A day of Internet life

Mark Meiss^{1,3}, Filippo Menczer^{1,2}, Alessandro Vespignani²



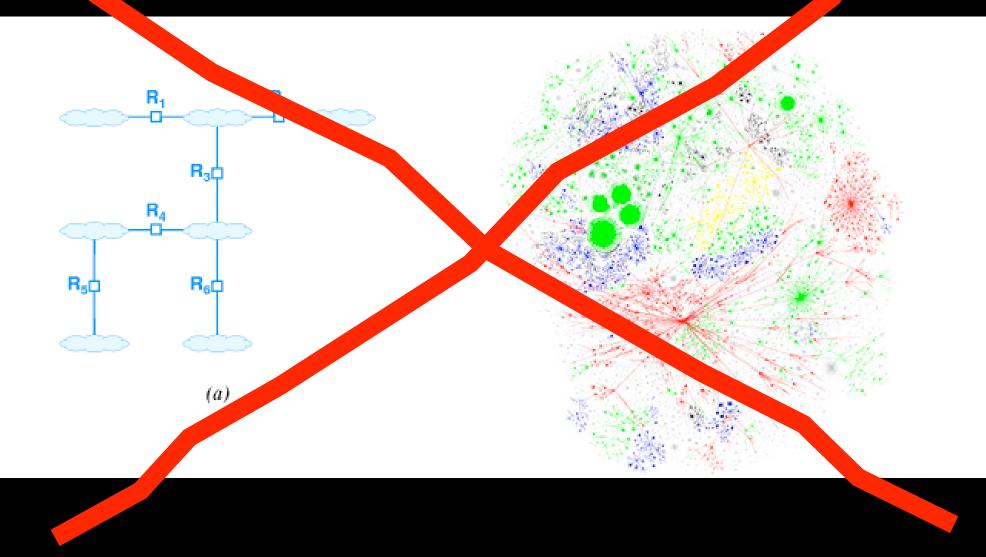
 (1) Department of Computer Science
(2) Department of Informatics
(3) Advanced Network Management Lab Indiana University, Bloomington

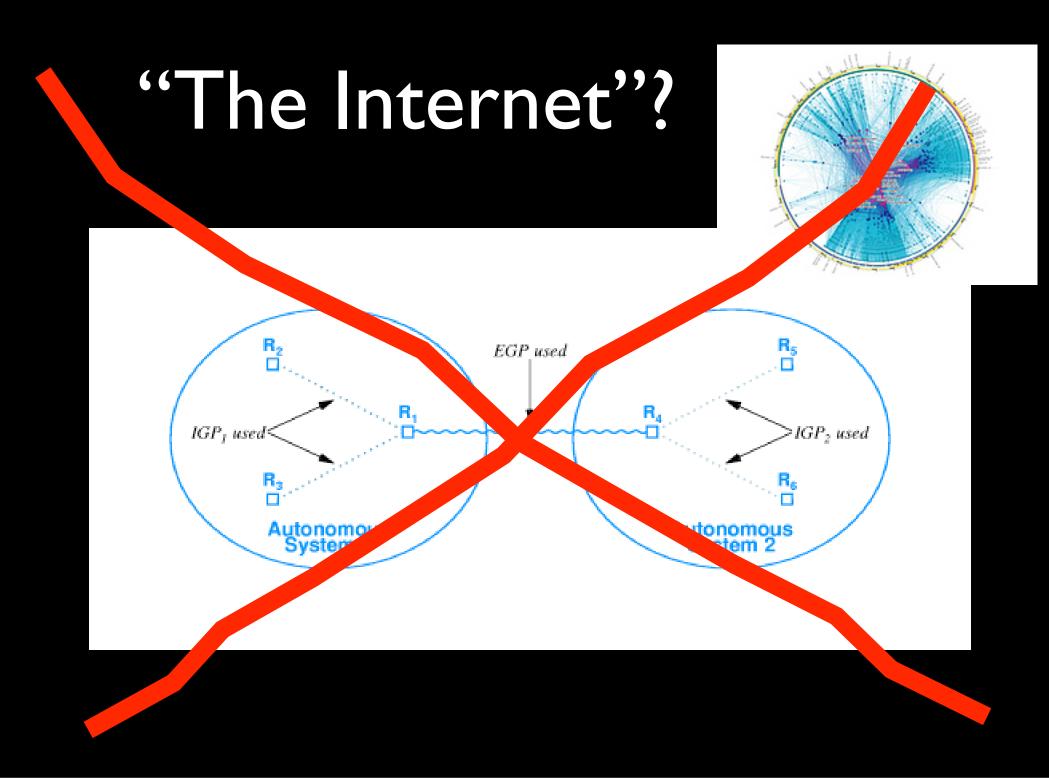




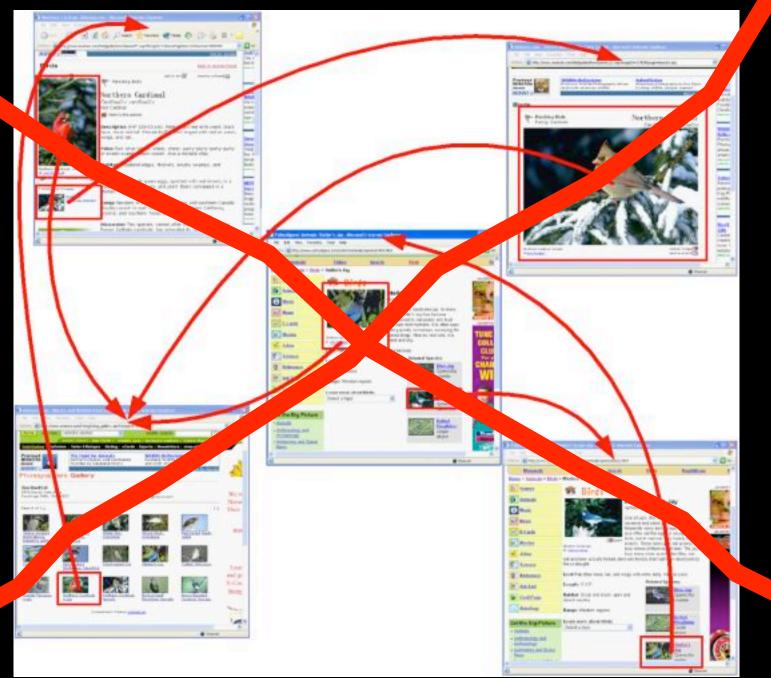


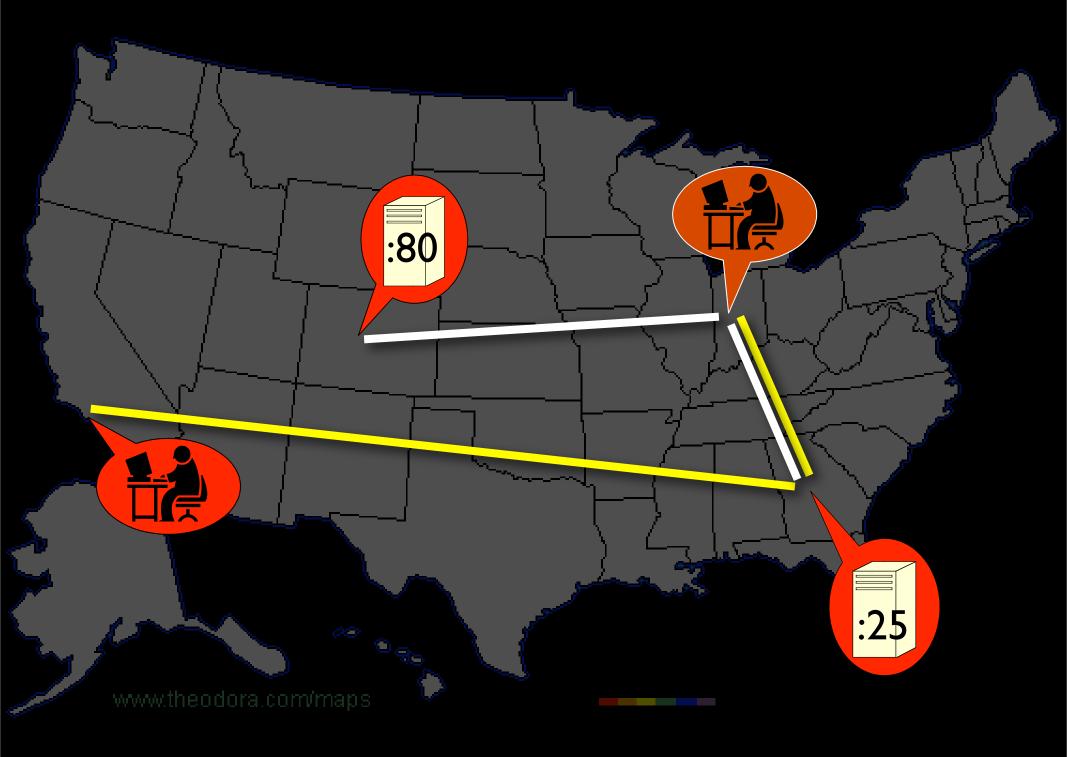
"The Internet"?





"The Internet"?

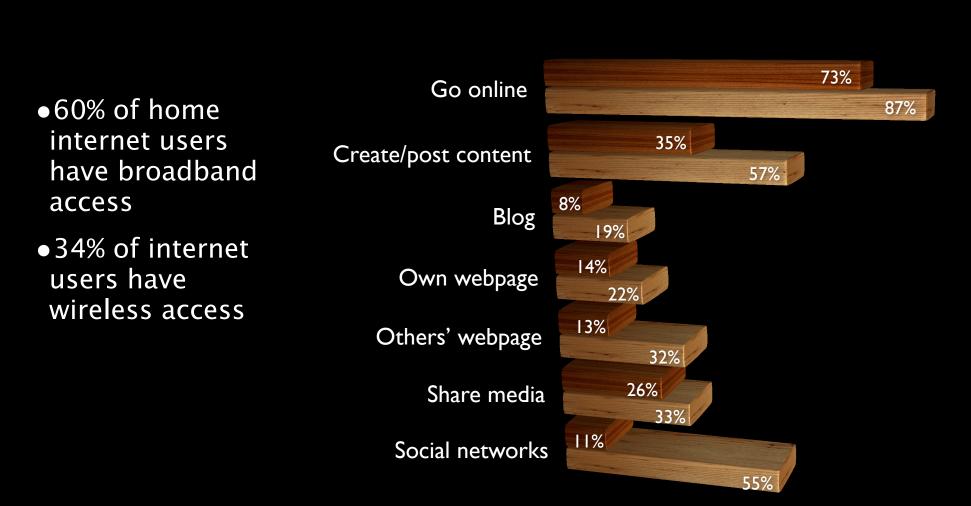




What do we do online?

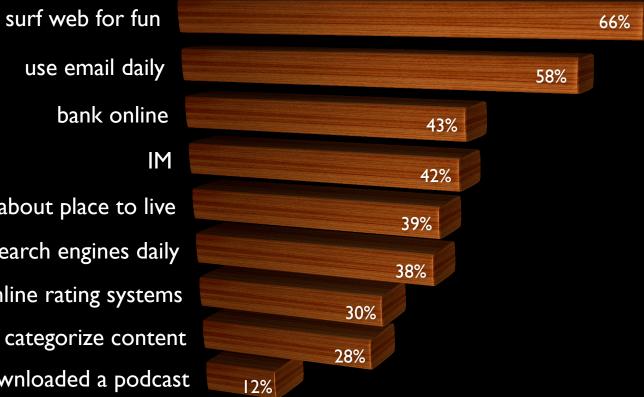
Adults

Teens



Source: PIP (pewinternet.org)

More stats...



bank online IM look for info about place to live use search engines daily use online rating systems tag or categorize content

downloaded a podcast

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

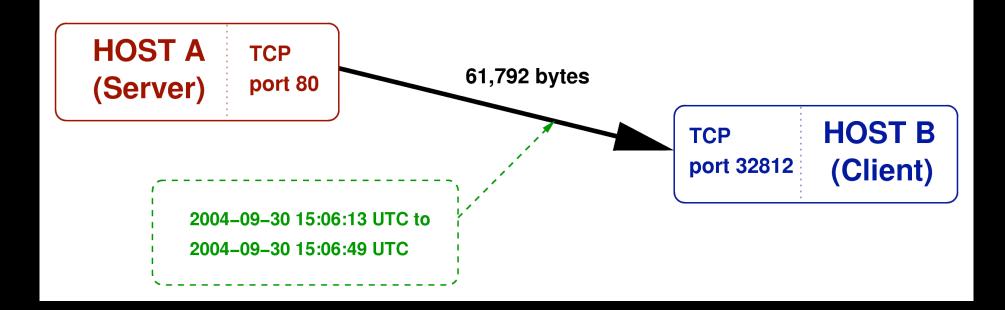
- Build networks from Internet traffic data
- Two proofs of concept:
 - Behavioral networks (host to host)
 - Application networks (app to app)

- 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

Internet2/Abilene

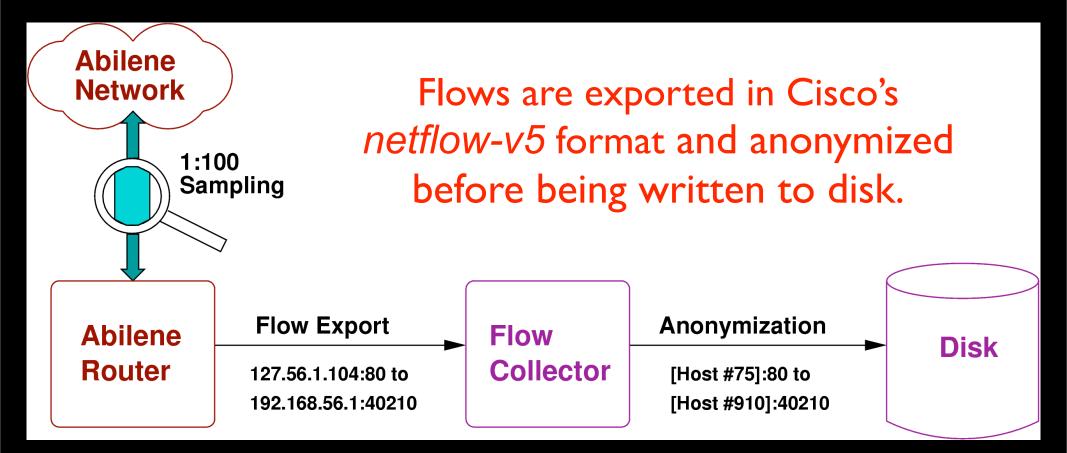


Network flow data



A successful TCP session contains *two* flows

Flow collection

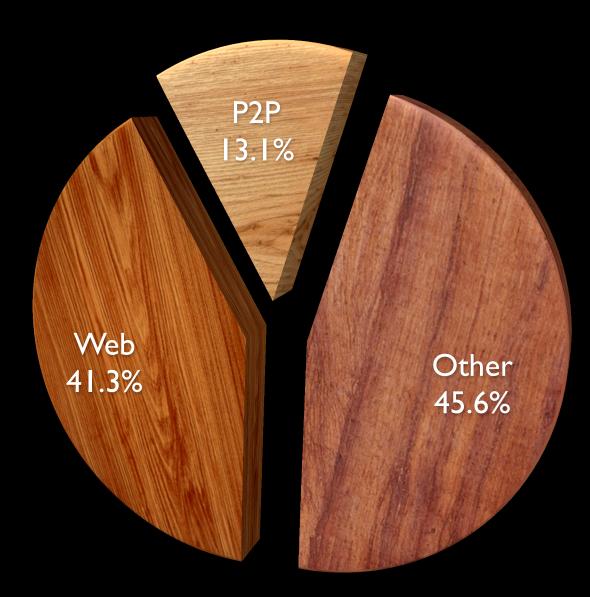


In a typical day: 35GB raw data at 3.1 Mbps

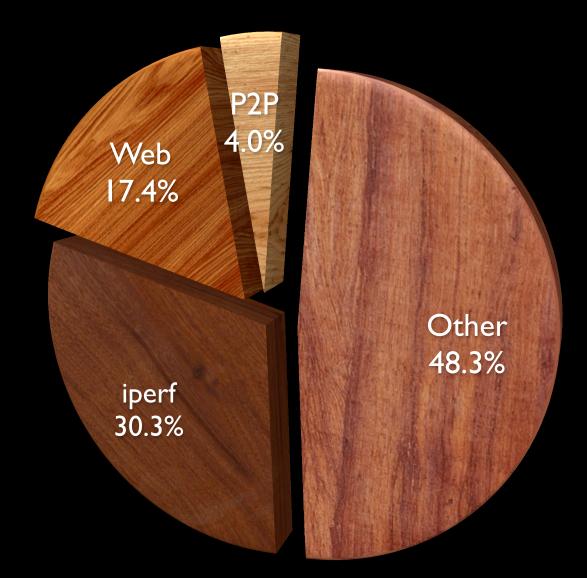
Data set for analysis

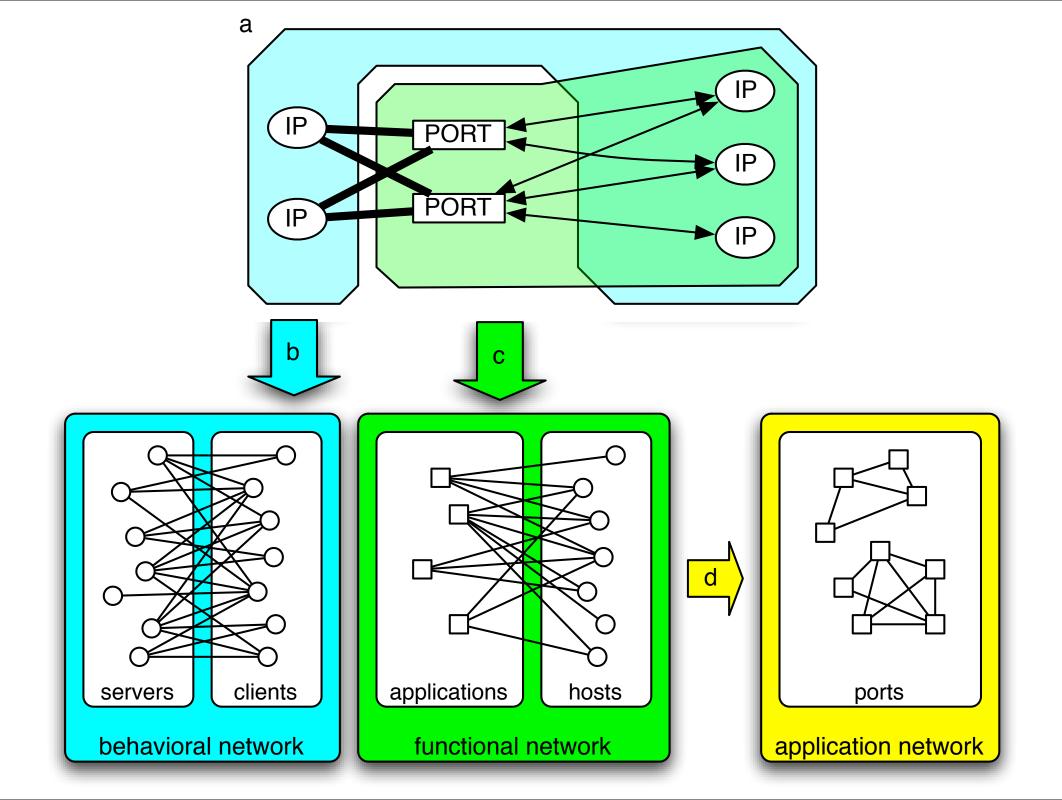
- Full 24-hour day of network flow data starting at 2005-04-14 05:00:00 UTC
 - a typical day
 - 600M flows
 - I5M unique hosts
 - 2TB (1% sampling)

Flows

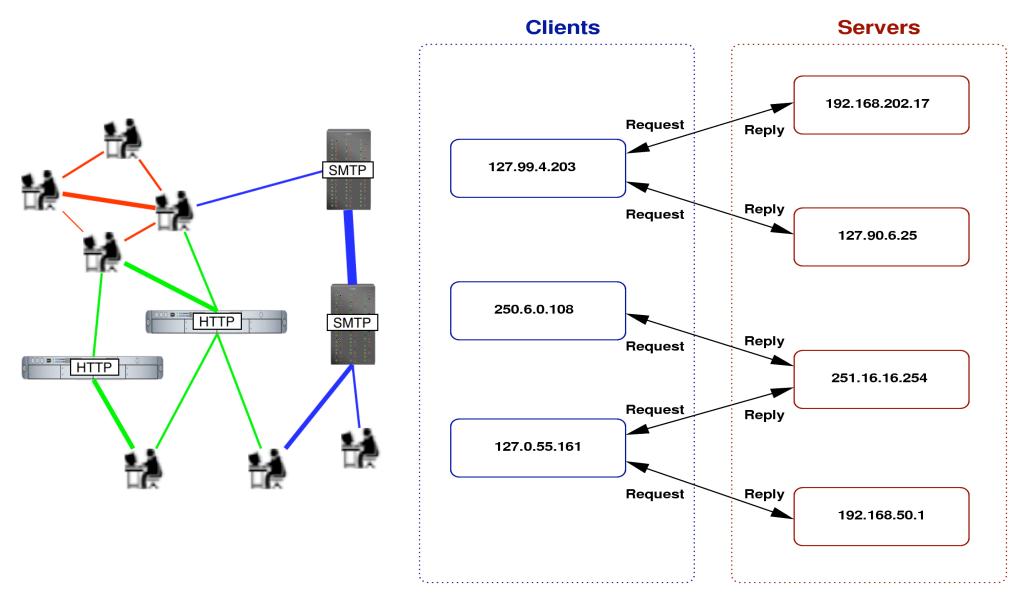


Traffic



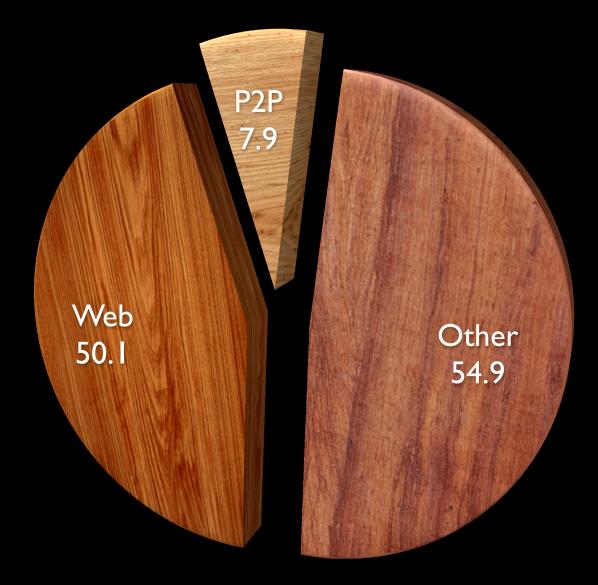


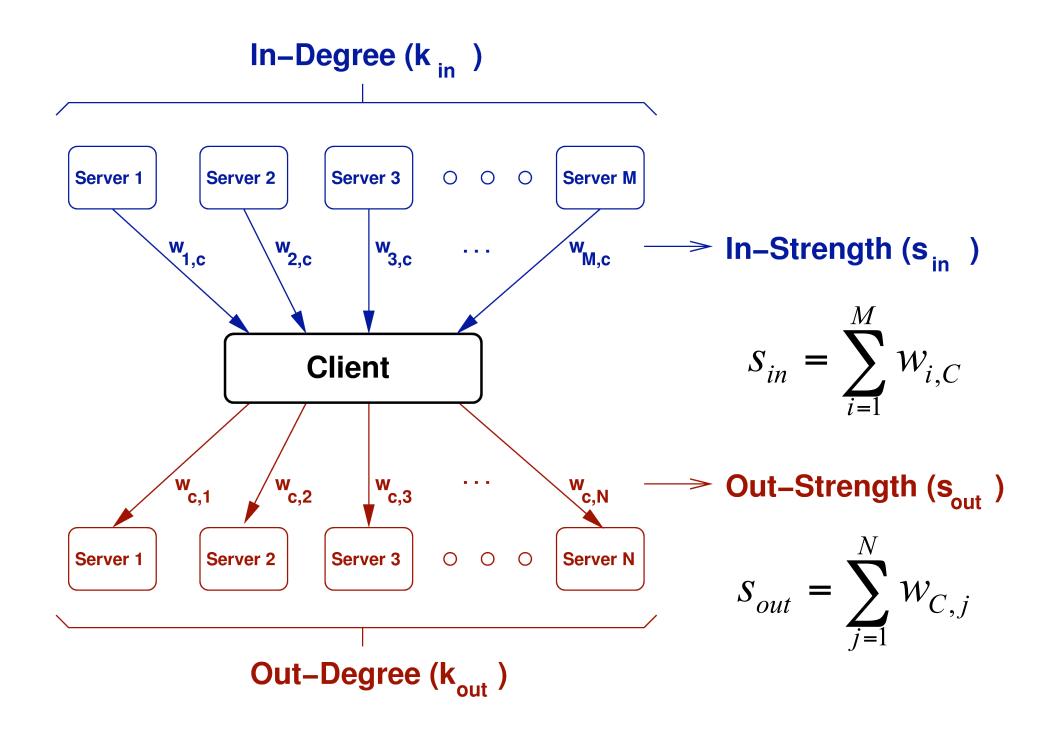
Behavioral networks



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

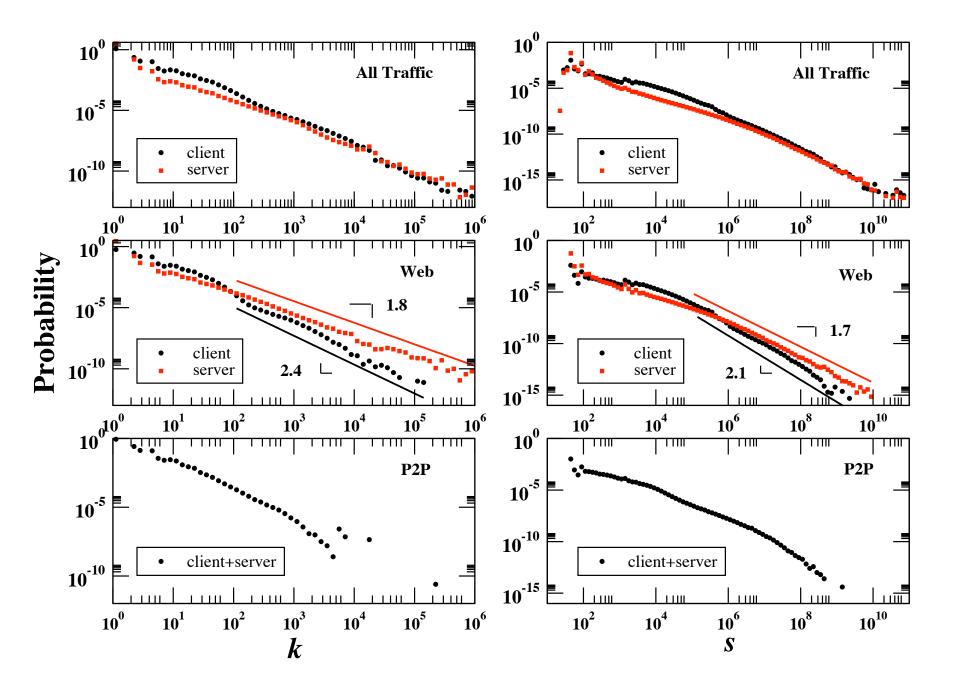
Edges (x 10⁶)



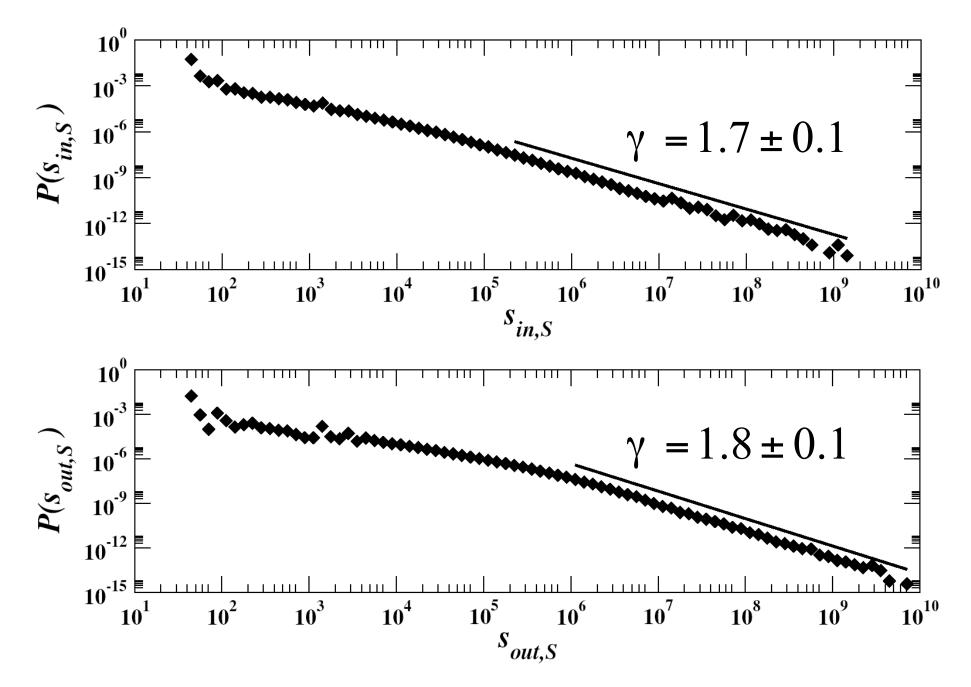


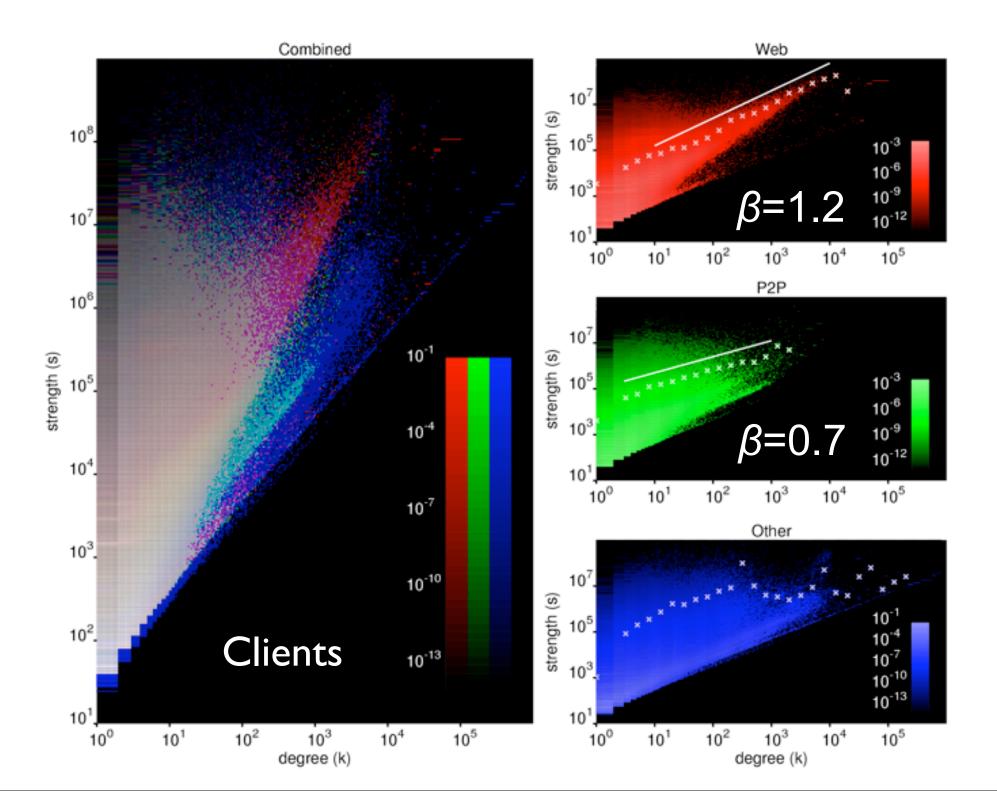
Degree

Strength

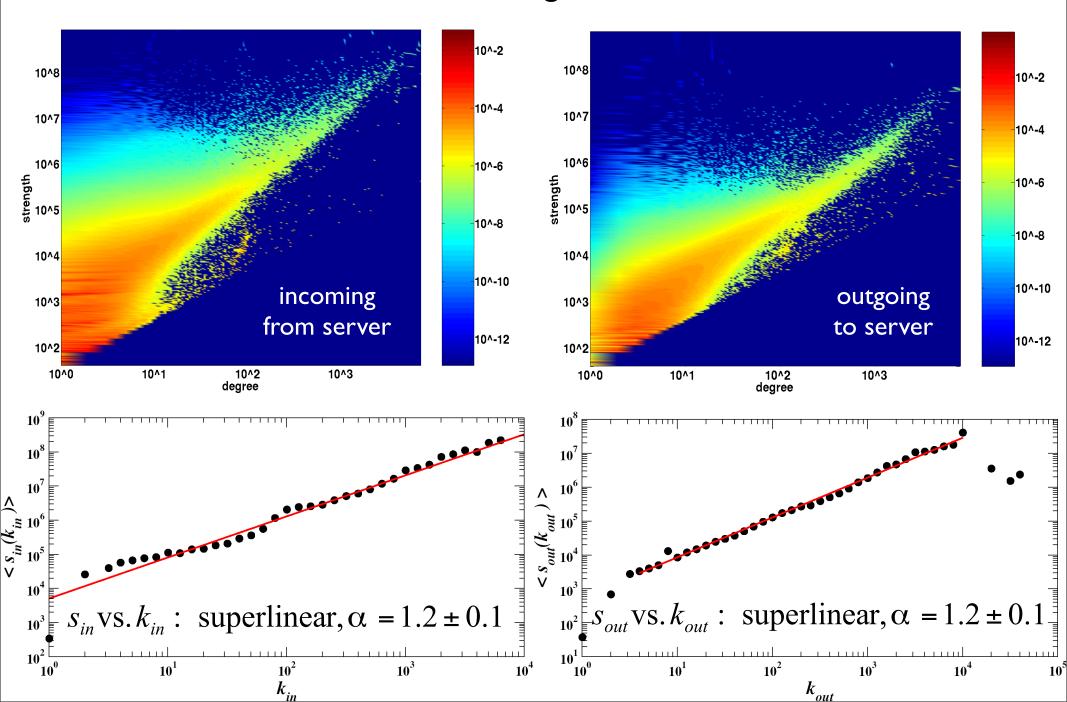


Web servers: Strength distributions



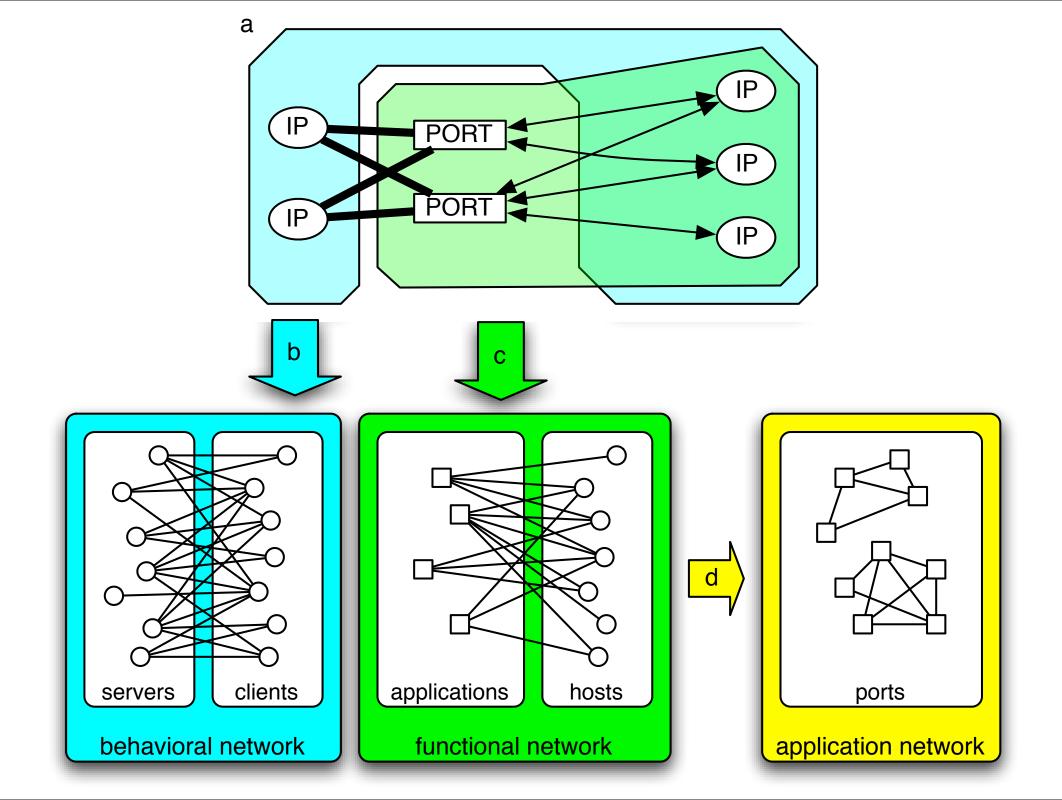


Web scalability problem: The more servers a client contacts, the more data it exchanges with **each** server!

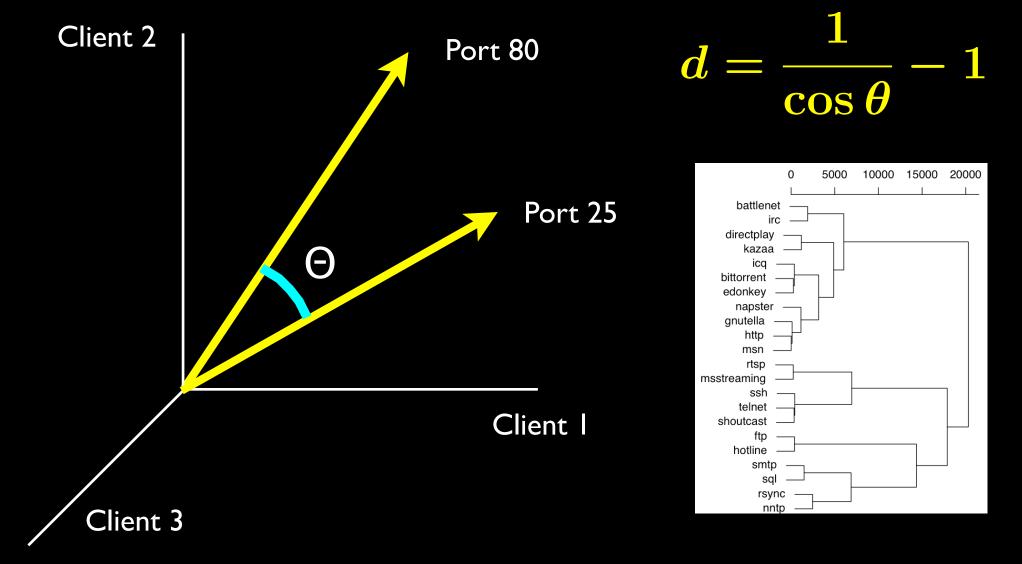


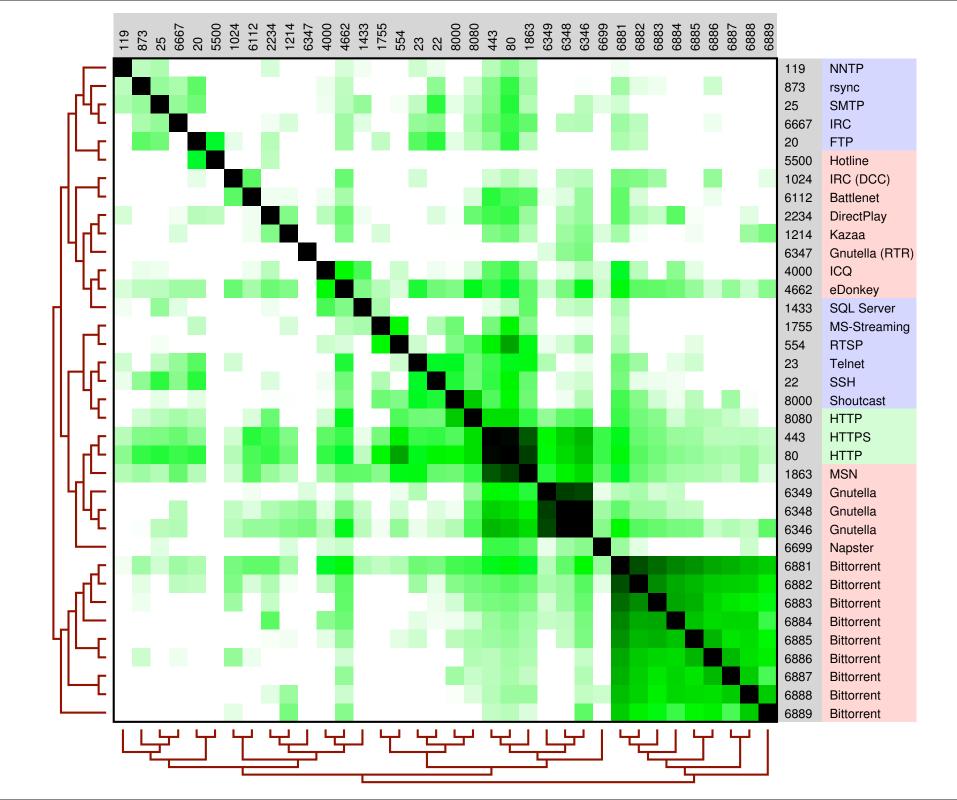
Summary

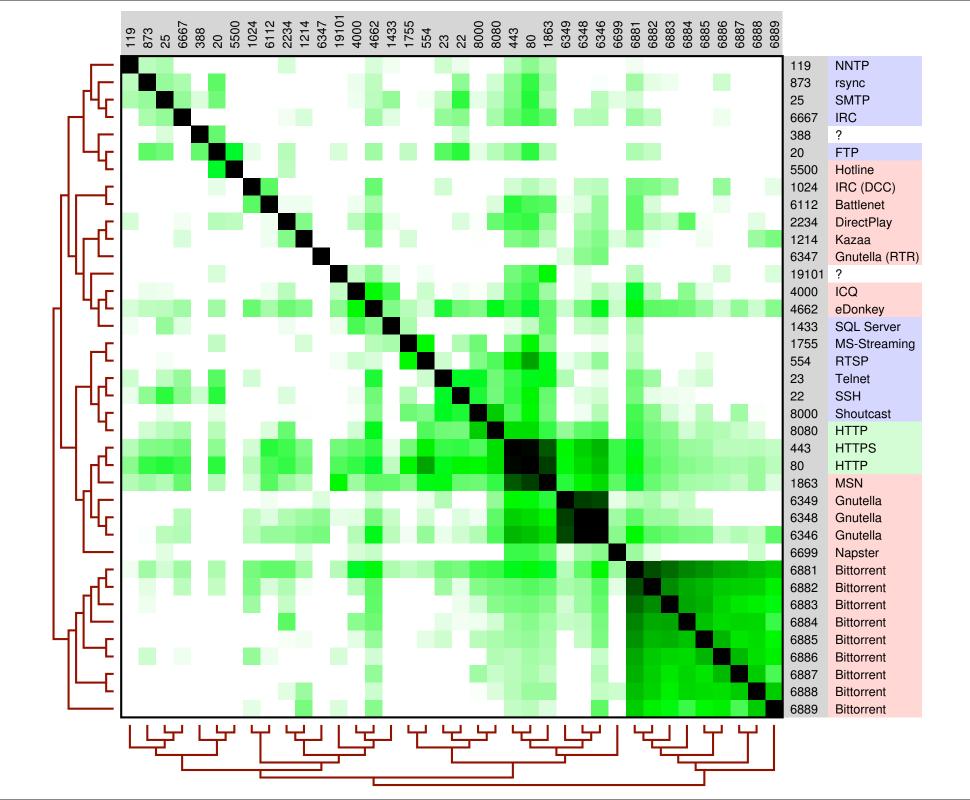
- **Power-law distributions** are found in all aspects of traffic networks: degree, strength, and weight distributions for both clients and servers
 - The strength distribution for Web servers lacks any mean value
 - The relationship between **degree** and **strength** for Web clients is super-linear
 - **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



From functional to application clusters







| Unknown port | Application |
|--------------|-------------|
| 388 | Unidata/LDM |
| 19101 | Clubbox |

| Port | Арр | 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

| 1.00 | | | |
|-----------|------------------------------|------------------|--|
| 15- | anale ((decerce)) | with then people | |
| ma | Del.icio.us | | |
| WIKIPEDIA | From Wikipedia, the free and | pologendia | |
| | | | |

The Free Encyclopedia arget of Main Page Cushmatry Portal Featured articles · Carrent executo a Success changes · Raidar atchi

 Insta Control Wikgedu · Annalised 440.0

Ge Seath of the late · What lists have Balatel charges · topland file · Special pages · Property and 4 Pairs pear bit.

| ٠ | Cit #11 alboh |
|---|-------------------|
| - | other Cangeragets |
| | Catelli |
| | From P |
| | Expañdi |

| | The commond state | anna kap Mikinda (ping) | |
|--|-------------------|-----------------------------|--|
| Del.icio.us | | | |
| From Wikipolia, the free excaptionalia | | | |
| The correct title of this article is delatonase. The initial letter is capitalized due to recting a entity | TRANK. | | |
| The website delictions (processing) as "delicious") is a social bosimulting, social software web | | del.icio.us | |
| service for storing and sharing web loostmarks. The site came online in take 2003 and was developed by Joshua Schachter, co-maintainer of Memopool, | 1.4 | del.icio.us | |
| Ewything perted to deticle as is publicly viewable by default, although a user can mark specific | | Generotus | |
| teolocalis as private, and mass-imported bookmarks are private by default. The public aspect is emphasized: it is not intended to be a tool for stoning private bookmark collections. Many people use | Commonial? | 80 | |
| their delicts as RSS local to publish "soliblegs" an their <u>writings</u> . | Type of site | Online usual built-starting | |
| A nam hierarchital keyword categorization system is used on delicits as where users can tag each of | Owner | Valuet Int. | |
| their benkmarks with a number of freely chosen keywords (cf. Icilizonomy). A contributed vorw of | Created by | Joshua Scharbrer | |

everyone's bookmarks with a given tag is available; far instance, the URL "Impublic in unitagivitied" displays all at the most recent links tapped "win", its collective nature also makes it possible to view bookmarks added by similar-minded users.

del.ico os has a hum page "hotto", "popula" and "hotent" pages that help to make the website an effective conveyor of pegulat internat moment and bends.

The website's unconventional domain name, known as a domain hack, its simple HTML interface with human readable URLs, as well as a REST API and RSS feeds for web syndication, have contributed to making it one of the most popular services of its kind.

Use of dolicits us is currently bee. The source code of the site is not available, but a user's new entered data is beenly downloadable through the SP (P

There exist a few open source closes of deficition with an de brown and submits in fit, as well as many discontening Social bookmarking services such as [TritList @, Full@ and Souri @

details us was acquired by Vahoot on Fislay, December 9, 2005. [1] @ Variant guesses suggest 8 was sold for somewhere between \$25-39 million US dollary Brinner worked. This would be Yaboo's second purchase of an internet company using a domain back, the go # being their first. Contrate Point

Not all dynamics are captured

Clicks?

:80

del.icio.us social bookmarking

» all your bookmarks in one place

» bookmark things for yourself and friends

» check out what other people are bookmarking

laism more

» get started «



www.BrazilanArtists.net, the A-2 of Brazilian Arts. Entertainment & Cultural Events in the UK Wkipedia Public domain image resources - Wkipedia.

Use YouTube to learn Photoshop

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





