Flexible Traffic and Host Profiling via DNS Rendezvous

SATIN 2011
April 4, 2011

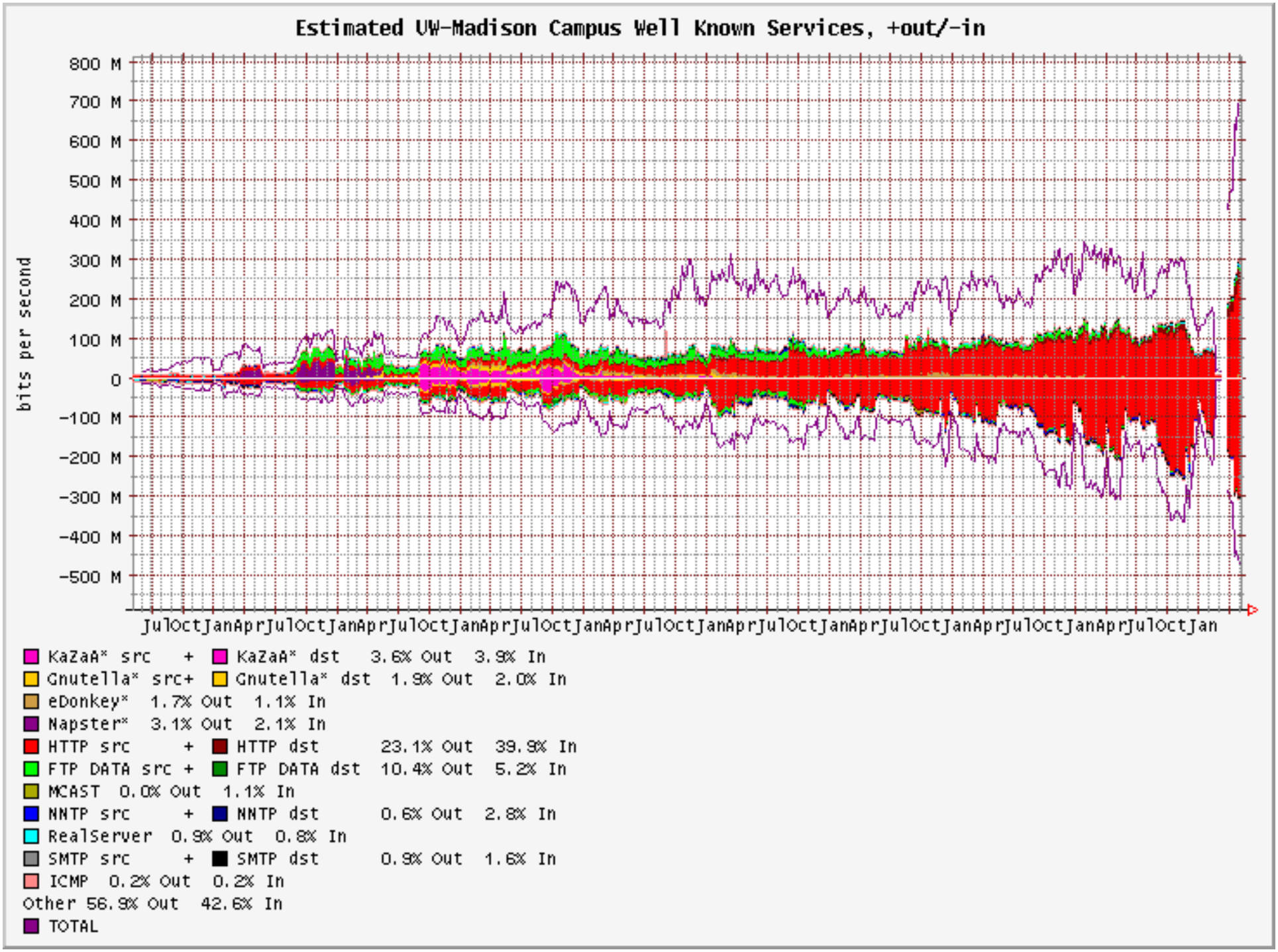
THE UNIVERSITY of WISCONSIN MADISON

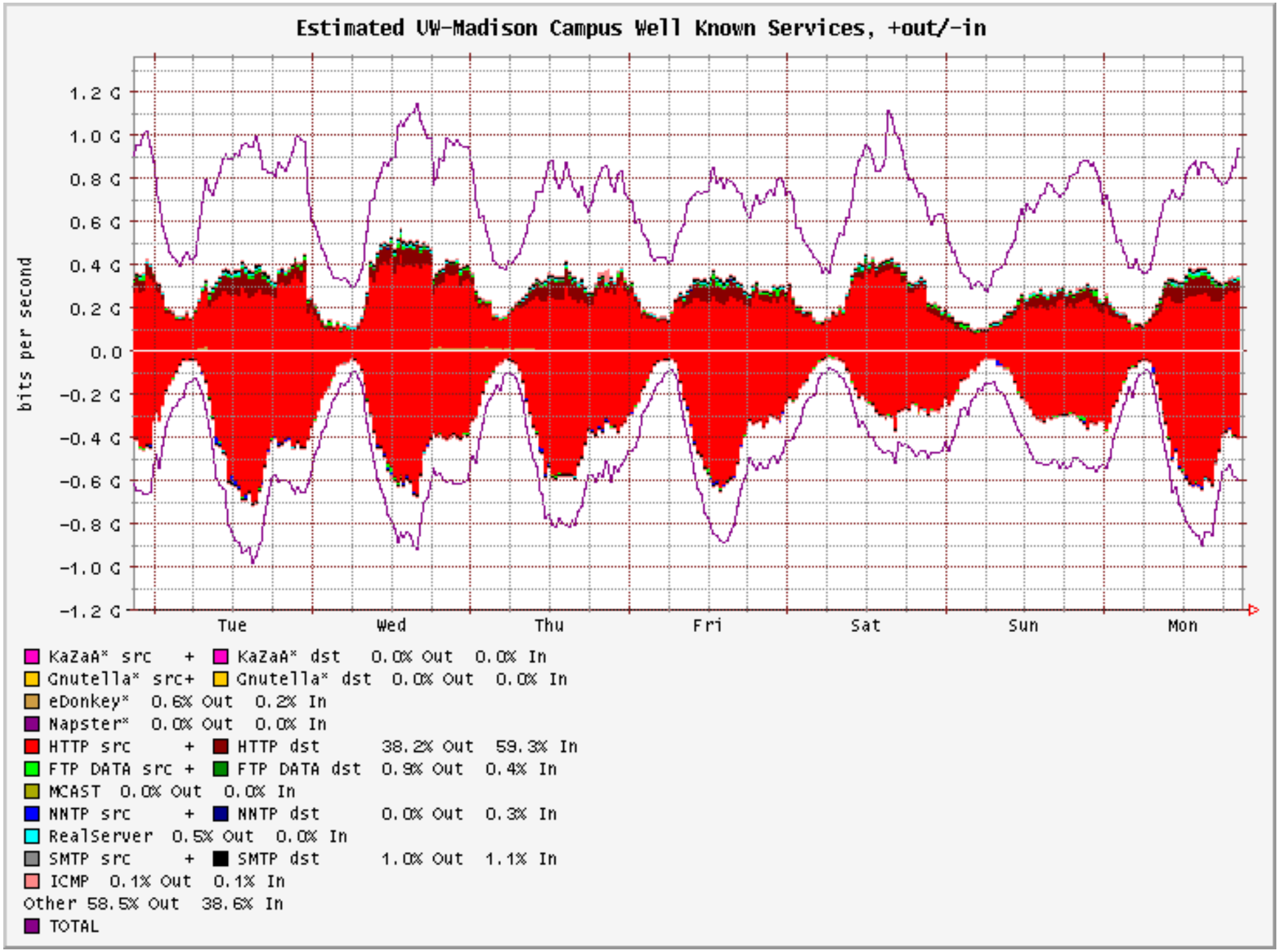
David Plonka
&
Paul Barford
{plonka,pb}@cs.wisc.edu

Traffic Classification Challenges

Accurate classification is an open problem; timely classification is desirable.

- New and evolving applications and protocol reuse
- Increased forwarding speeds and higher-capacity links
- Obscured or encrypted traffic, to sidestep service limitations or for user privacy


Estimated UW-Madison Campus Well Known Services, +out/-in
bits per second
800 M
700 M
600 M
500 M
400 M
300 M
200 M
100 M
0
-100 M
-200 M
-300 M
-400 M
-500 M
JulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJanAprJulOctJan
KaZaA* src + KaZaA* dst 3.6% Out 3.9% In
Gnutella* src+ Gnutella* dst 1.9% Out 2.0% In
eDonkey* 1.7% Out 1.1% In
Napster* 3.1% Out 2.1% In
HTTP src + HTTP dst 23.1% Out 39.9% In
FTP DATA src + FTP DATA dst 10.4% Out 5.2% In
MCAST 0.0% Out 1.1% In
NNTP src + NNTP dst 0.6% Out 2.8% In
RealServer 0.9% Out 0.8% In
SMTP src + SMTP dst 0.9% Out 1.6% In
ICMP 0.2% Out 0.2% In
Other 56.9% Out 42.6% In
TOTAL


Estimated UW-Madison Campus Well Known Services, +out/-in
bits per second
1.2 G
1.0 G
0.8 G
0.6 G
0.4 G
0.2 G
0.0
-0.2 G
-0.4 G
-0.6 G
-0.8 G
-1.0 G
-1.2 G
Tue
Wed
Thu
Fri
Sat
Sun
Mon
KaZaA* src + KaZaA* dst 0.0% Out 0.0% In
Gnutella* src+ Gnutella* dst 0.0% Out 0.0% In
eDonkey* 0.6% Out 0.2% In
Napster* 0.0% Out 0.0% In
HTTP src + HTTP dst 38.2% Out 59.3% In
FTP DATA src + FTP DATA dst 0.9% Out 0.4% In
MCAST 0.0% Out 0.0% In
NNTP src + NNTP dst 0.0% Out 0.3% In
RealServer 0.5% Out 0.0% In
SMTP src + SMTP dst 1.0% Out 1.1% In
ICMP 0.1% Out 0.1% In
Other 58.5% Out 38.6% In
TOTAL

Prior Classification Work

- Transport-based analysis
  - *e.g.,* FlowScan [Plonka, '00], [Fullmer, *et al.,* '00]
- Payload-based analysis
  - *e.g.,* Snort [Roesch, '99], [Dews, *et al.,* '03]
  - Examine payloads for specific features
- Behavioral analysis
  - *e.g.,* BLINC [Karagiannis, *et al.,* '05]
  - Consider social/functional/transport characteristics
- Statistical/machine-learning-based analysis
  - *e.g.,* [Erman, *et al.,* '06]
  - Apply standard methods to transport features

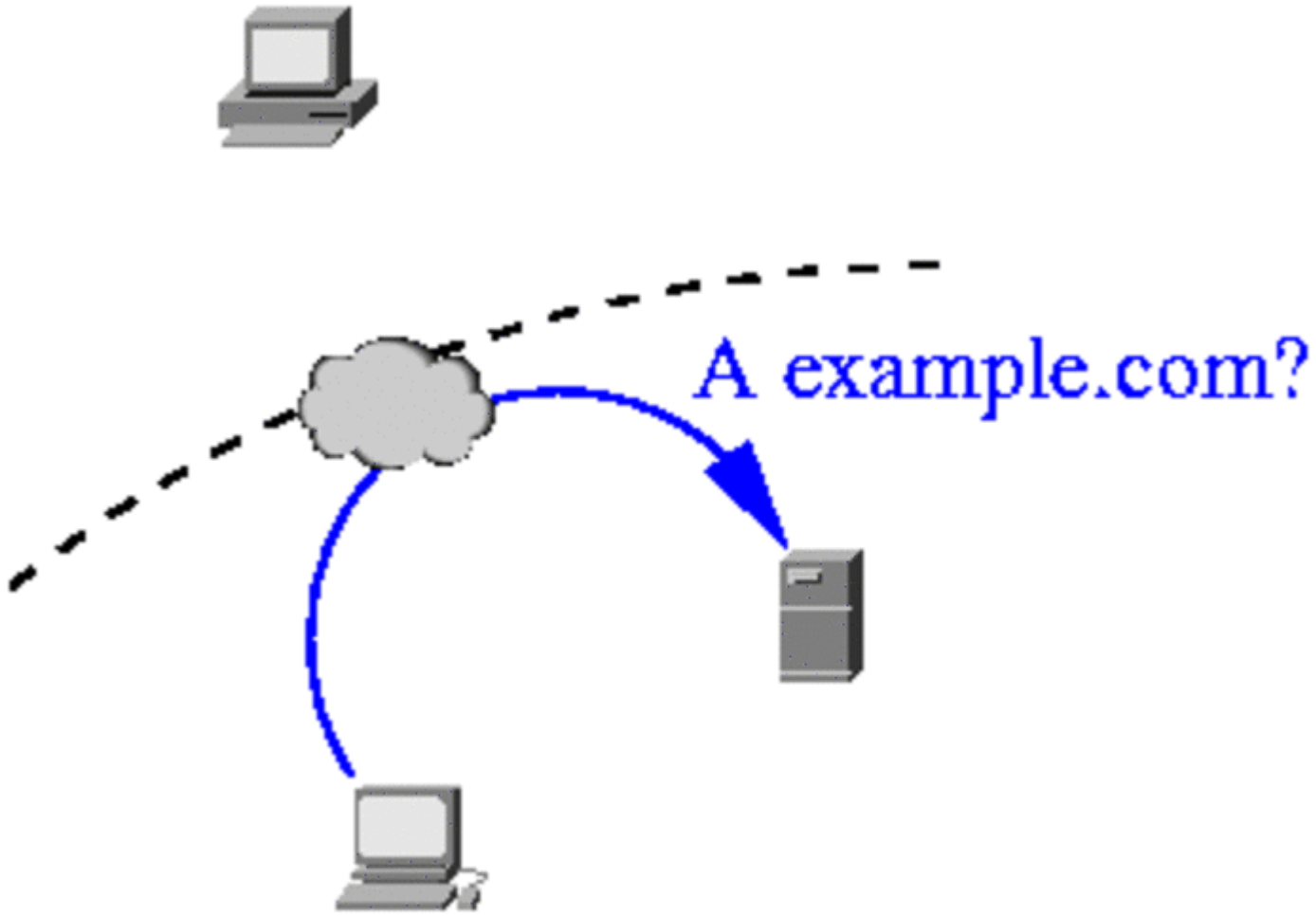
DNS Rendezvous-based Classification

- *rendezvous,* meaning “present yourselves”

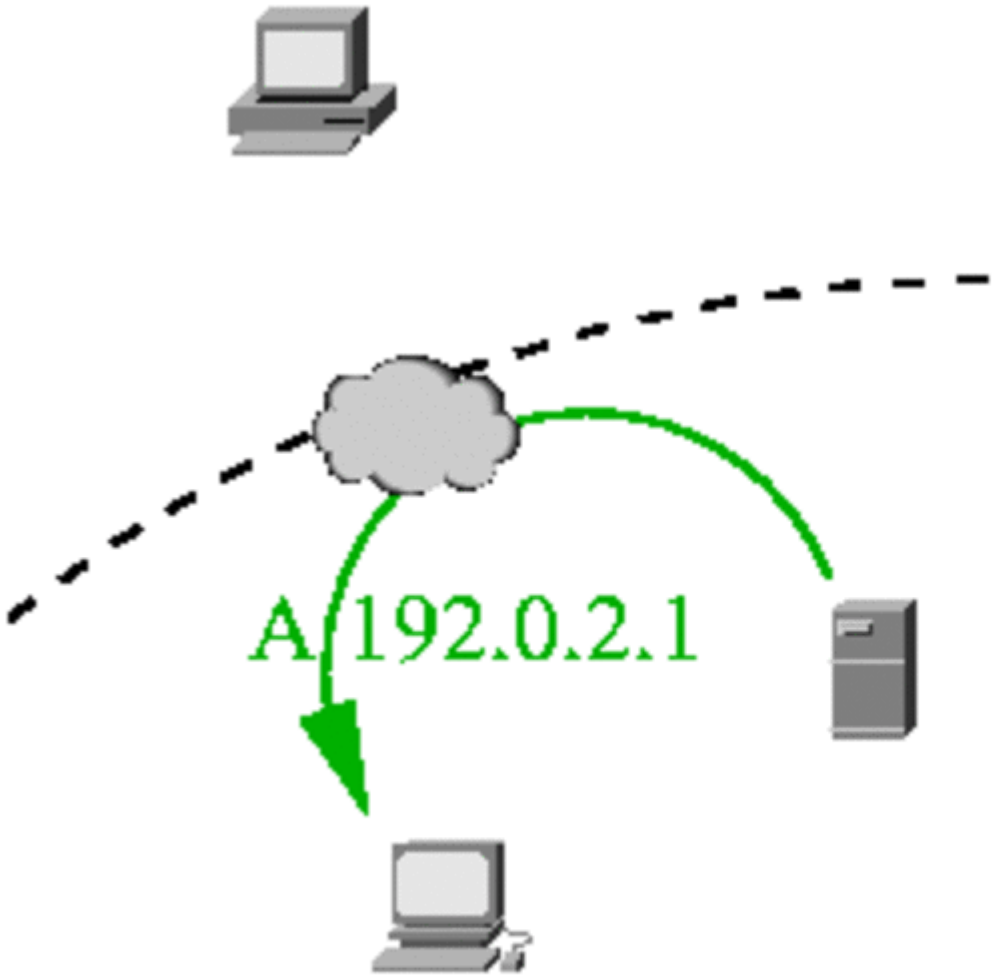
- ***Premise***: Internet hosts regularly use the DNS to find remote IP addresses of the hosts with which they might interact.
  - It is an ***easily separable*** “clear text” protocol.

- Hypothesis: We can inform and improve traffic classification by considering,
  “How does this host know that peer IP address?”

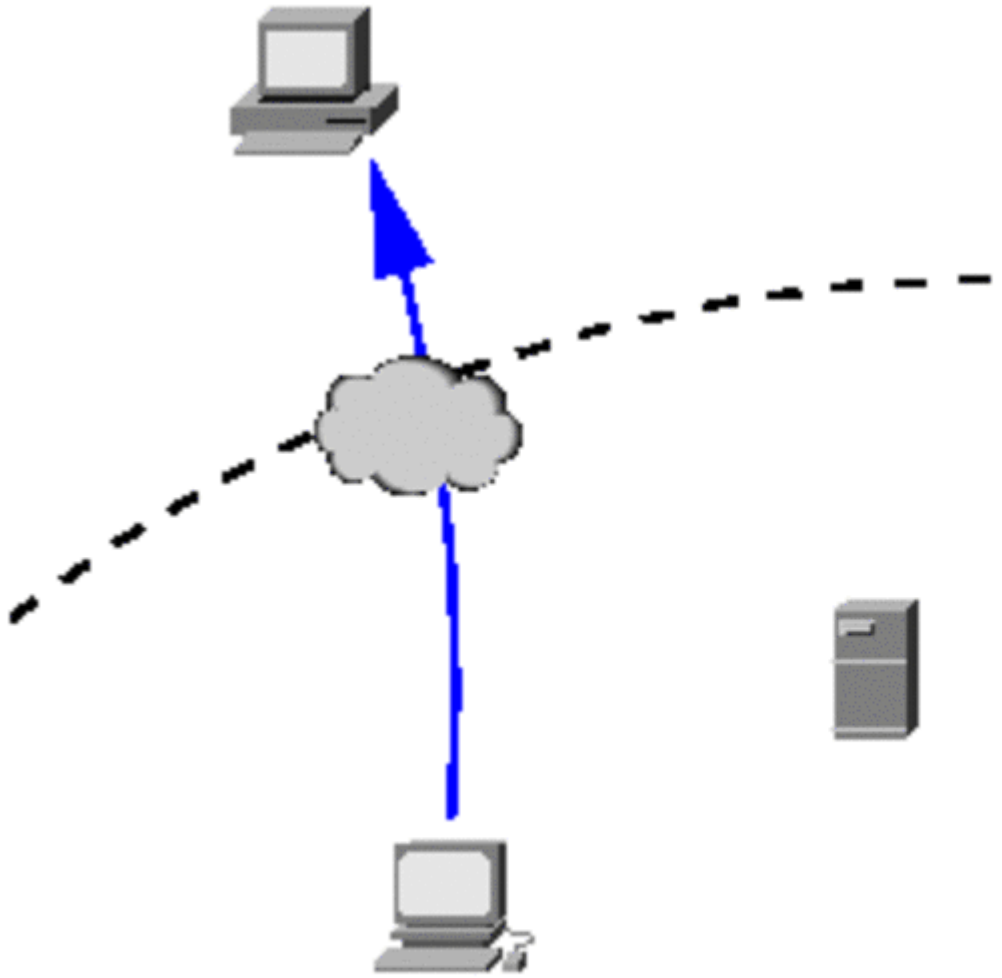
DNS Rendezvous: (1) Query



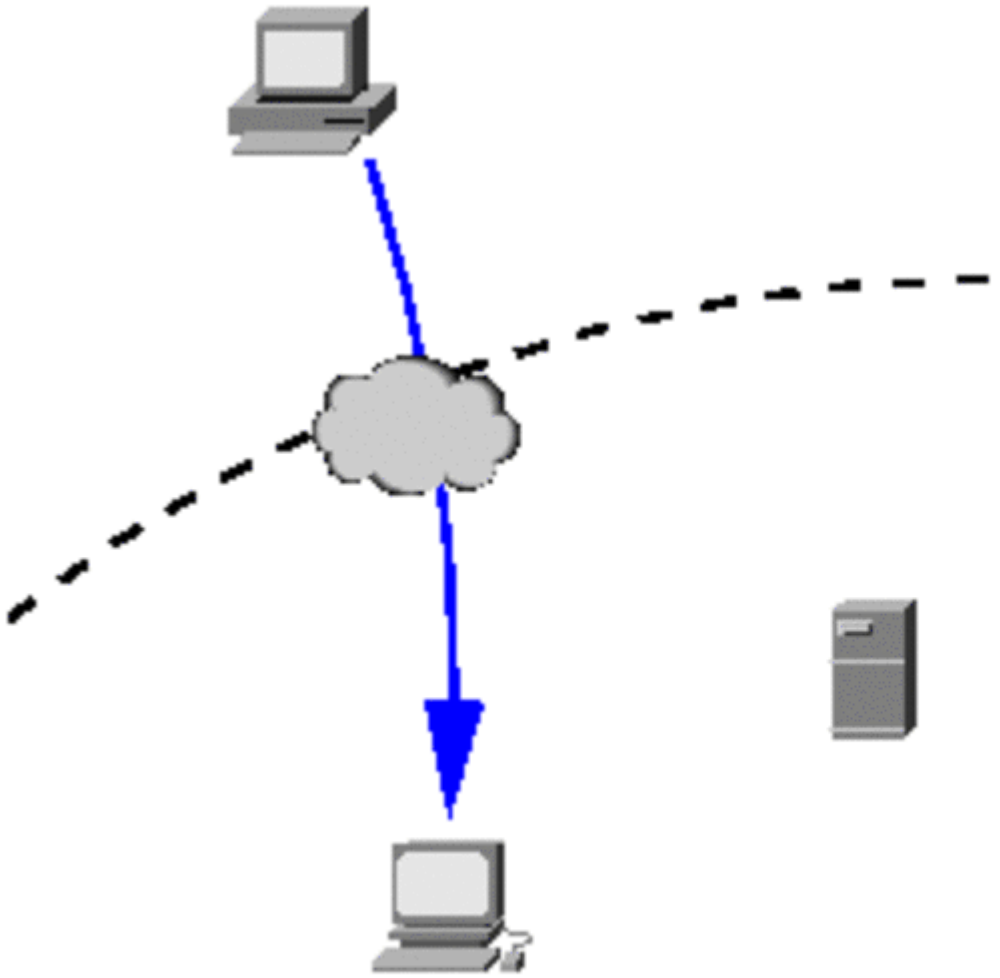
DNS Rendezvous: (2) Response



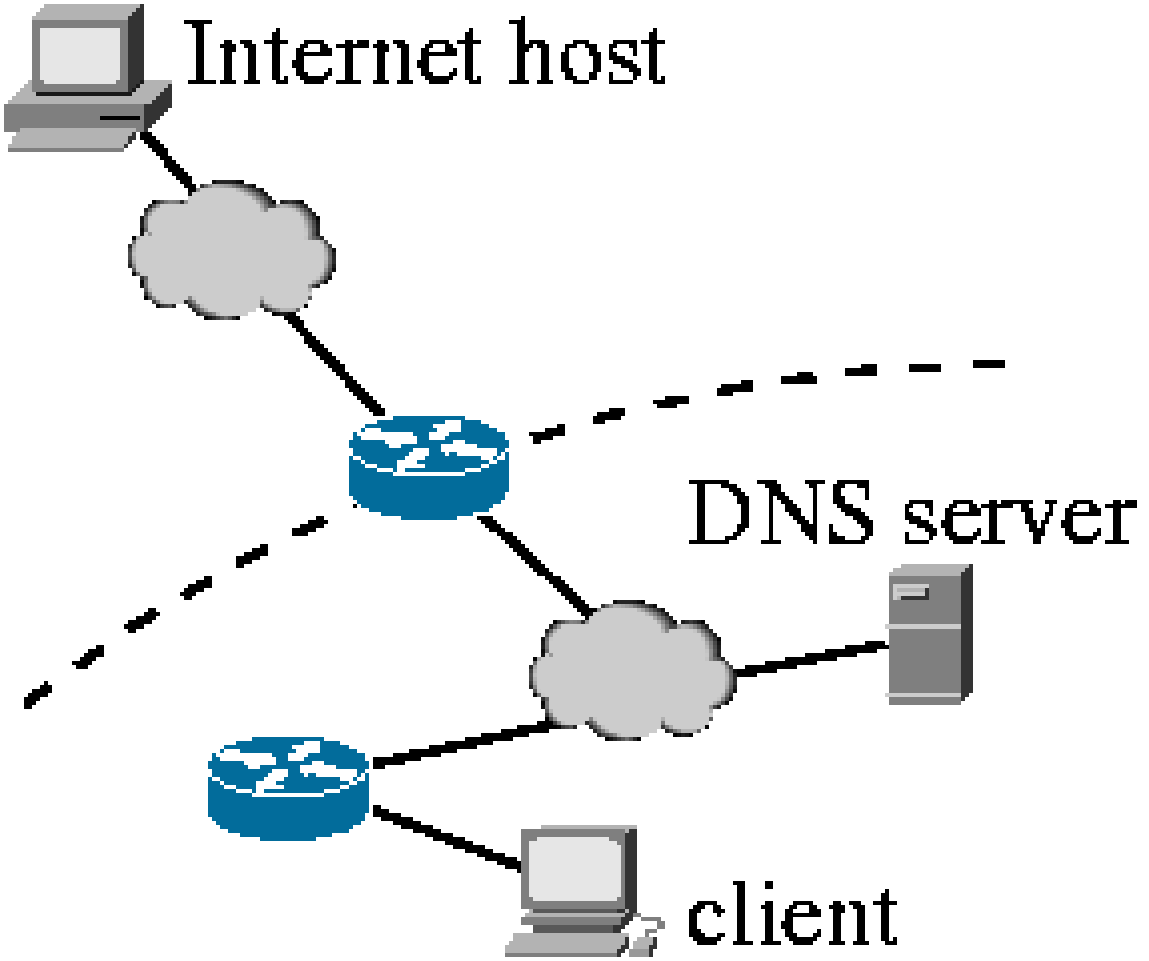
DNS Rendezvous: (3) Outbound



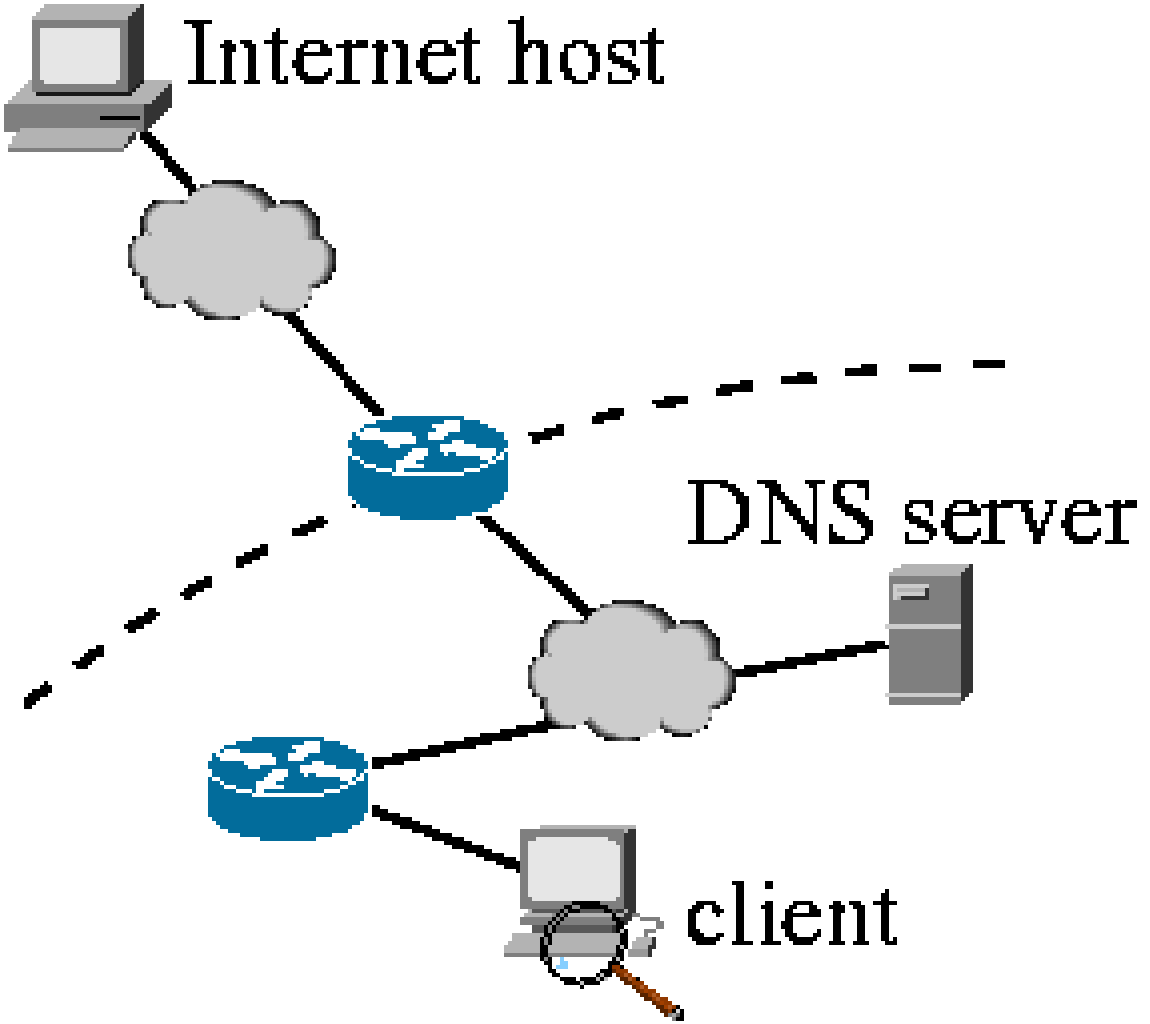
DNS Rendezvous: (4) Inbound



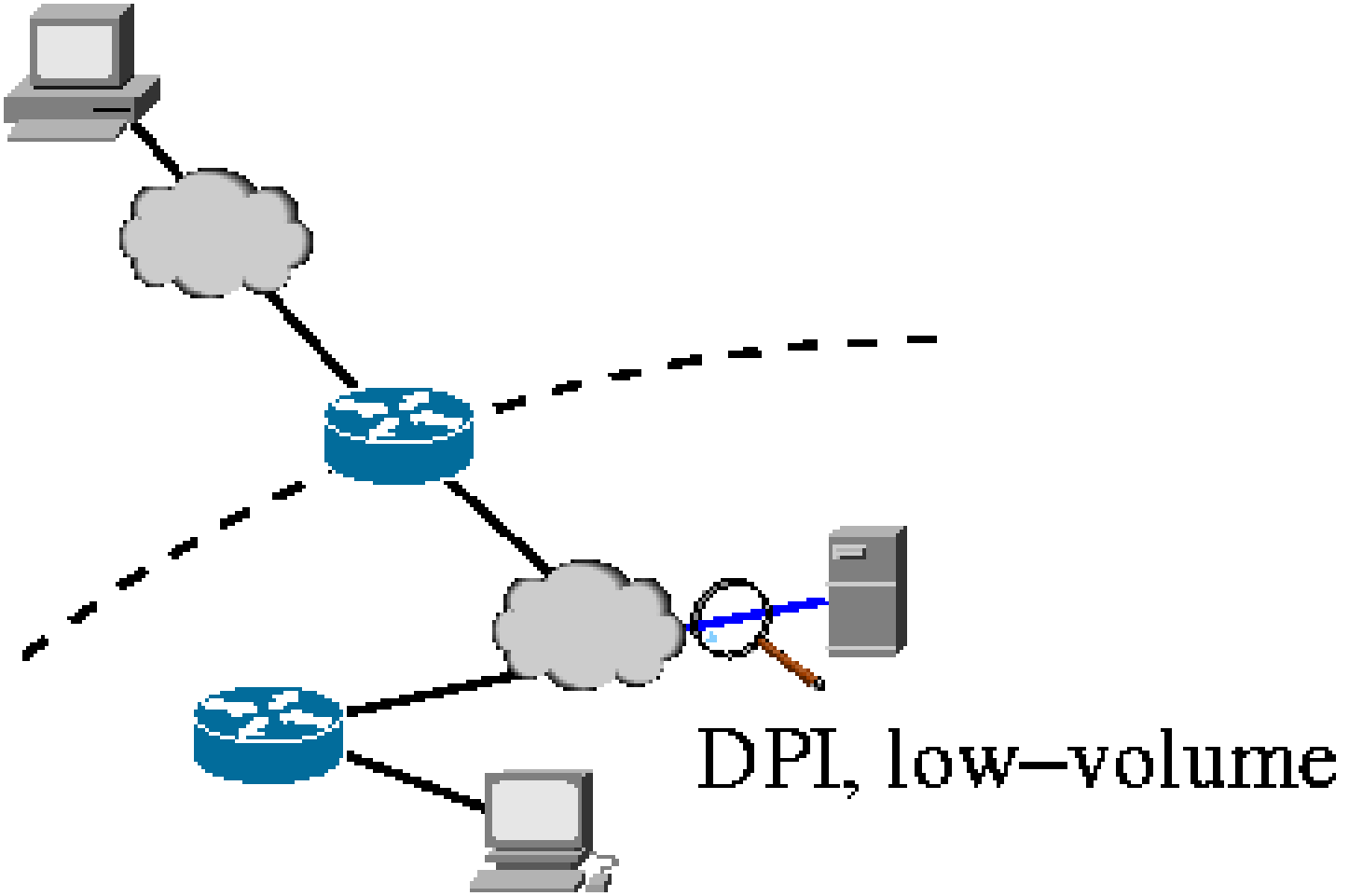
Traffic Observation Points



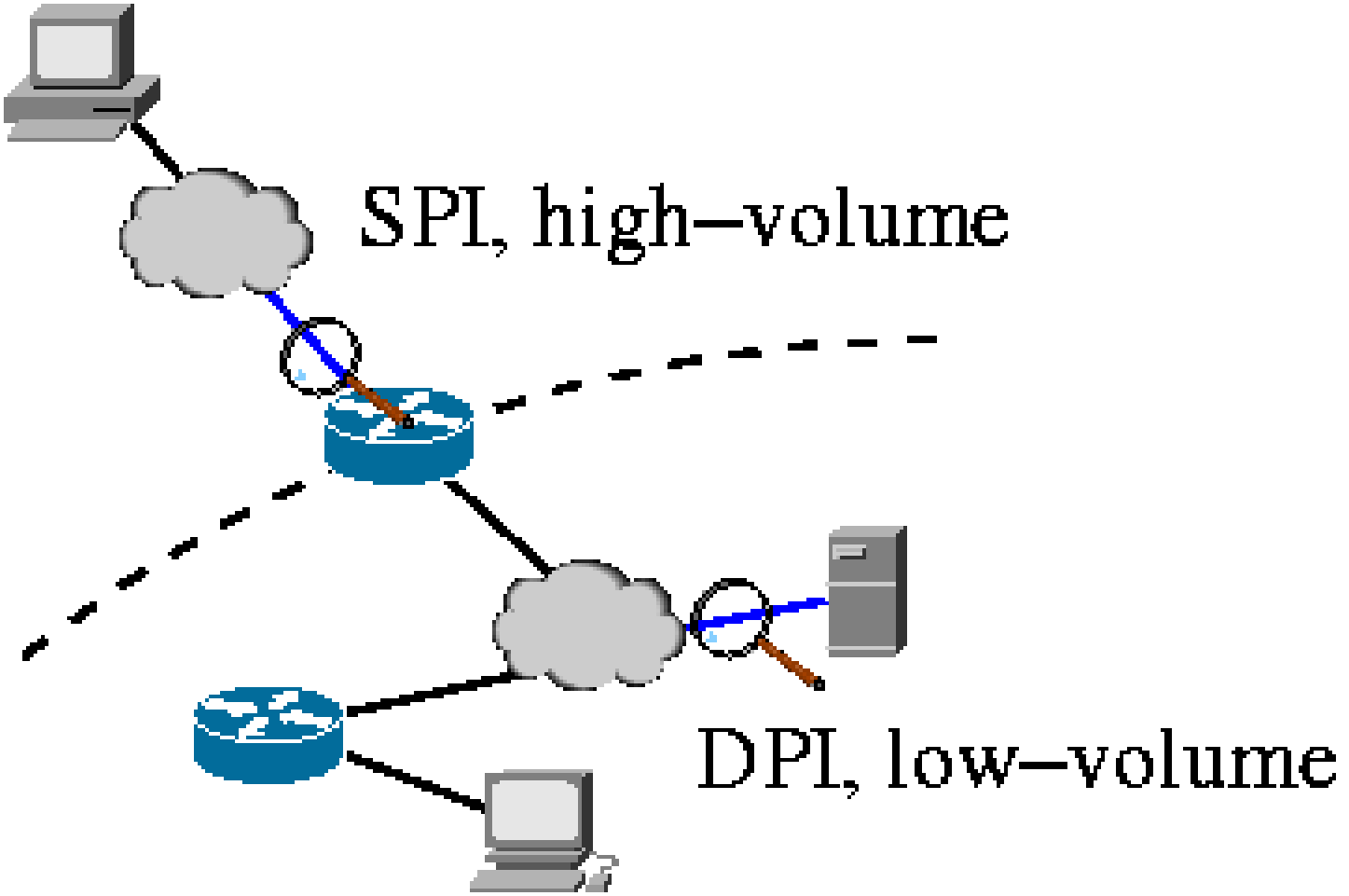
Traffic Observation Points



Traffic Observation Points

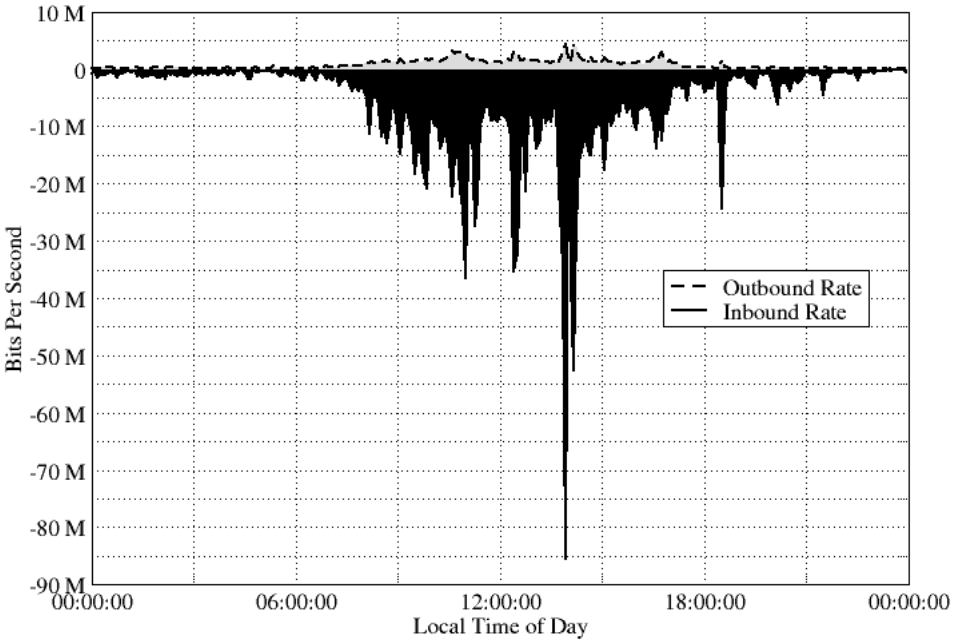


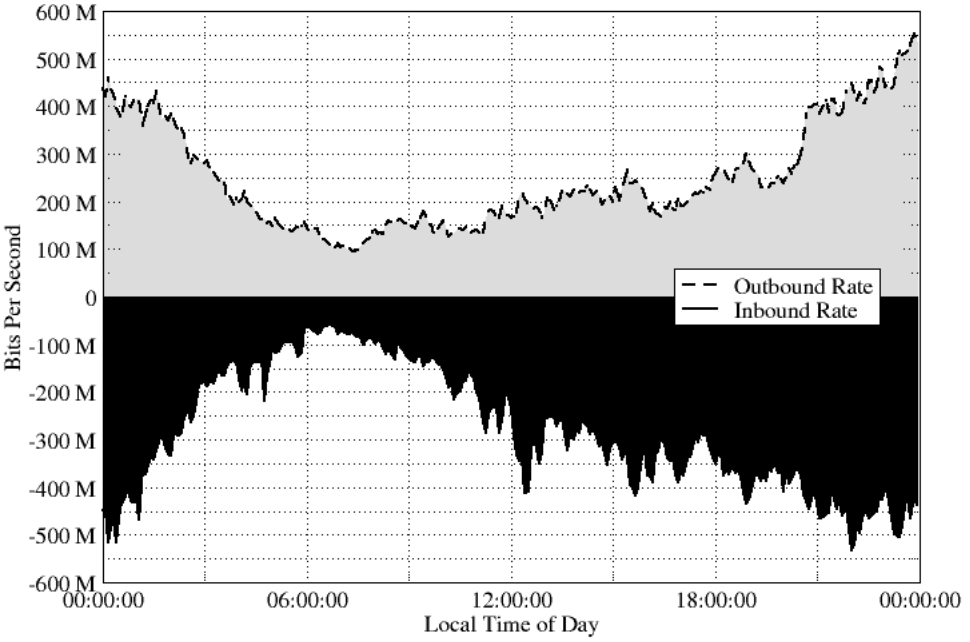
Traffic Observation Points



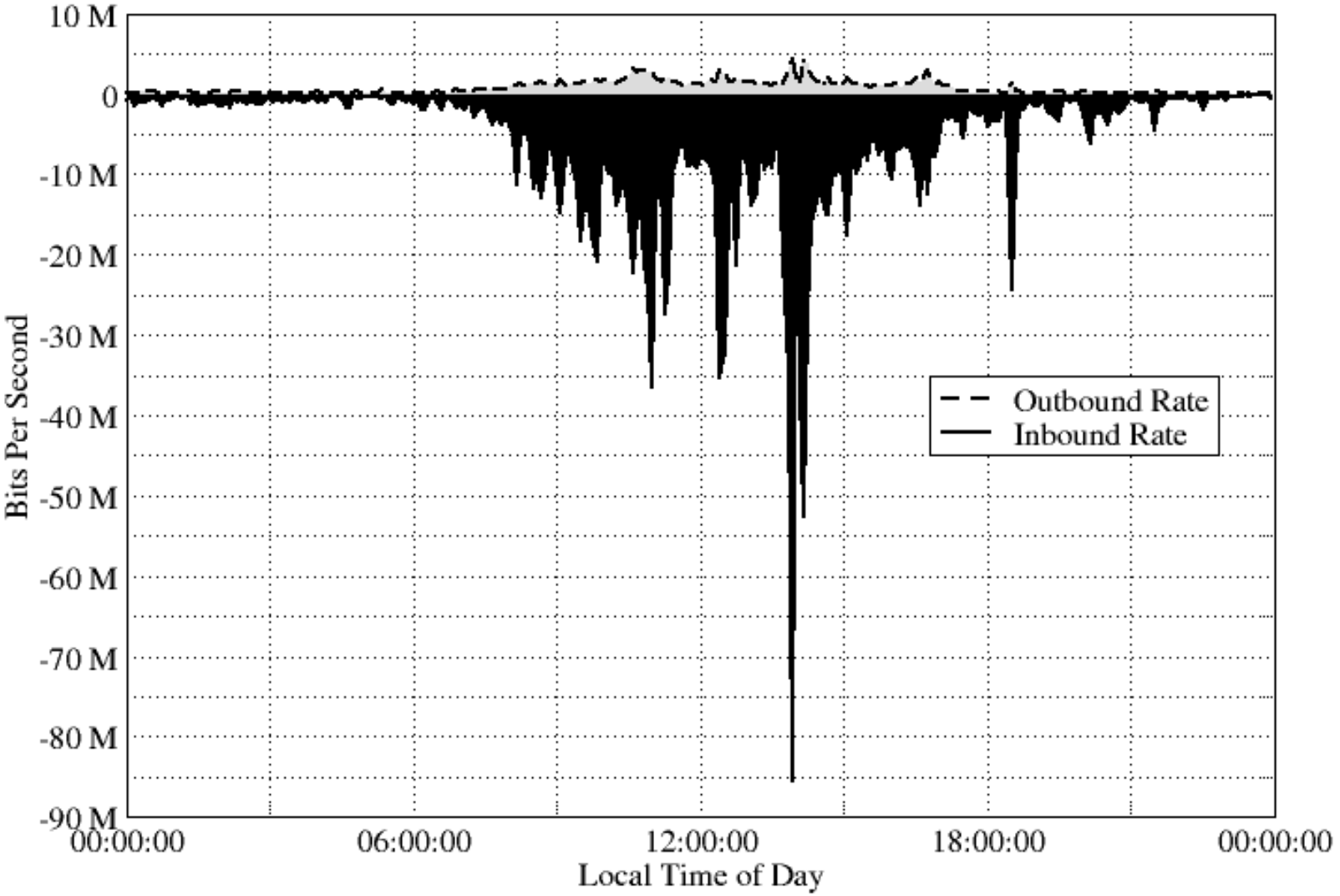
Characteristics of Data Sets

Data Set	Date	Day	Duration	Clients	Unique NOERROR FQDNs	DNS Reply Pkts	Average DNS Reply Utilization	Average Wide-Area Outbound / Inbound Utilization
Office	2009-04-17	Fri	24h	614	19.4 K	560 K	12.2 Kbps	753 Kbps / 5.66 Mbps
Residential	2009-04-17	Fri	24h	9,819 (5,344)	(143 K)	15.7 M	360 Kbps	244 Mbps / 276 Mbps

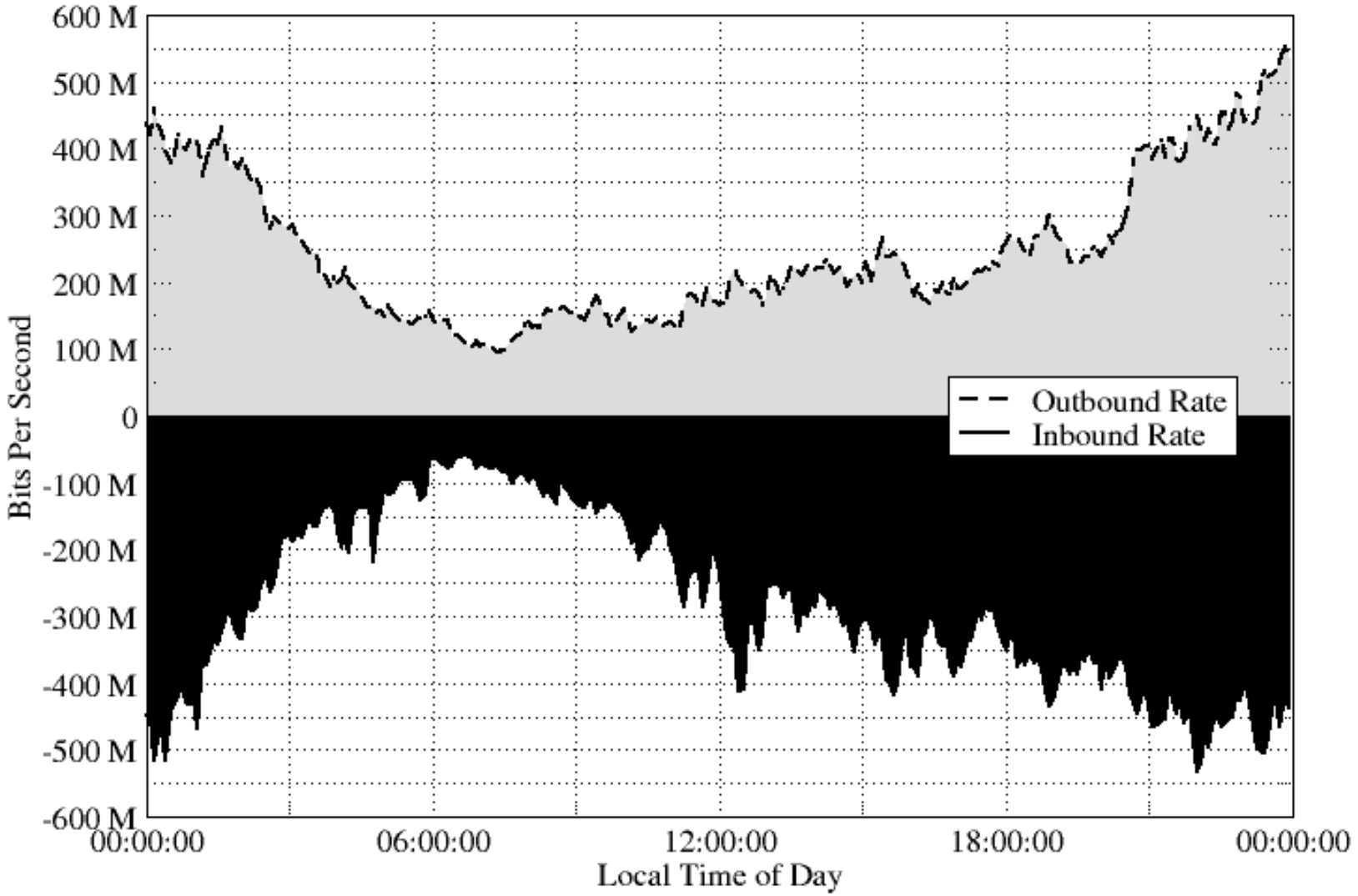




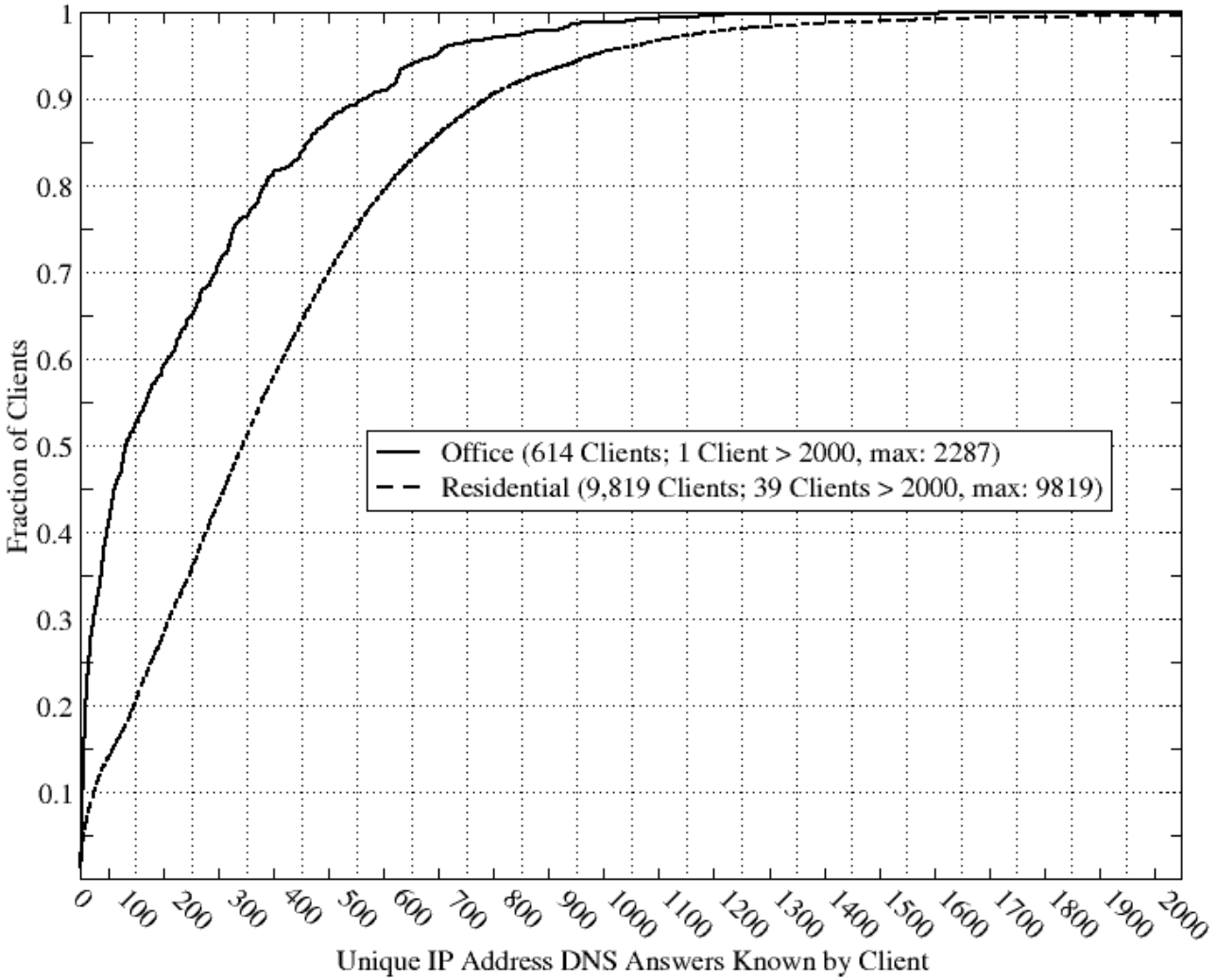
Office Wide-Area Traffic



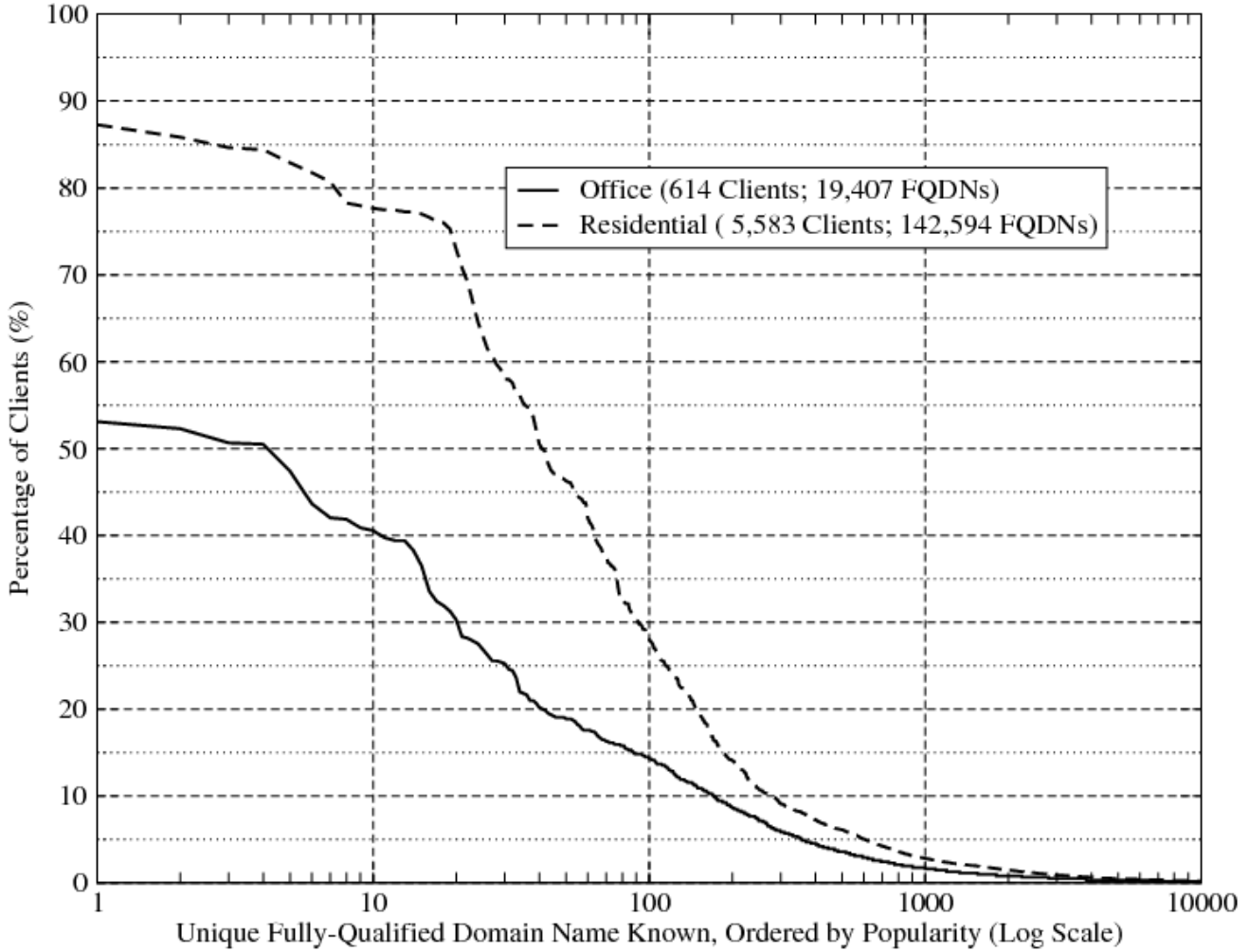
Residential Wide-Area Traffic



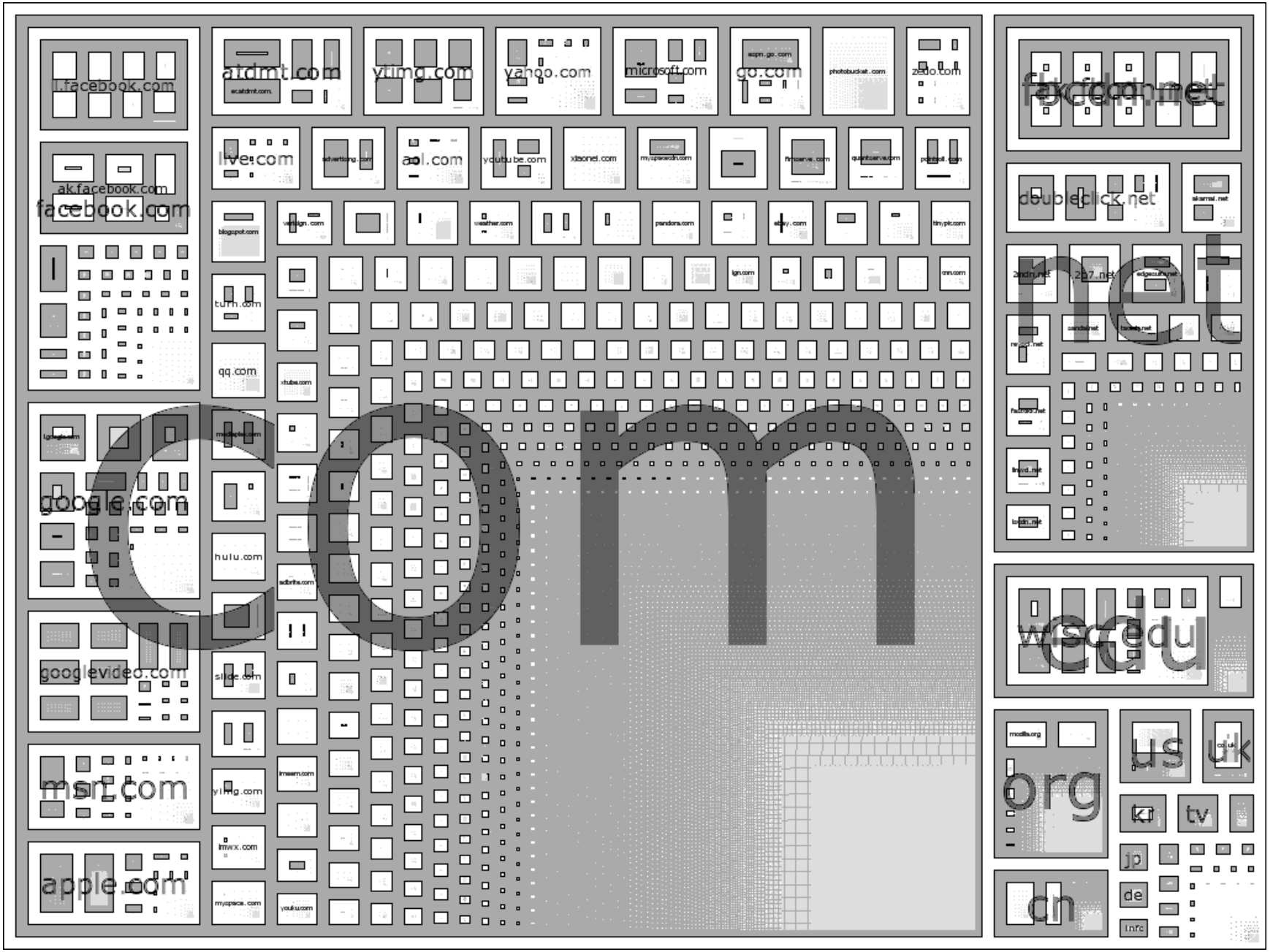
DNS Rendezvous Traffic Analysis: # of IP addrs known via DNS per client (1 day, CDF)



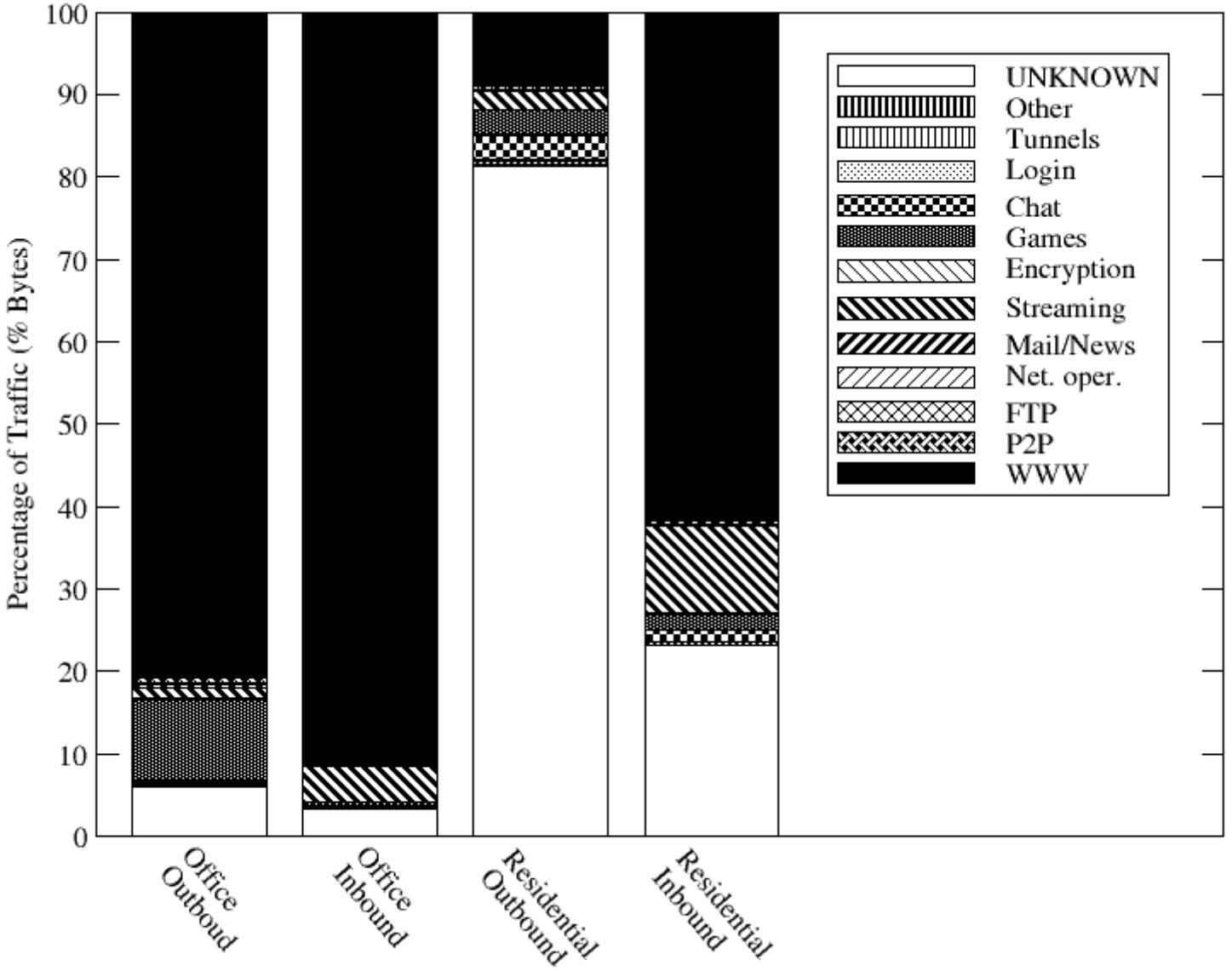
DNS Rendezvous Traffic Analysis: FQDN Popularity by client (1 day)



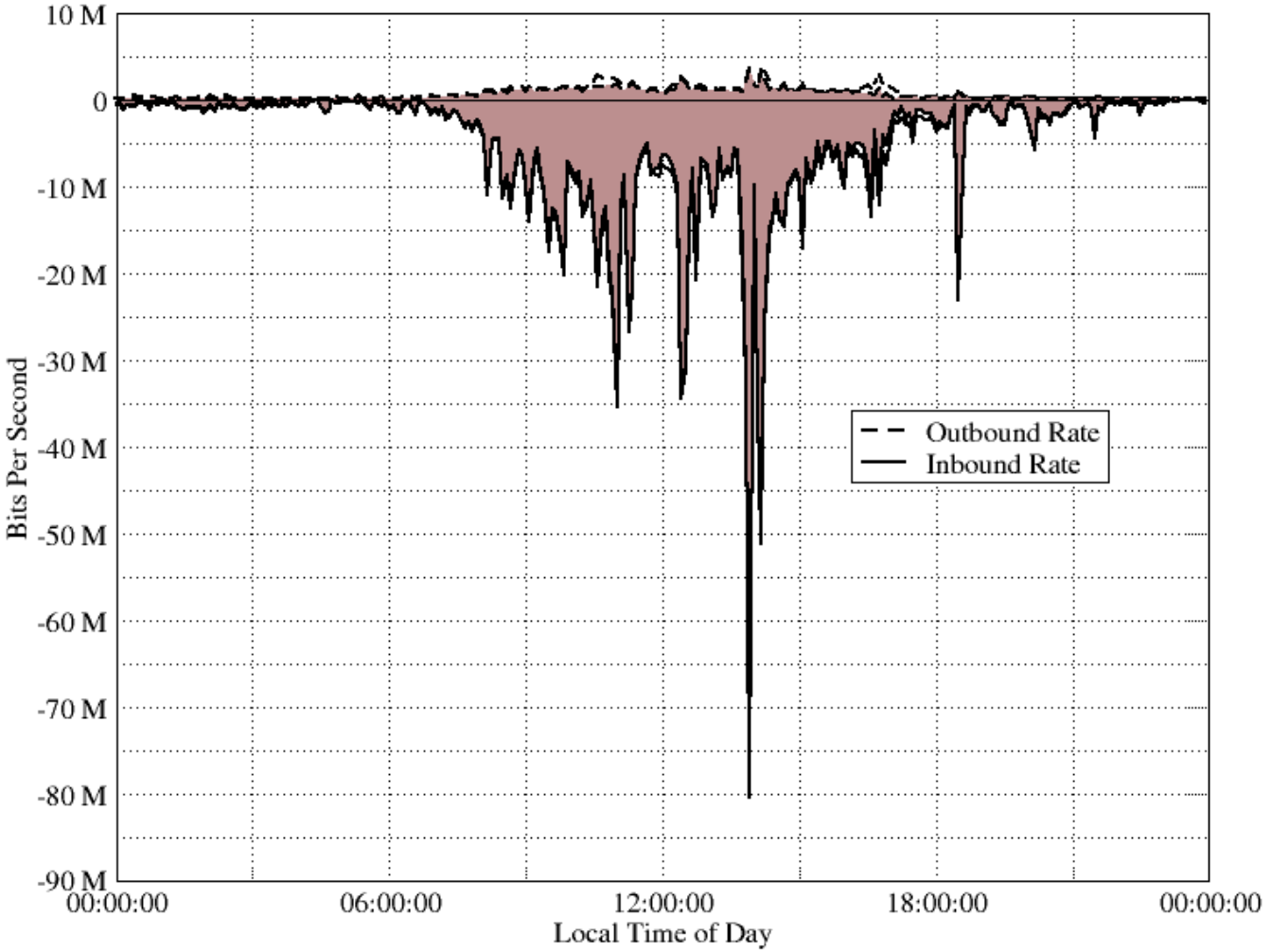
Residential: Domain Popularity



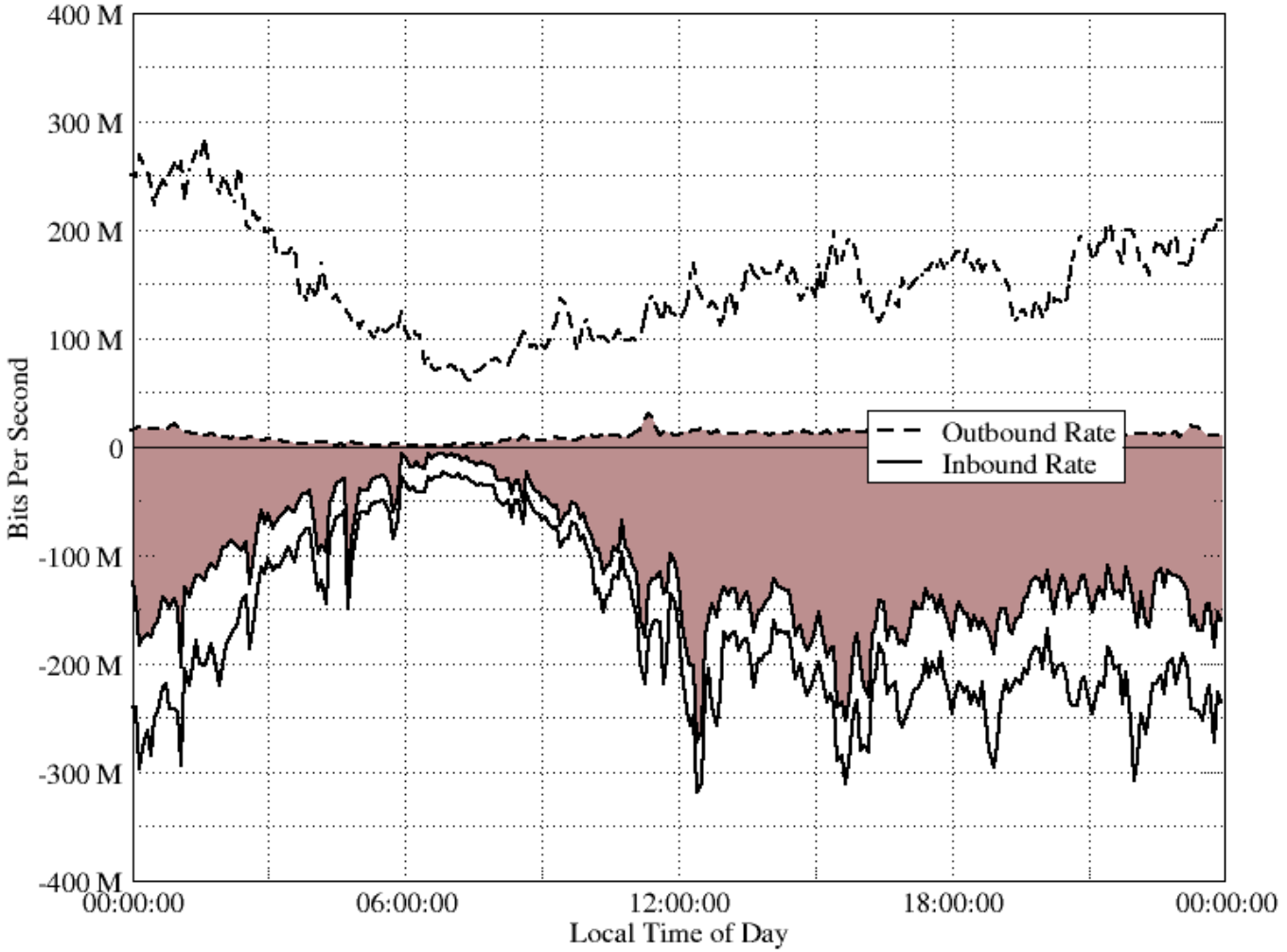
Target Traffic Classification: Port-based method



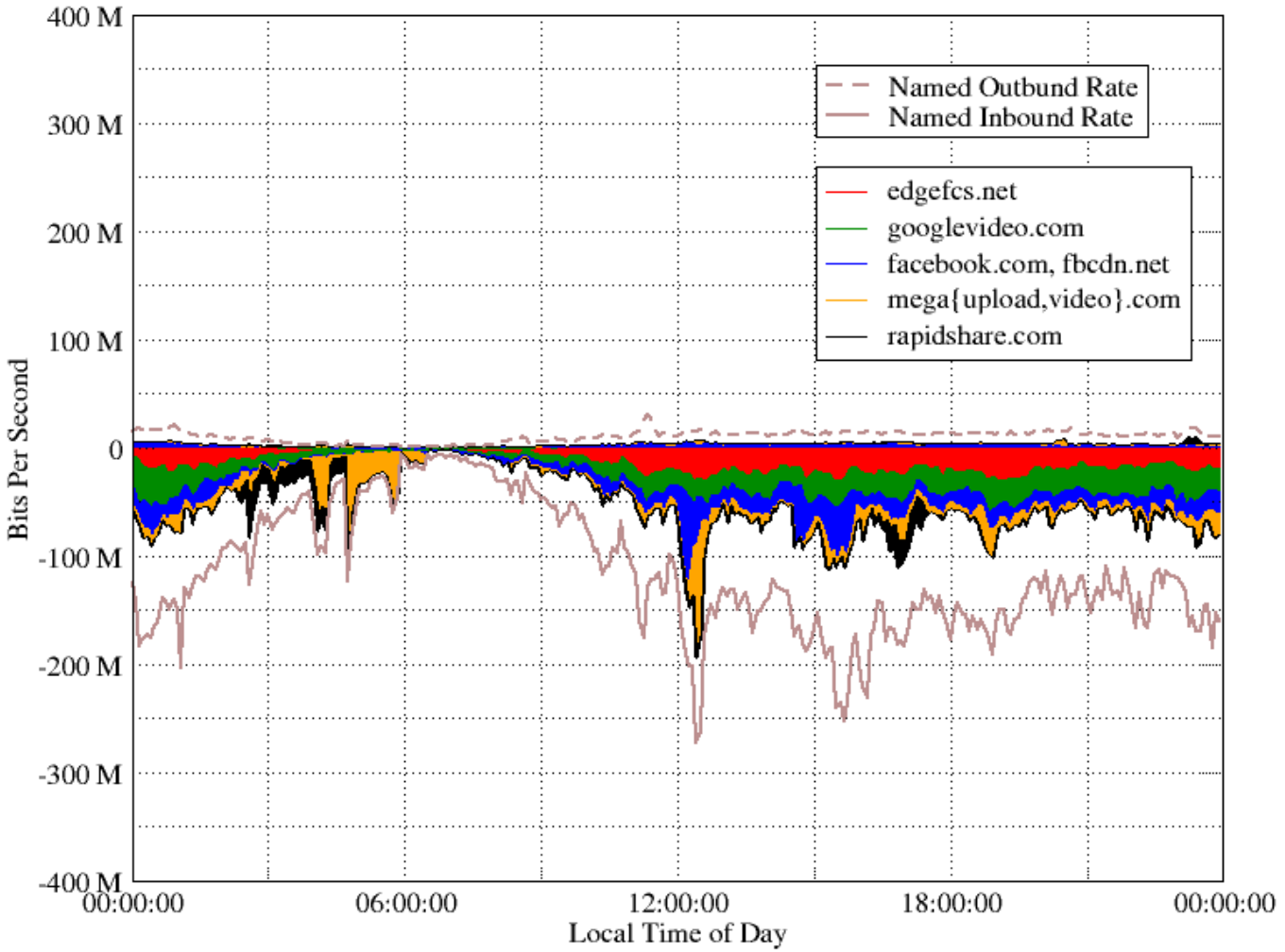
Office Target Traffic Classification: “named” and “unnamed”



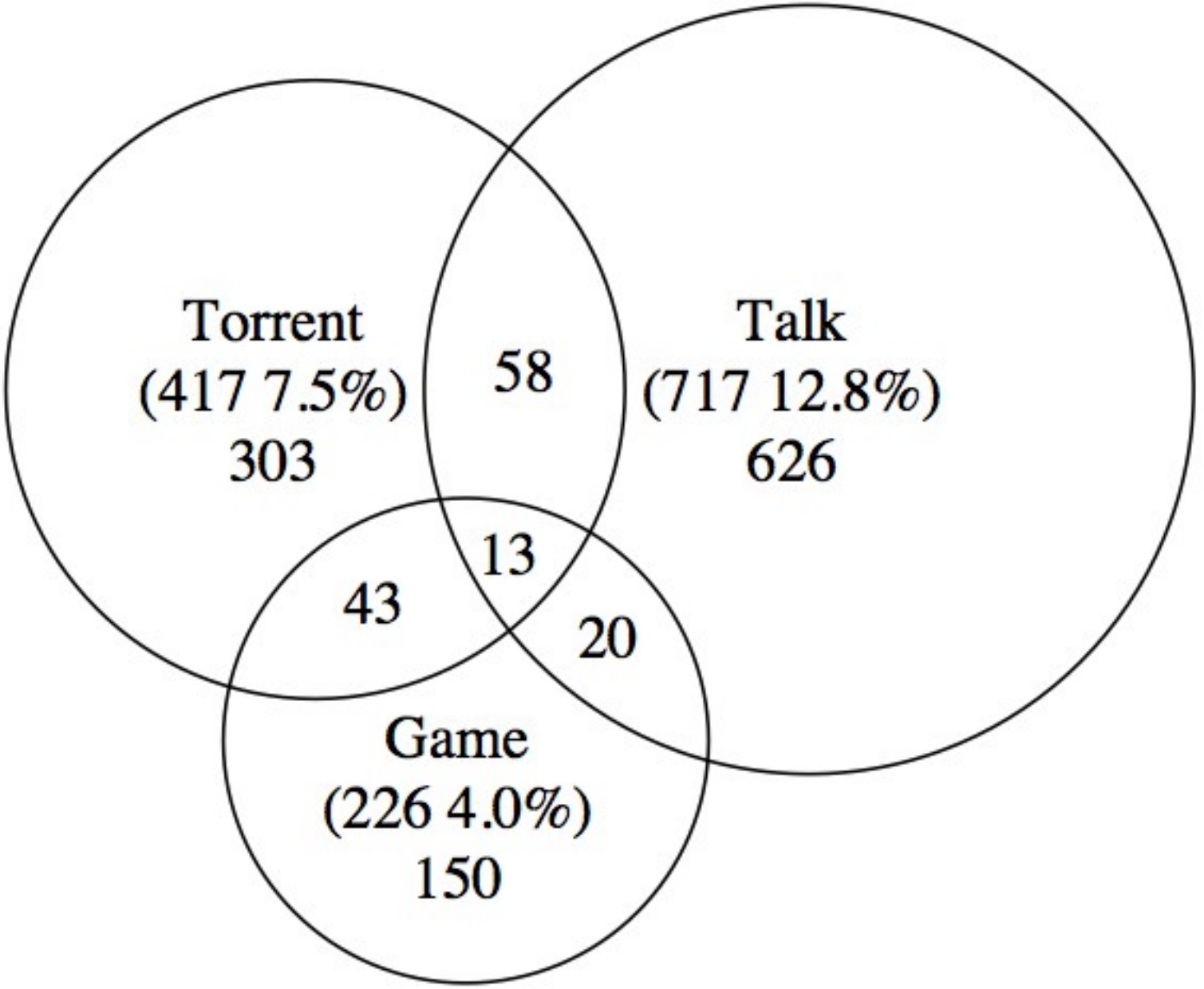
Residential Target Traffic Classification: “named” and “unnamed”



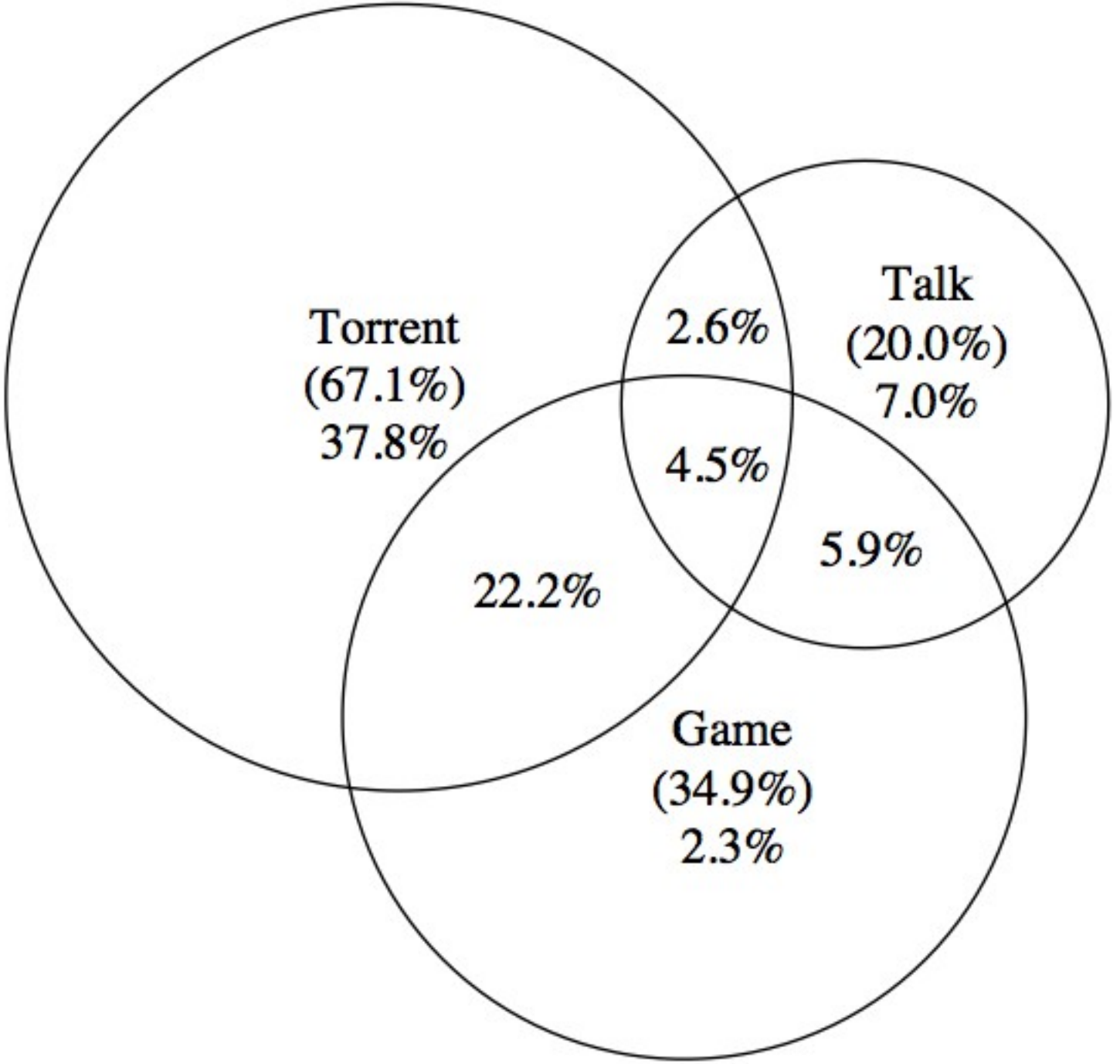
Residential Target Traffic Classification: “named” by popular domains



Residential Hosts Classification by P2P Host Profile (1 day)



“unnamed” Target Traffic by P2P Profile



Results Summary: Traffic Classified (% bytes)

Data Set	Port-known	DNS-named and Port-known	DNS-named	DNS-named and DNS-Profiled
Office Out	93.9%	80.5%	81.8%	91.9%
Office In	96.6%	91.8%	93.2%	95.4%
Residential Out	18.6%	6.2%	6.7%	83.5%
Residential In	76.9%	58.3%	67.9%	88.2%

Discussion & Future Work

- In what circumstances can we **trust** DNS rendezvous information for traffic classification?
- Employ DNS rendezvous-based classification to compare IPv4 and **IPv6** service performance.
- Tap rendezvous methods other than the DNS; e.g., application-specific methods (WWW, P2P); are they **separable** and clear?
- Should we alter rendezvous protocols to better inform classification and packet treatments?

Flexible Traffic and Host Profiling via DNS Rendezvous

FIN

THE UNIVERSITY of WISCONSIN MADISON

David Plonka
&
Paul Barford
{plonka,pb}@cs.wisc.edu