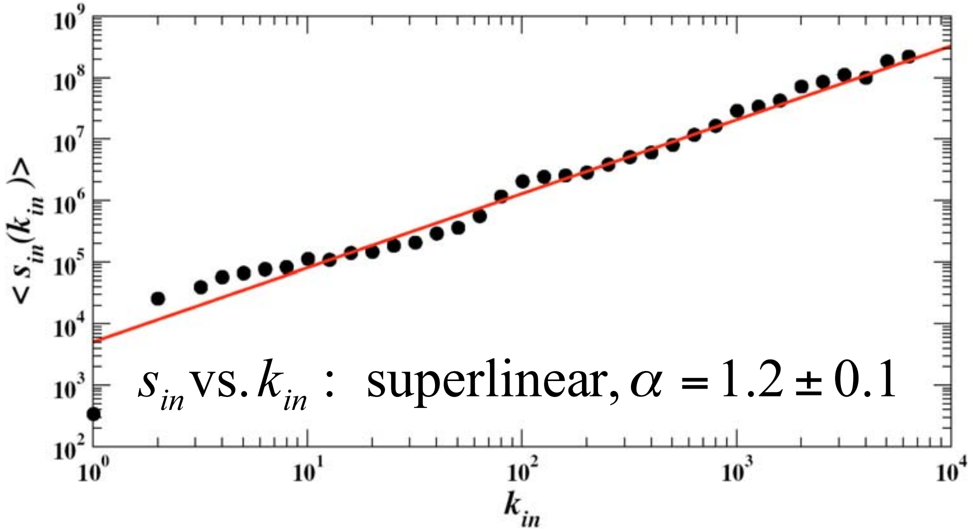


Clients: Strength vs. Degree

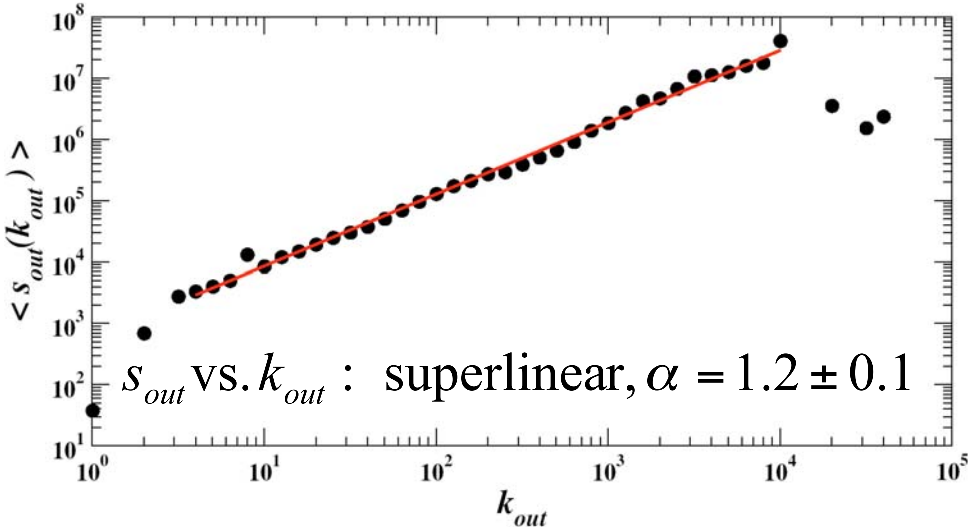




Clients: Strength vs. Degree



s_{in} vs. k_{in} : superlinear, $\alpha = 1.2 \pm 0.1$

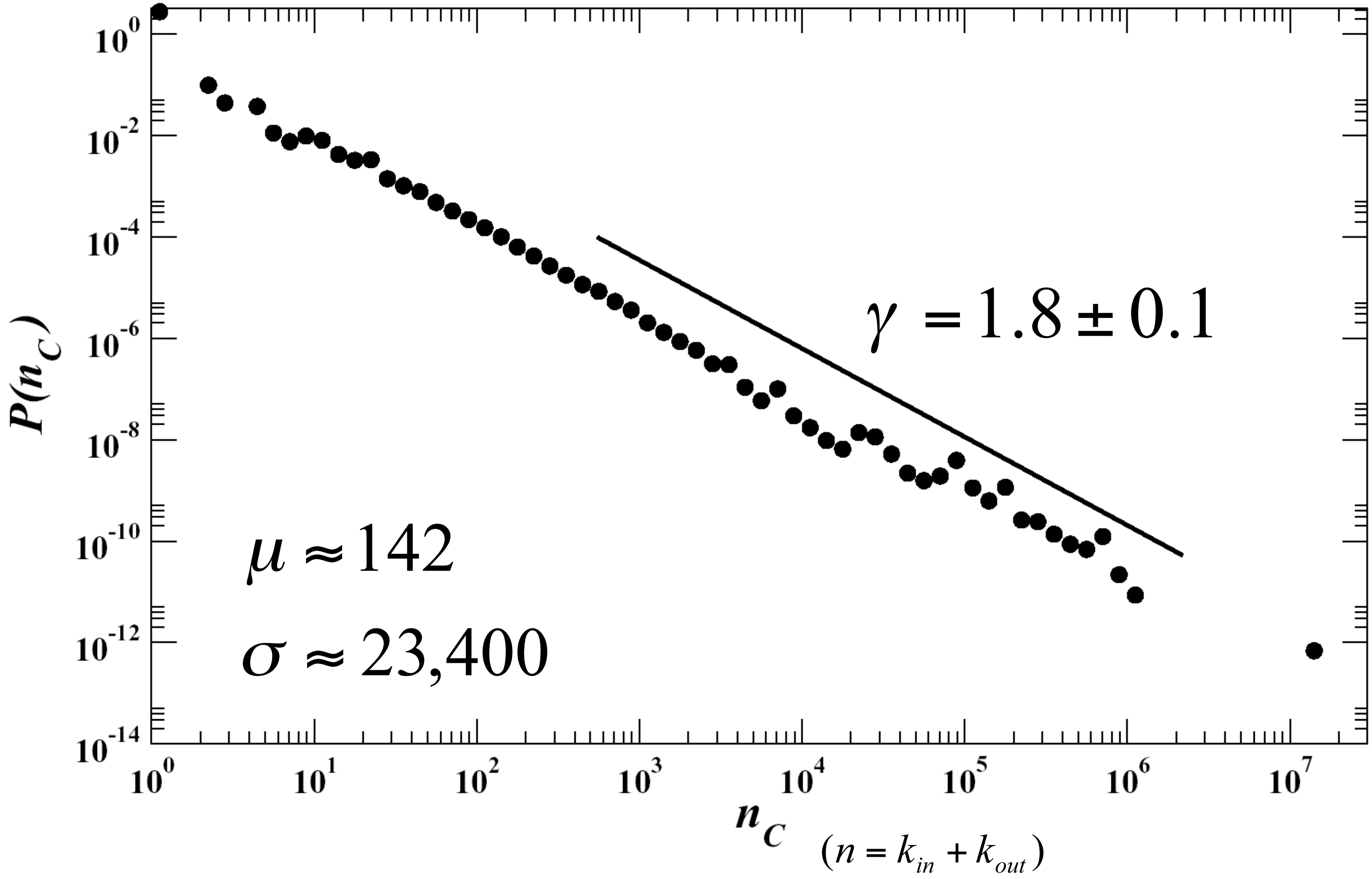


s_{out} vs. k_{out} : superlinear, $\alpha = 1.2 \pm 0.1$

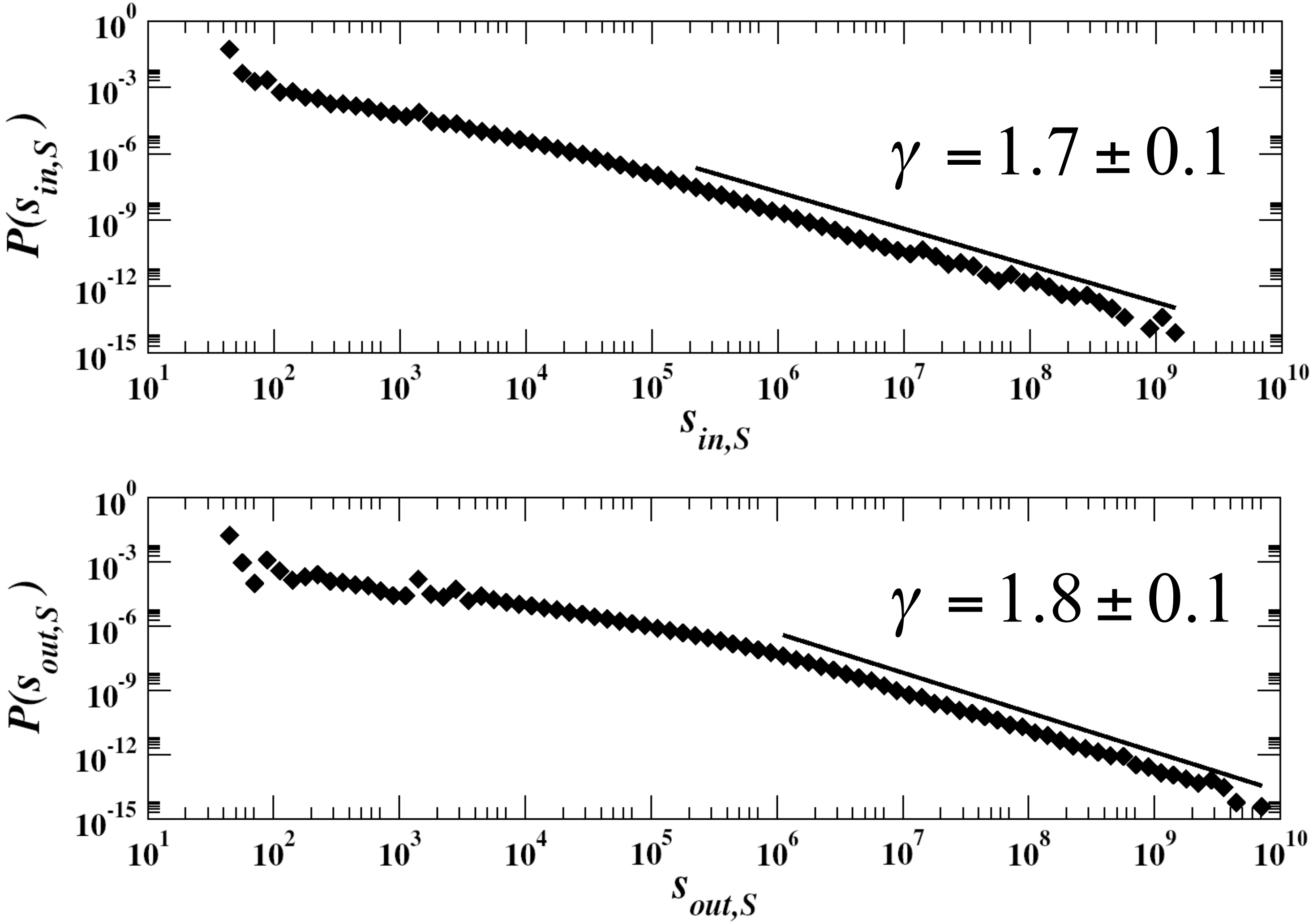
Super-linear behavior in clients

- As the ***number of servers*** in contact with a Web client ***increases***, so does the ***amount of traffic*** exchanged with ***each server***
- This points to difficulties in designing scalable client applications
- We are developing techniques to differentiate different types of client (browsers, crawlers, scanners, etc.)

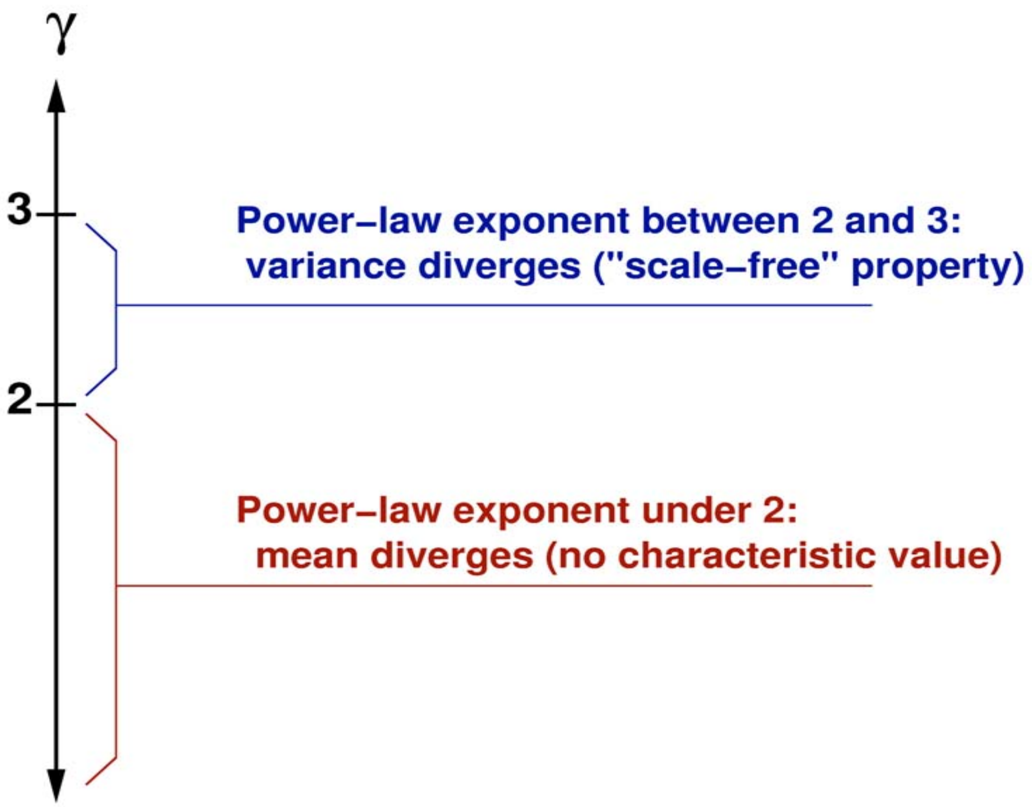
Servers: Degree distribution



Servers: Strength distributions

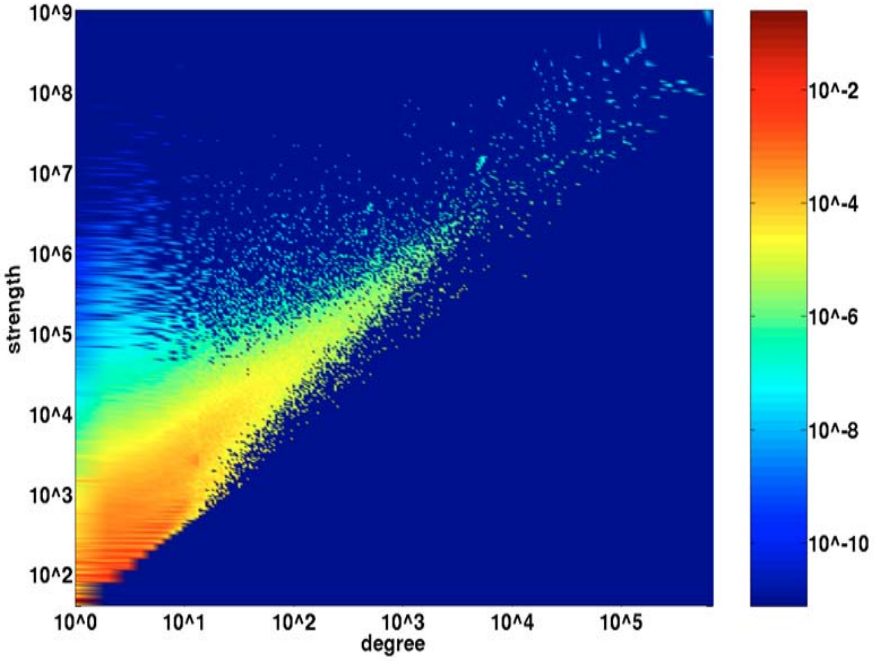


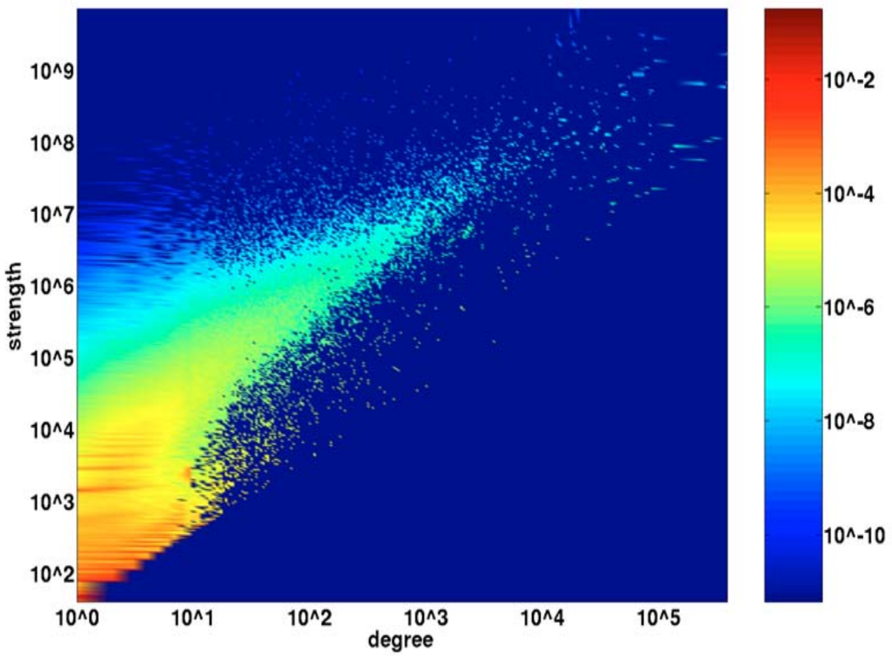
Unbounded fluctuations in server strength


γ
3
2
Power–law exponent between 2 and 3:
variance diverges ("scale–free" property)
Power–law exponent under 2:
mean diverges (no characteristic value)

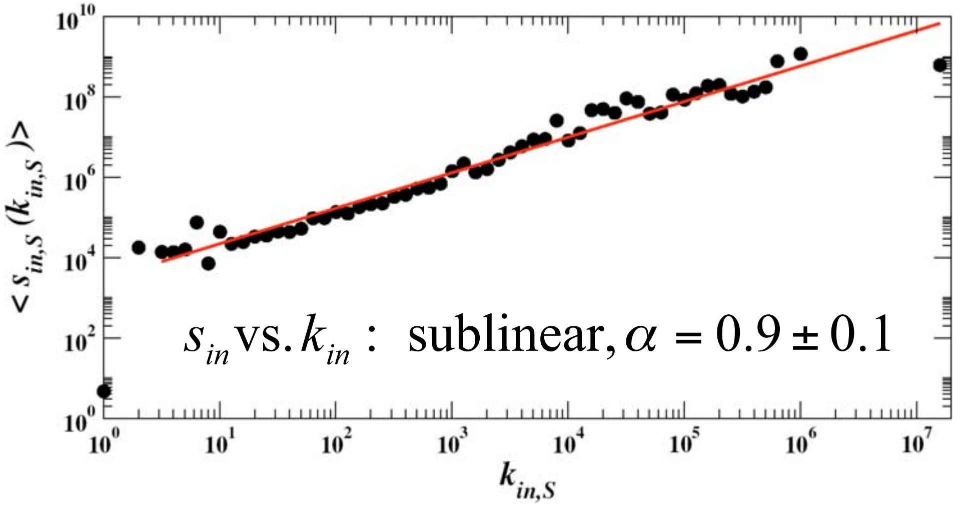
- In our sample, γ is definitely under 2
  - Mean is dependent on size of sample
  - No clear scale for a general-purpose Web server

Servers: Strength vs. Degree

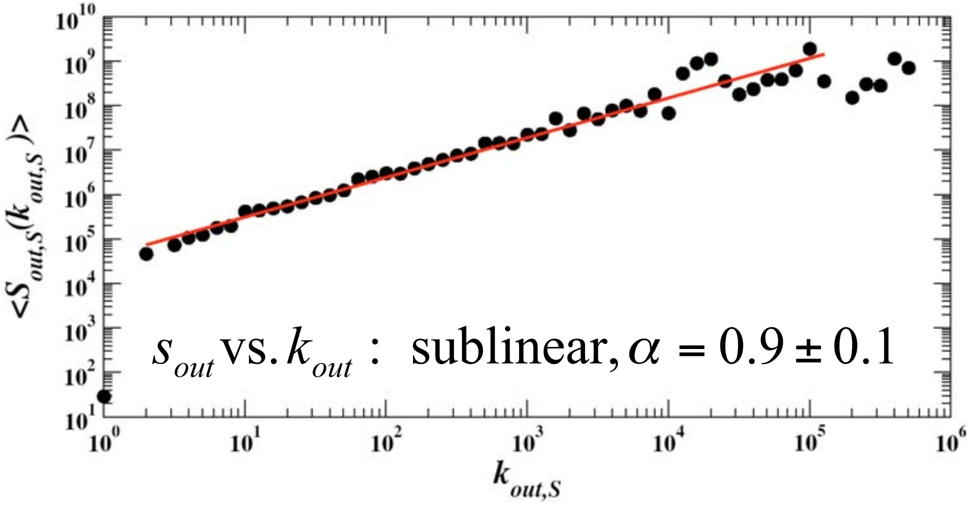




Servers: Strength vs. Degree



s_{in} vs. k_{in} : sublinear, $\alpha = 0.9 \pm 0.1$



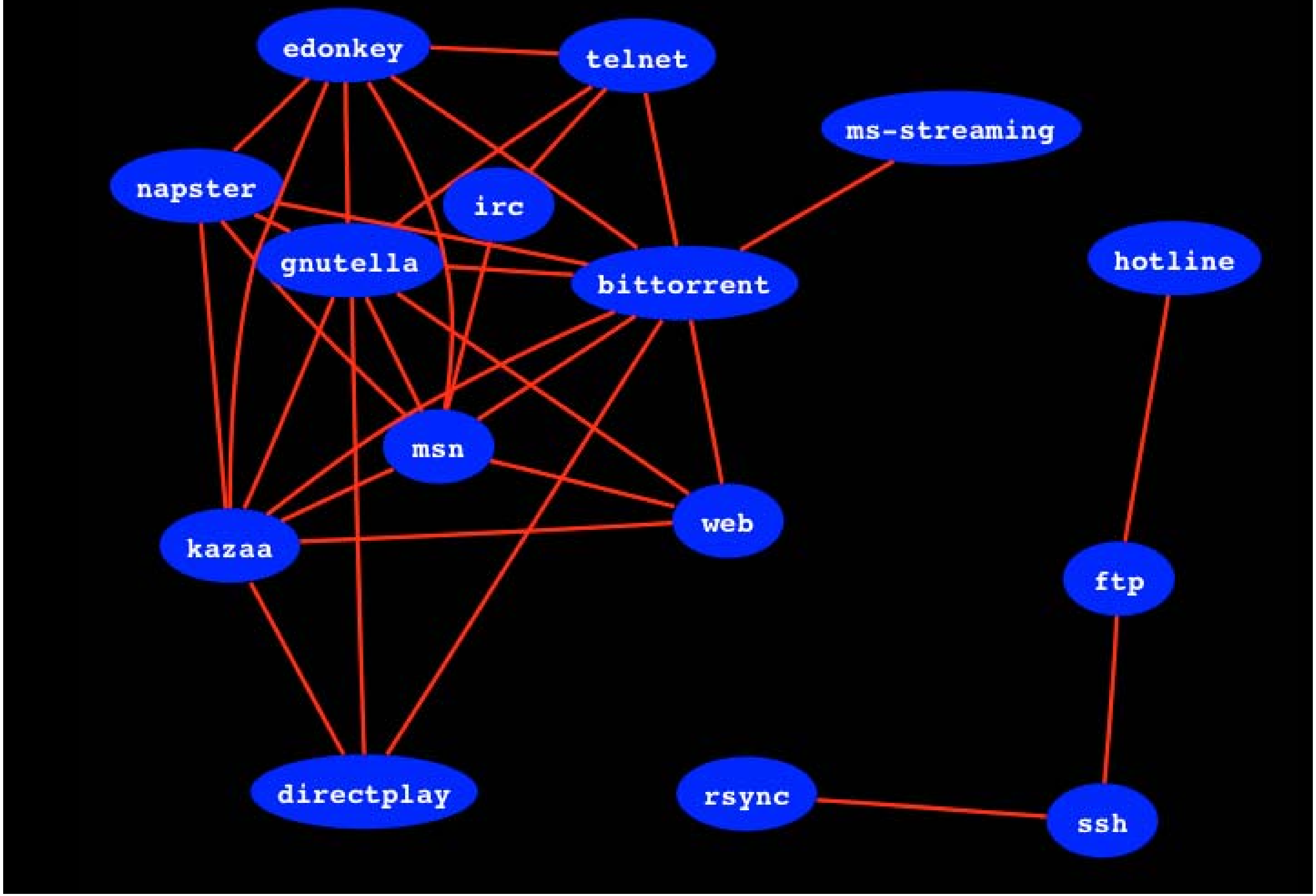
s_{out} vs. k_{out} : sublinear, $\alpha = 0.9 \pm 0.1$

Summary

- ***Power-law distributions*** are found in all aspects of the Web behavioral network
  - □ Degree, strength, and weight distributions for both clients and servers
- The relationship between ***degree*** and ***strength*** for Web clients is ***super-linear***
- The ***strength*** distribution for Web ***servers*** lacks any ***mean value***
- Models must be able to account for these heavy-tailed distributions and the non-linear coupling between degree and strength

Current Work

- Confirmation of analysis with ***more recent data***
  - Data gathered between 2005-04-08 and 2005-04-15 show the same characteristics
- Extension of analysis to ***other applications*** (especially peer-to-peer)
- Analysis of ***correlation of use*** of major network applications.


edonkey
telnet
ms-streaming
napster
irc
gnutella
bittorrent
hotline
msn
web
kazaa
ftp
directplay
rsync
ssh

Future work

- Ongoing repetition of analysis on future data sets to identify long-term trends
- Classification of Web clients according to their purpose: *browsers, crawlers, scanners,* etc.
    - This may provide insight into scalable design
- Using flow data to improve the performance of search engines

Questions and comments