Unconstrained Profiling of Internet Endpoints via Information on the Web

("Googling" the Internet)

Ionut Trestian[1]

Soups Ranjan[2]

Aleksandar Kuzmanovic[1]

Antonio Nucci[2]

[1] Northwestern University

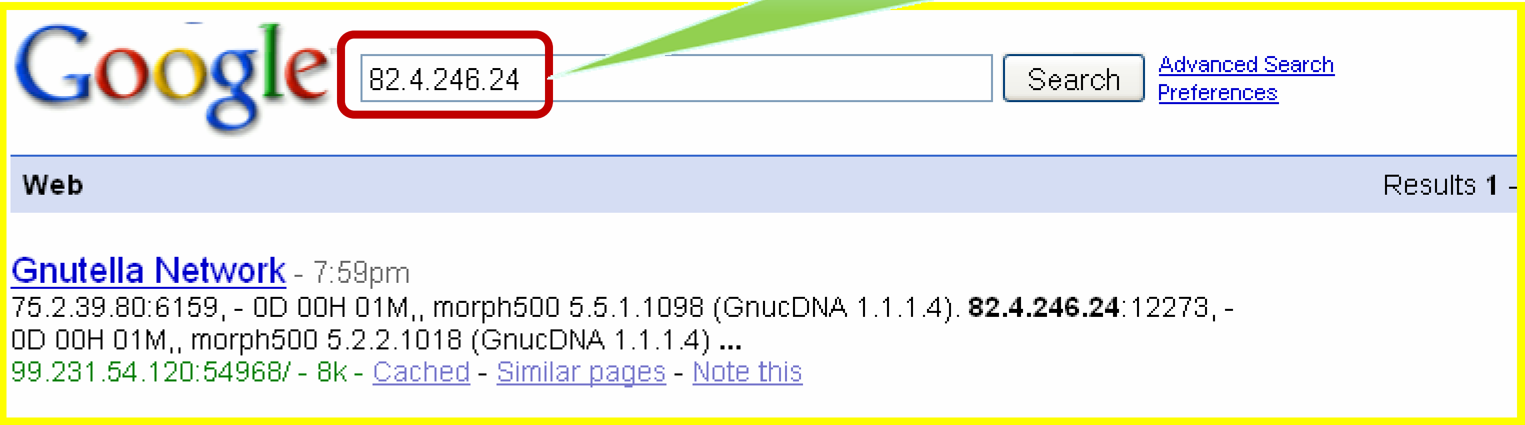
[2] Narus Inc.

http://networks.cs.northwestern.edu **http://www.narus.com**

Introduction

- Can we use Google for networking research?

Huge amount of endpoint information available on the web

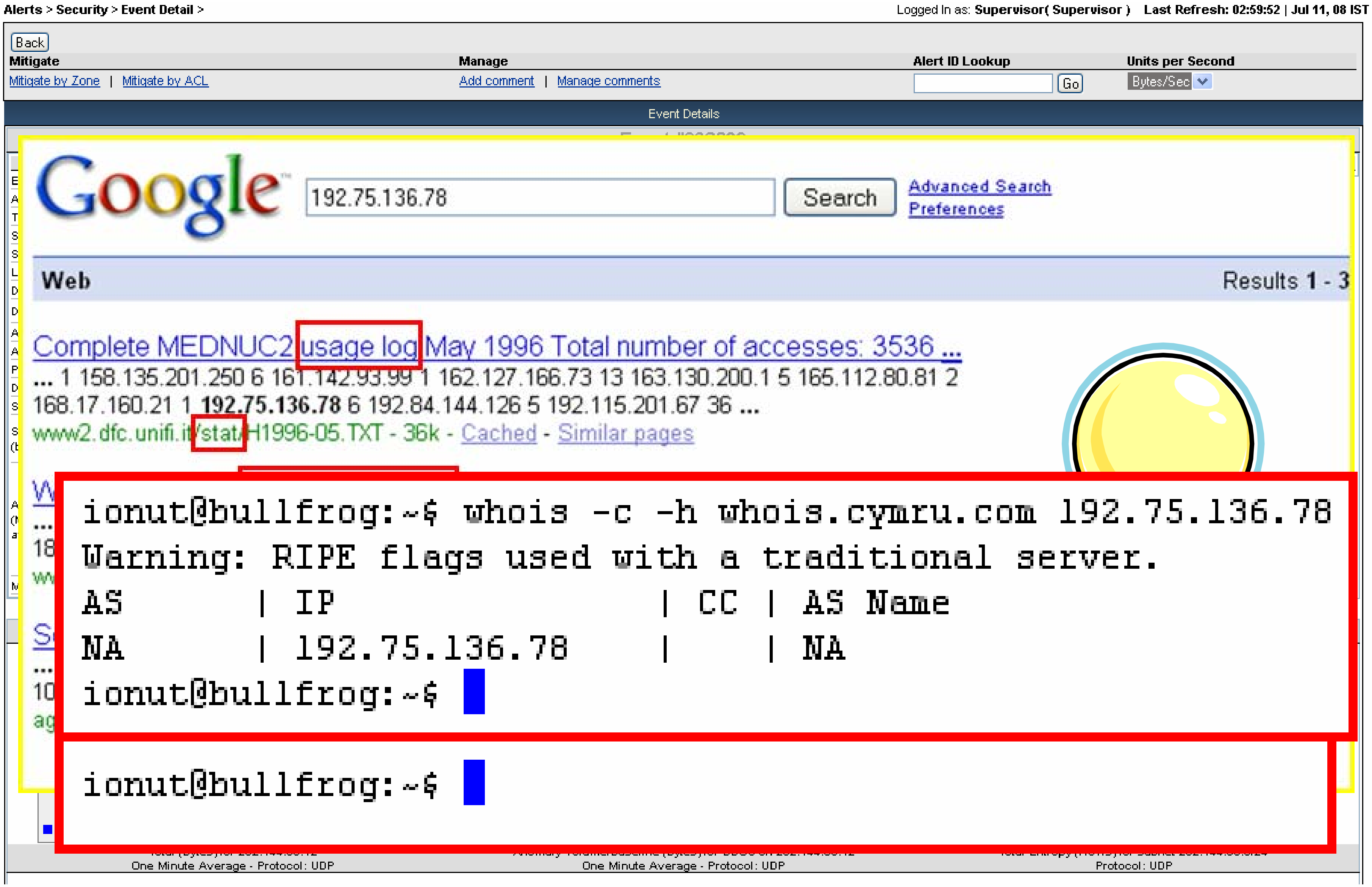


82.4.246.24

Gnutella Network - 7:59pm
75.2.39.80:6159, - 0D 00H 01M,, morph500 5.5.1.1098 (GnucDNA 1.1.1.4). **82.4.246.24**:12273, - 0D 00H 01M,, morph500 5.2.2.1018 (GnucDNA 1.1.1.4) **...**
99.231.54.120:54968/ - 8k - Cached - Similar pages - Note this

Can we systematically exploit search engines to harvest endpoint information available on the Internet?

Application: Googling IP-addresses for Network Forensics


Alerts > Security > Event Detail >
Logged In as: Supervisor(Supervisor) Last Refresh: 02:59:52 | Jul 11, 08 IST
Back
Mitigate
Mitigate by Zone | Mitigate by ACL
Manage
Add comment | Manage comments
Alert ID Lookup
Go
Units per Second
Bytes/Sec
Event Details
Google
192.75.136.78
Search
Advanced Search
Preferences
Web
Results 1 - 3

Complete MEDNUC2 usage log May 1996 Total number of accesses: 3536 ...

... 1 158.135.201.250 6 161.142.93.99 1 162.127.166.73 13 163.130.200.1 5 165.112.80.81 2 168.17.160.21 1 **192.75.136.78** 6 192.84.144.126 5 192.115.201.67 36 ...

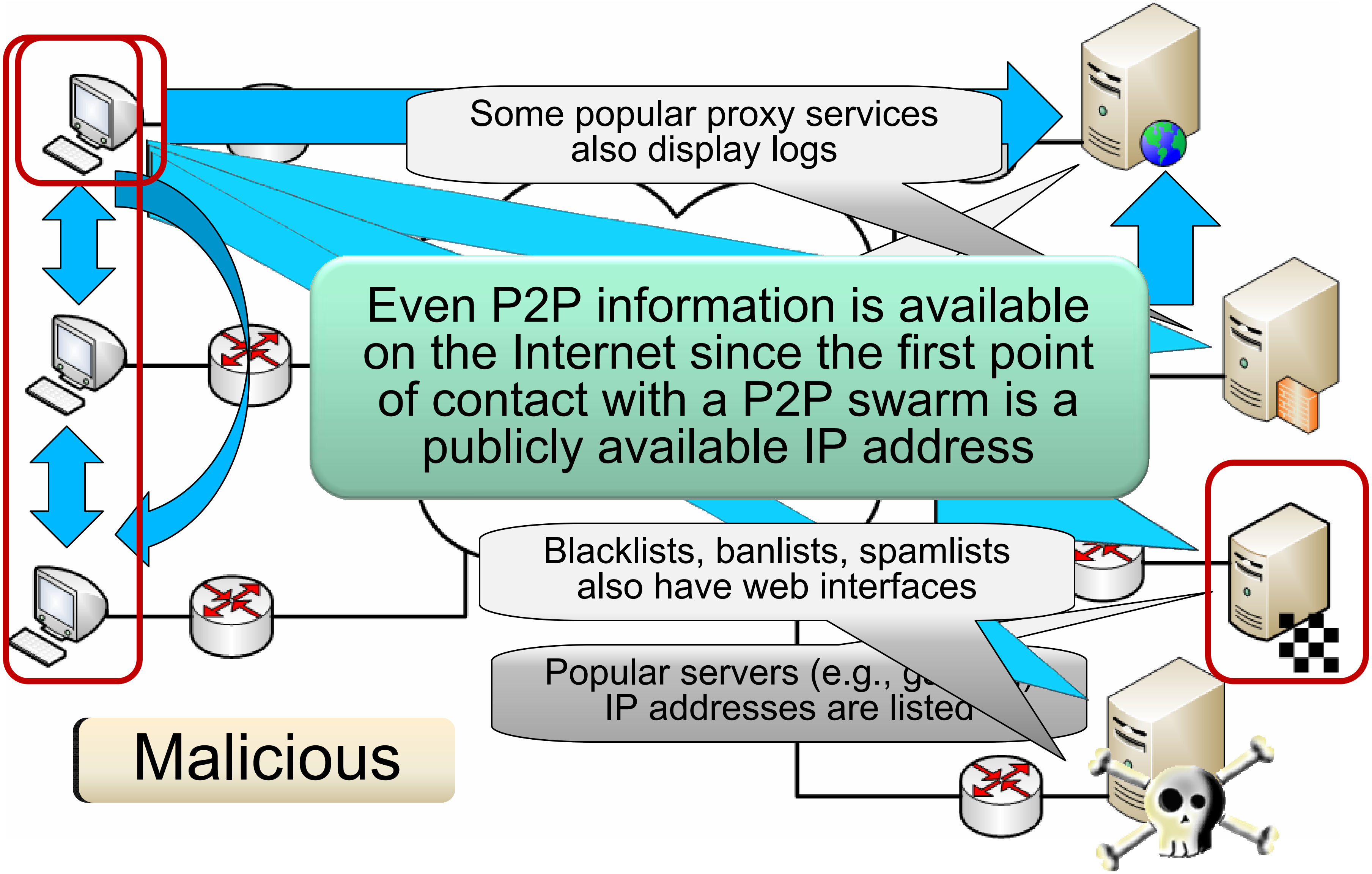
www2.dfc.unifi.it/stat/H1996-05.TXT - 36k - Cached - Similar pages

```
ionut@bullfrog:~$ whois -c -h whois.cymru.com 192.75.136.78
Warning: RIPE flags used with a traditional server.
AS      | IP               | CC | AS Name
NA      | 192.75.136.78    |    | NA
ionut@bullfrog:~$ █
```

```
ionut@bullfrog:~$ █
```

One Minute Average - Protocol: UDP
One Minute Average - Protocol: UDP
Protocol: UDP

Where Does the Information Come From?



Some popular proxy services also display logs

Even P2P information is available on the Internet since the first point of contact with a P2P swarm is a publicly available IP address

Blacklists, banlists, spamlists also have web interfaces

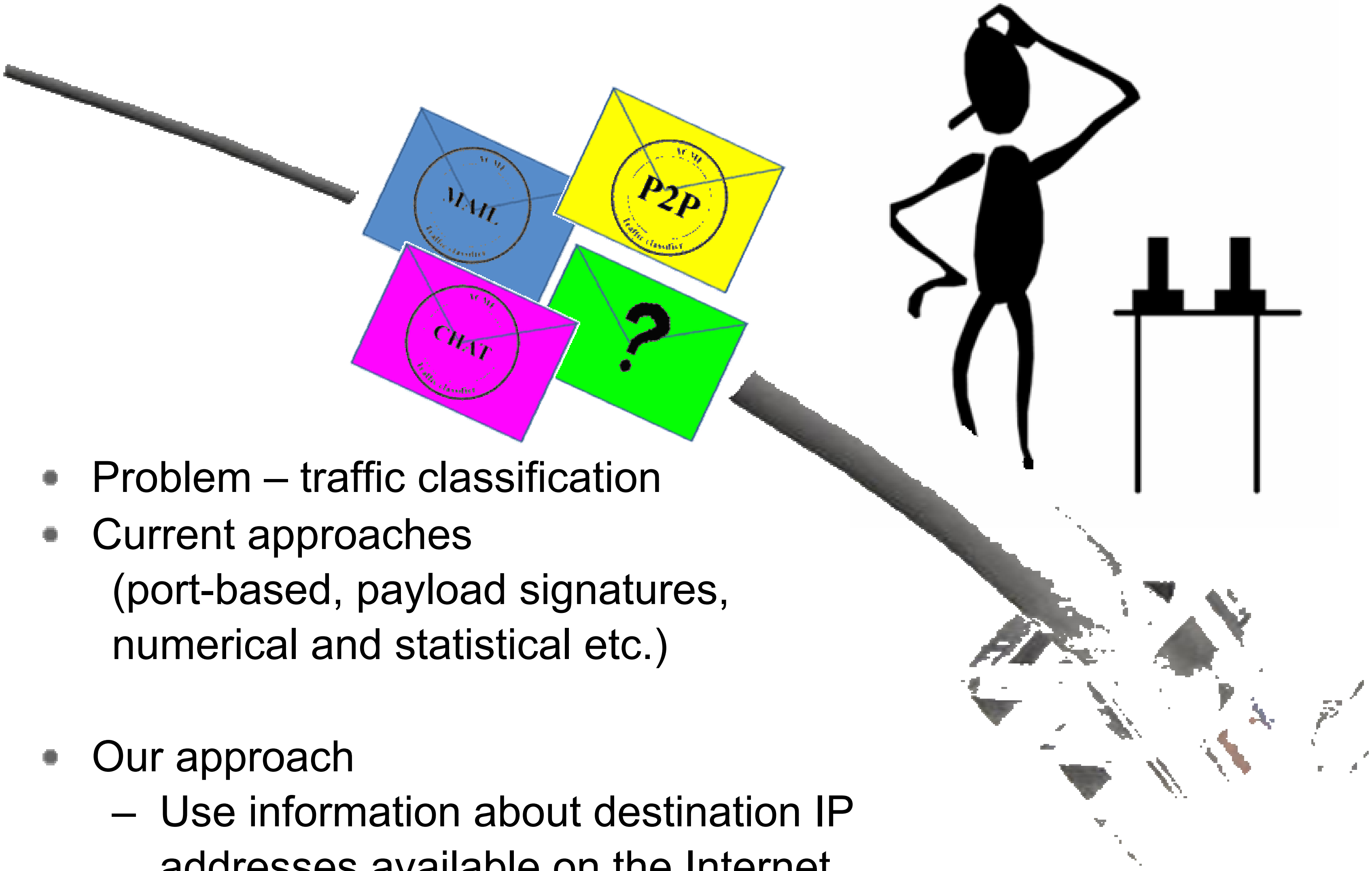
Popular servers (e.g., g) IP addresses are listed

Malicious

Detecting Application Usage Trends

Can we infer what applications people are using across the world without having access to network traces?

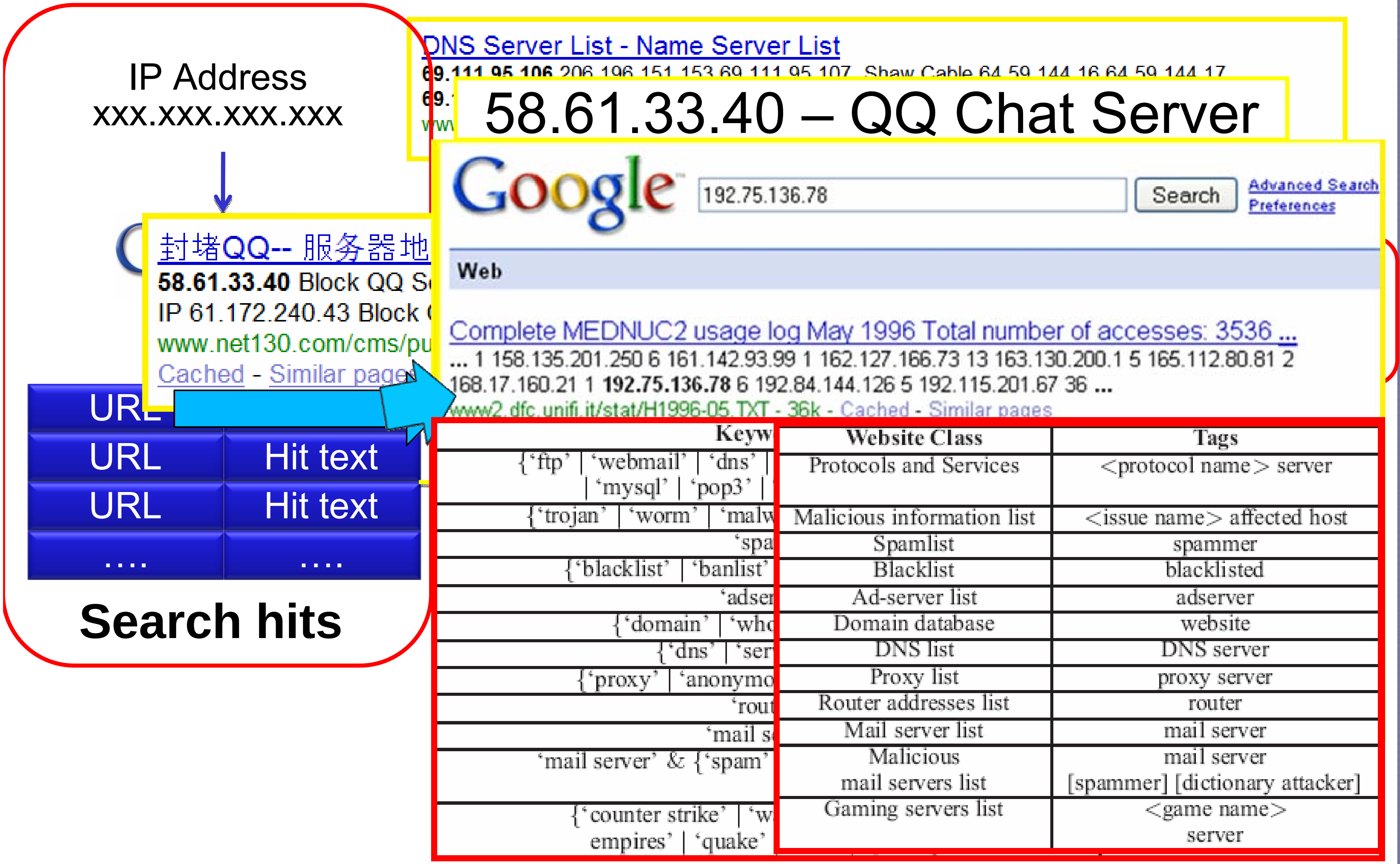
Traffic Classification



- Problem – traffic classification
- Current approaches
  (port-based, payload signatures, numerical and statistical etc.)

- Our approach
  - Use information about destination IP addresses available on the Internet

Methodology – Web Classifier and IP Tagging



封堵QQ-- 服务器地
58.61.33.40 Block QQ S
IP 61.172.240.43 Block (
www.net130.com/cms/pu
Cached - Similar pages

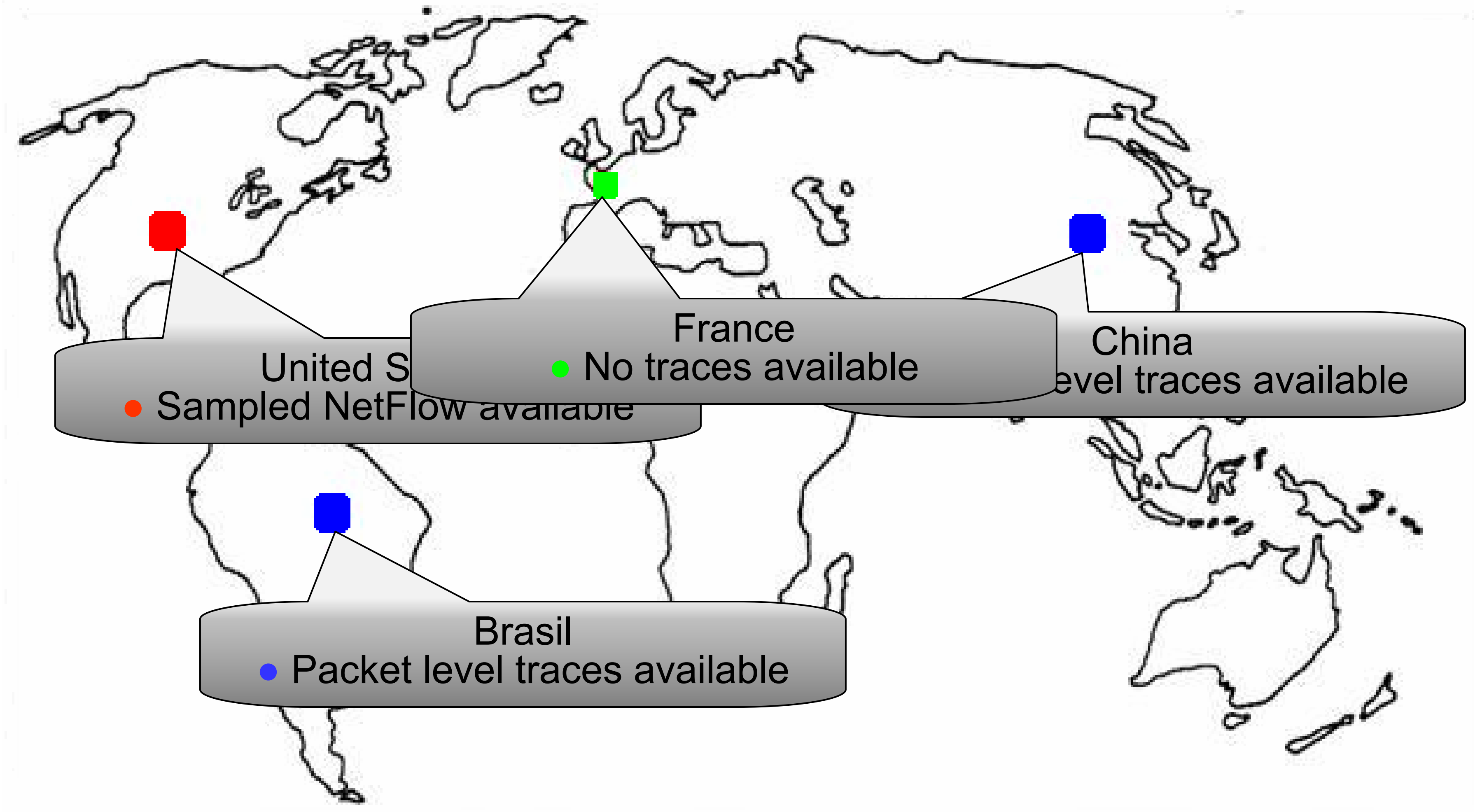
DNS Server List - Name Server List
69.111.95.106 206.196.151.153 69.111.95.107 Shaw Cable 64.59.144.16 64.59.144.17
69.
ww

58.61.33.40 – QQ Chat Server

Complete MEDNUC2 usage log May 1996 Total number of accesses: 3536 ...
... 1 158.135.201.250 6 161.142.93.99 1 162.127.166.73 13 163.130.200.1 5 165.112.80.81 2 168.17.160.21 1 **192.75.136.78** 6 192.84.144.126 5 192.115.201.67 36 ...
www2.dfc.unifi.it/stat/H1996-05.TXT - 36k - Cached - Similar pages

Keyw	Website Class	Tags
{'ftp' | 'webmail' | 'dns' | | 'mysql' | 'pop3' | '	Protocols and Services	<protocol name> server
{'trojan' | 'worm' | 'malw	Malicious information list	<issue name> affected host
'spa	Spamlist	spammer
{'blacklist' | 'banlist'	Blacklist	blacklisted
'adser	Ad-server list	adserver
{'domain' | 'whc	Domain database	website
{'dns' | 'ser	DNS list	DNS server
{'proxy' | 'anonymo	Proxy list	proxy server
'rout	Router addresses list	router
'mail s	Mail server list	mail server
'mail server' & {'spam'	Malicious mail servers list	mail server [spammer] [dictionary attacker]
{'counter strike' | 'w empires' | 'quake'	Gaming servers list	<game name> server

Evaluation – Ground Truth from Traces



United S

- Sampled NetFlow available

France

- No traces available

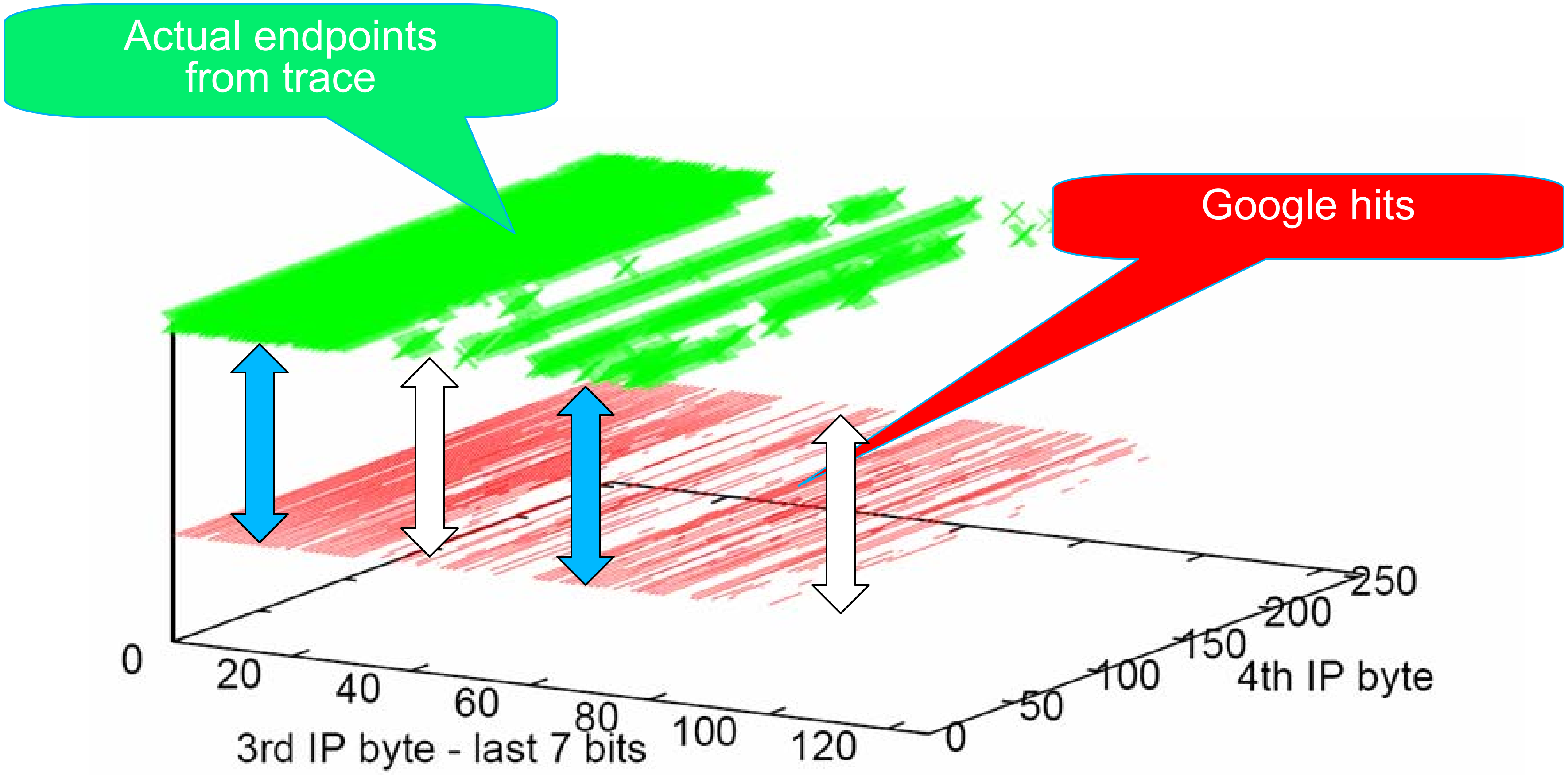
China

evel traces available

Brasil

- Packet level traces available

Inferring Active IP Ranges in Target Networks

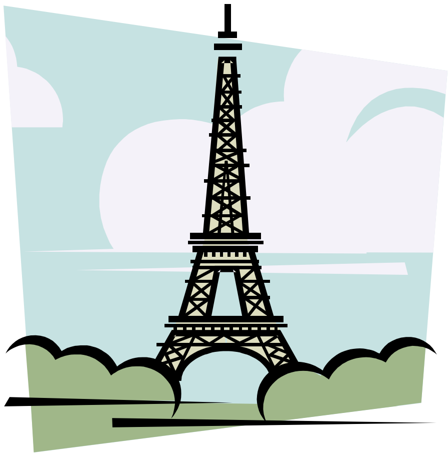


Actual endpoints from trace

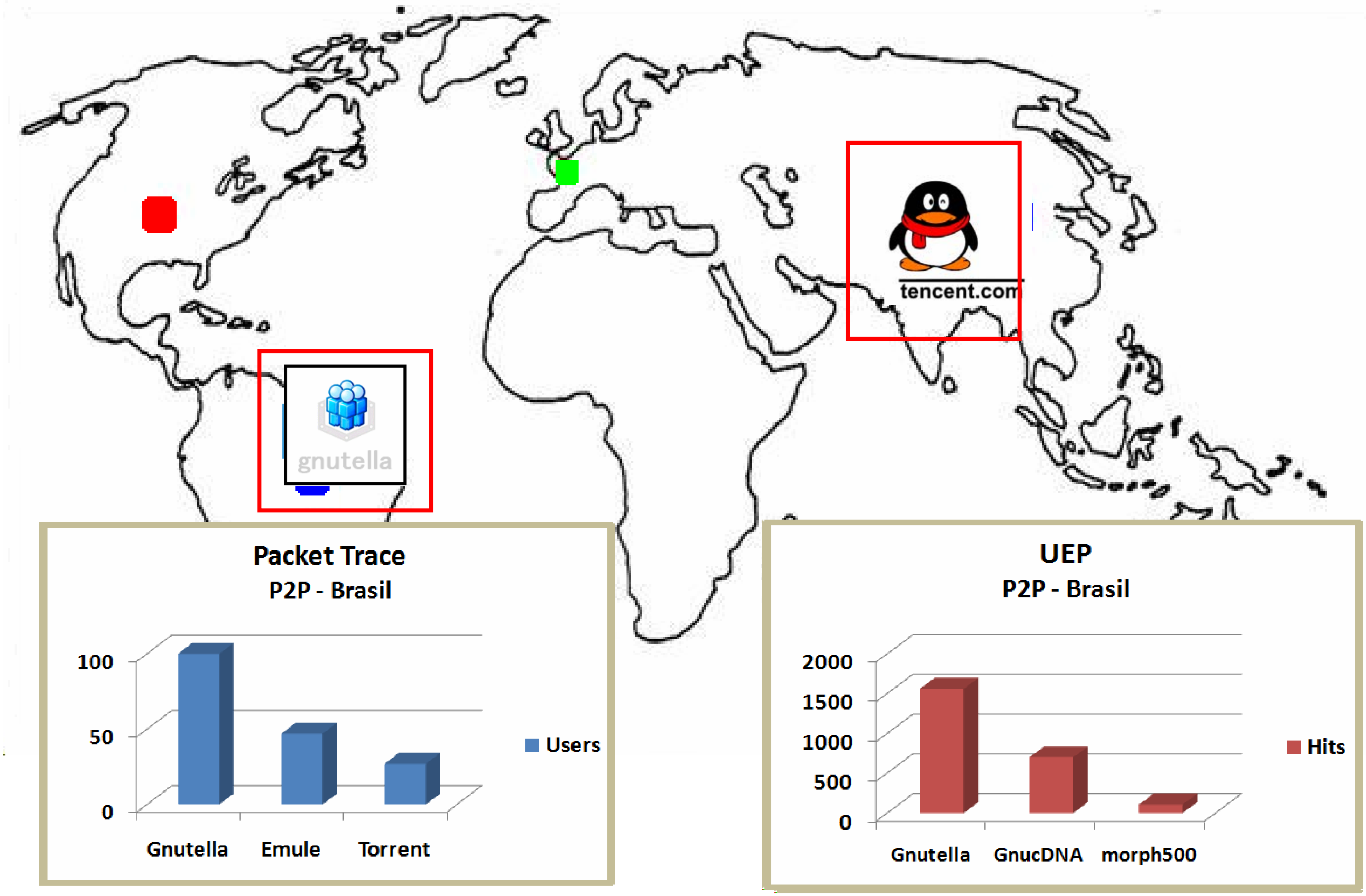
Google hits

XXX.163.0.0/17 network range
Overlap is around 77%

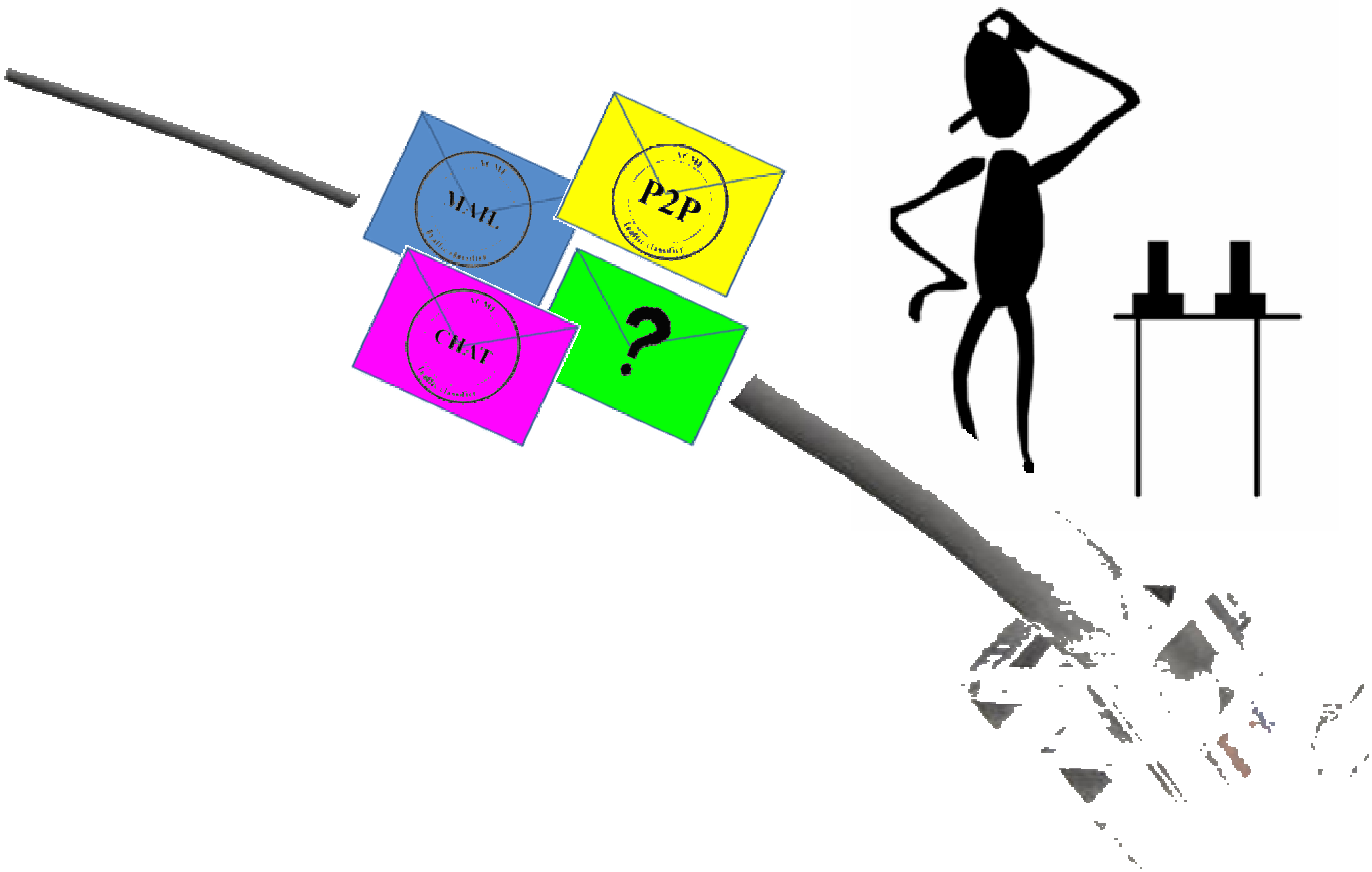
Application Usage Trends



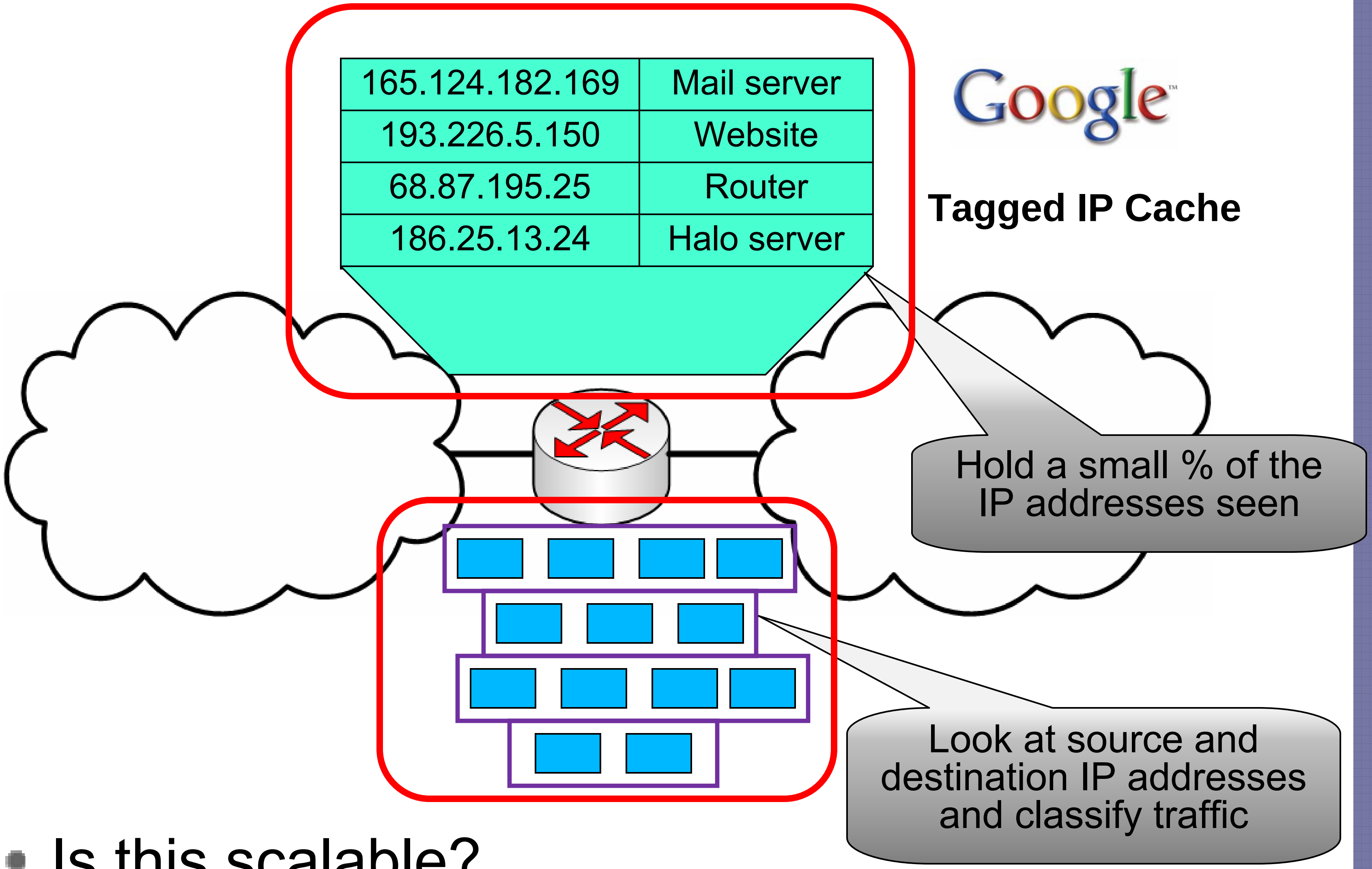
Correlation Between Network Traces and UEP



Traffic Classification



Traffic Classification



165.124.182.169	Mail server
193.226.5.150	Website
68.87.195.25	Router
186.25.13.24	Halo server

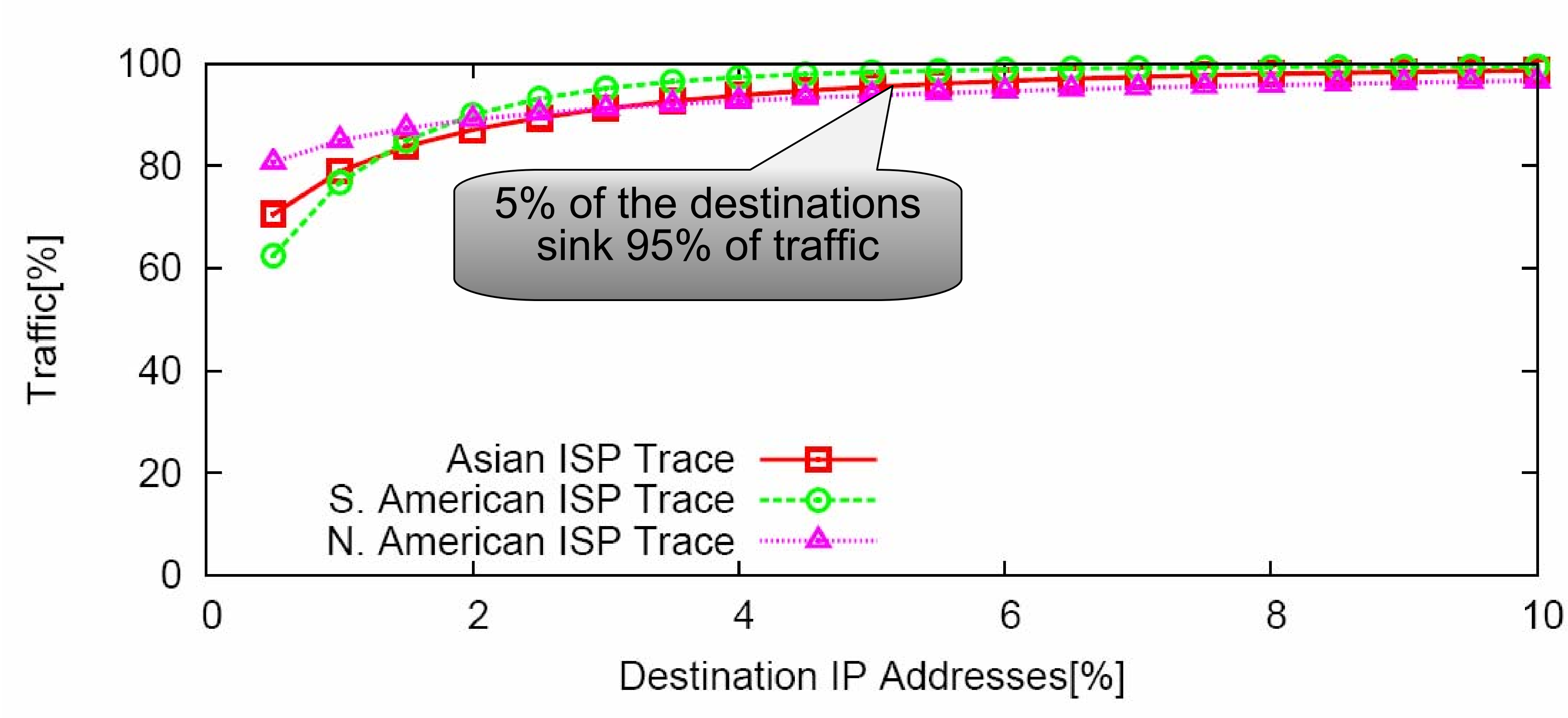
Tagged IP Cache

Hold a small % of the IP addresses seen

Look at source and destination IP addresses and classify traffic

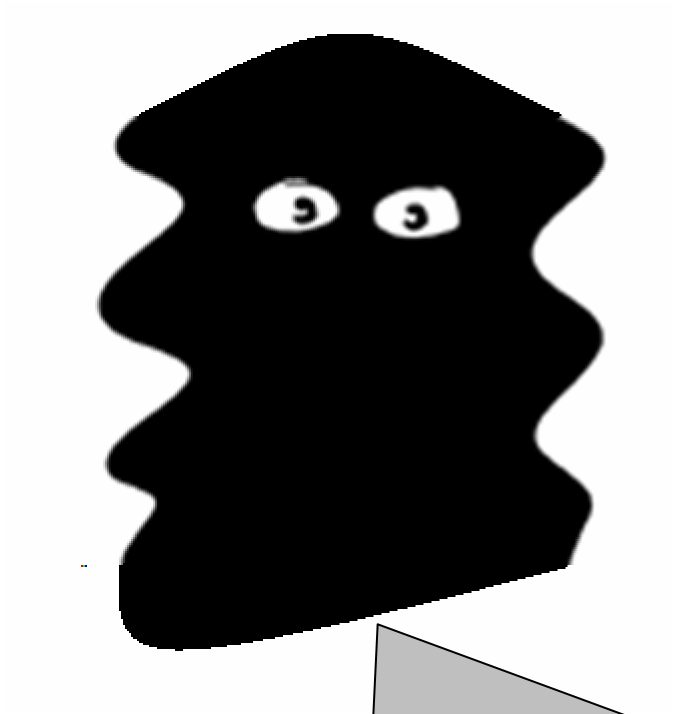
- Is this scalable?

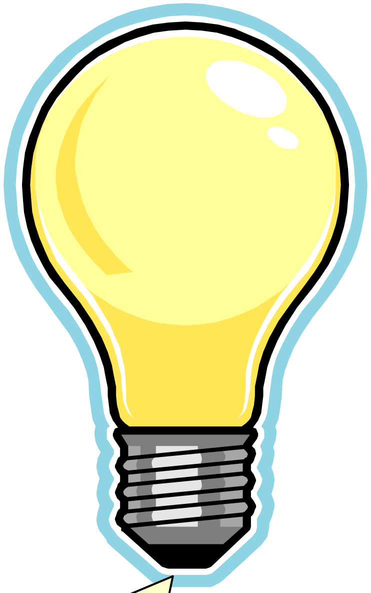
Traffic Classification



5% of the destinations sink 95% of traffic

BLINC vs. UEP

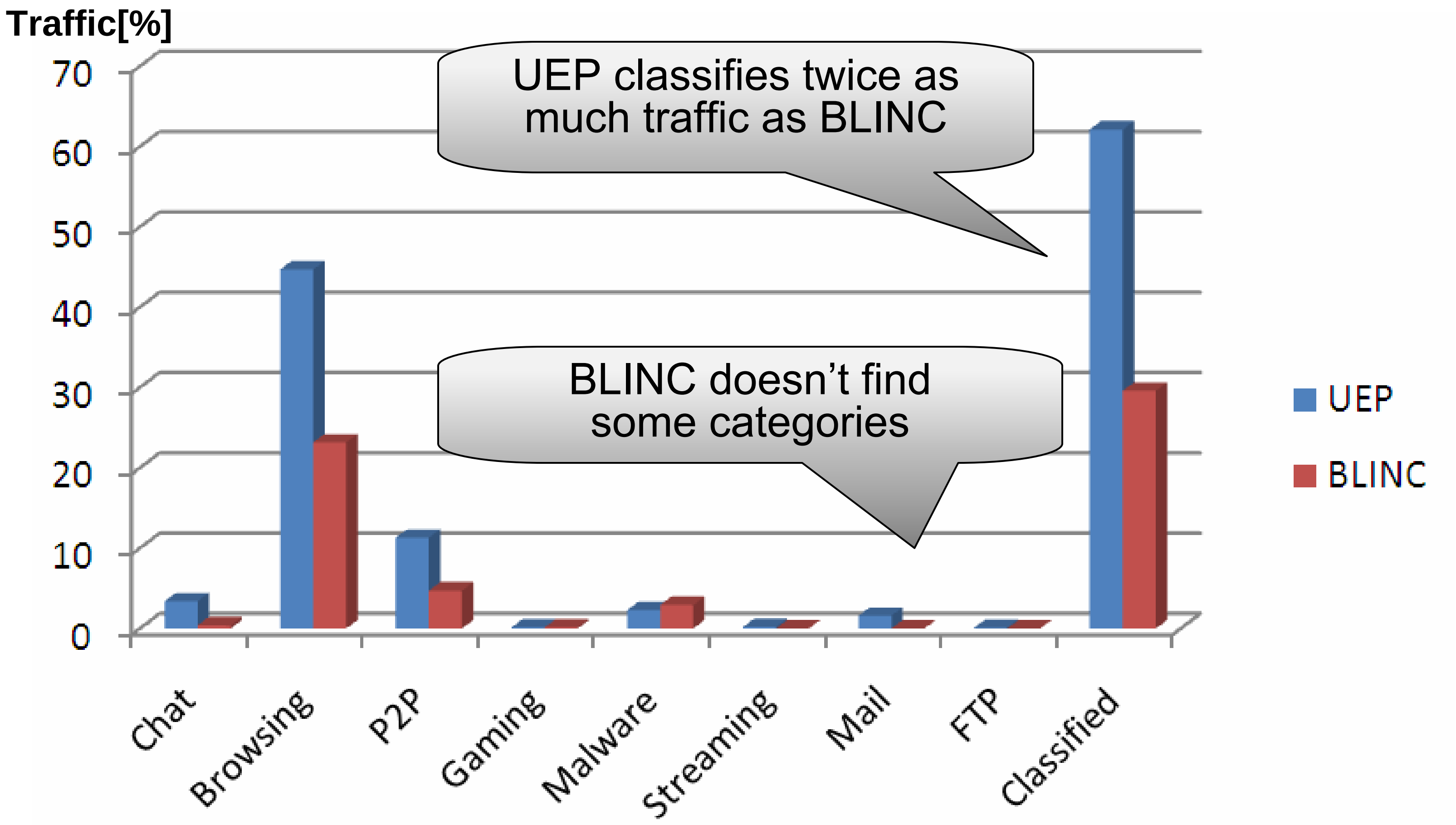




BLINC (SIGCOMM 2005, NANOG 35)

- Works “in the dark” (doesn’t examine payload)
- Uses “graphlets” to identify traffic patterns
- Uses thresholds to further classify traffic

BLINC vs. UEP (cont.)

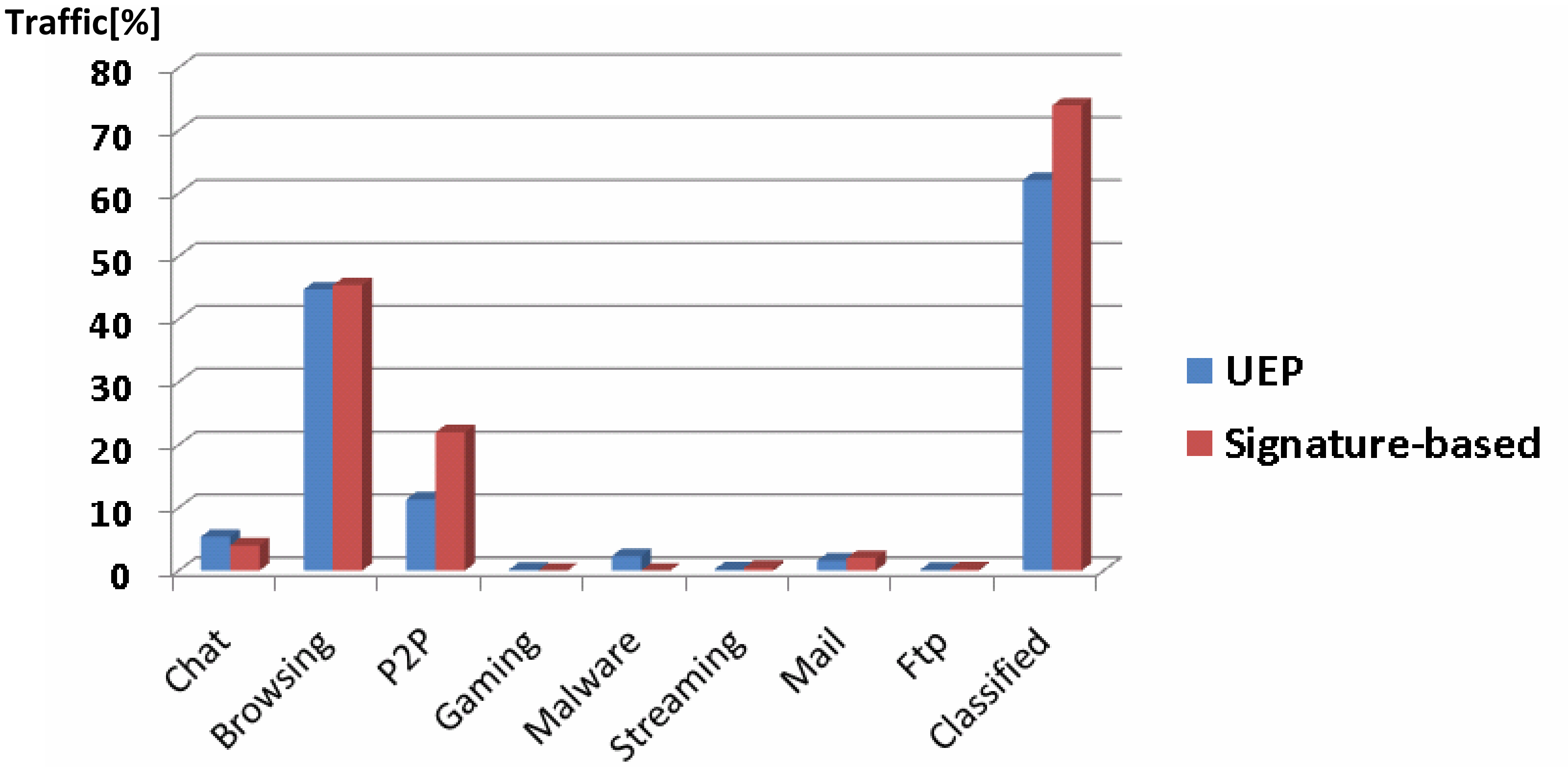


UEP classifies twice as much traffic as BLINC

BLINC doesn't find some categories

UEP also provides better semantics
Classes can be further divided into different services

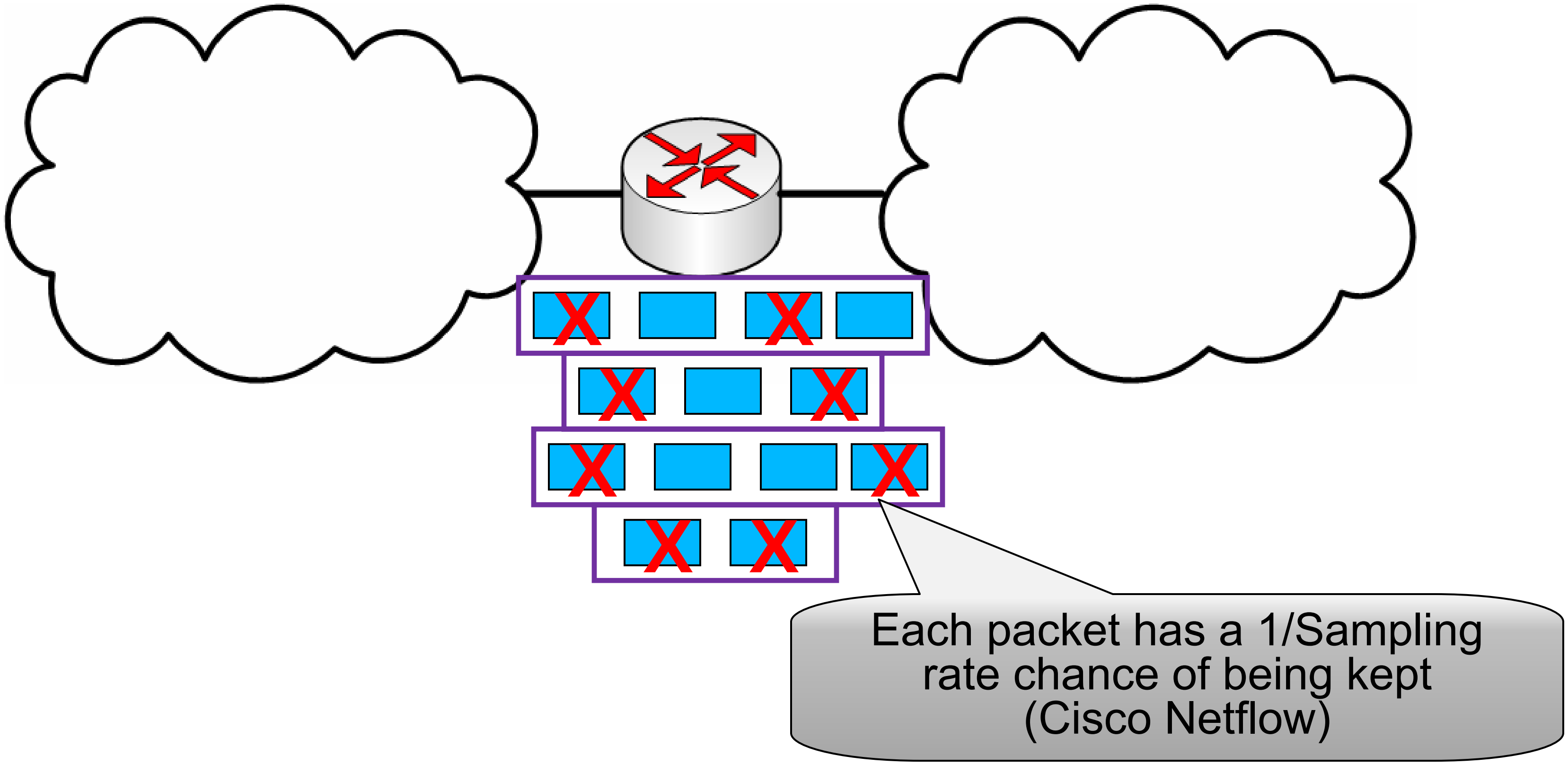
UEP vs. Signature-based



- Unconstrained Endpoint Profiling based Traffic Classification
  - Based on ip-addresses
- L7 signature based
- UEP has comparable performance

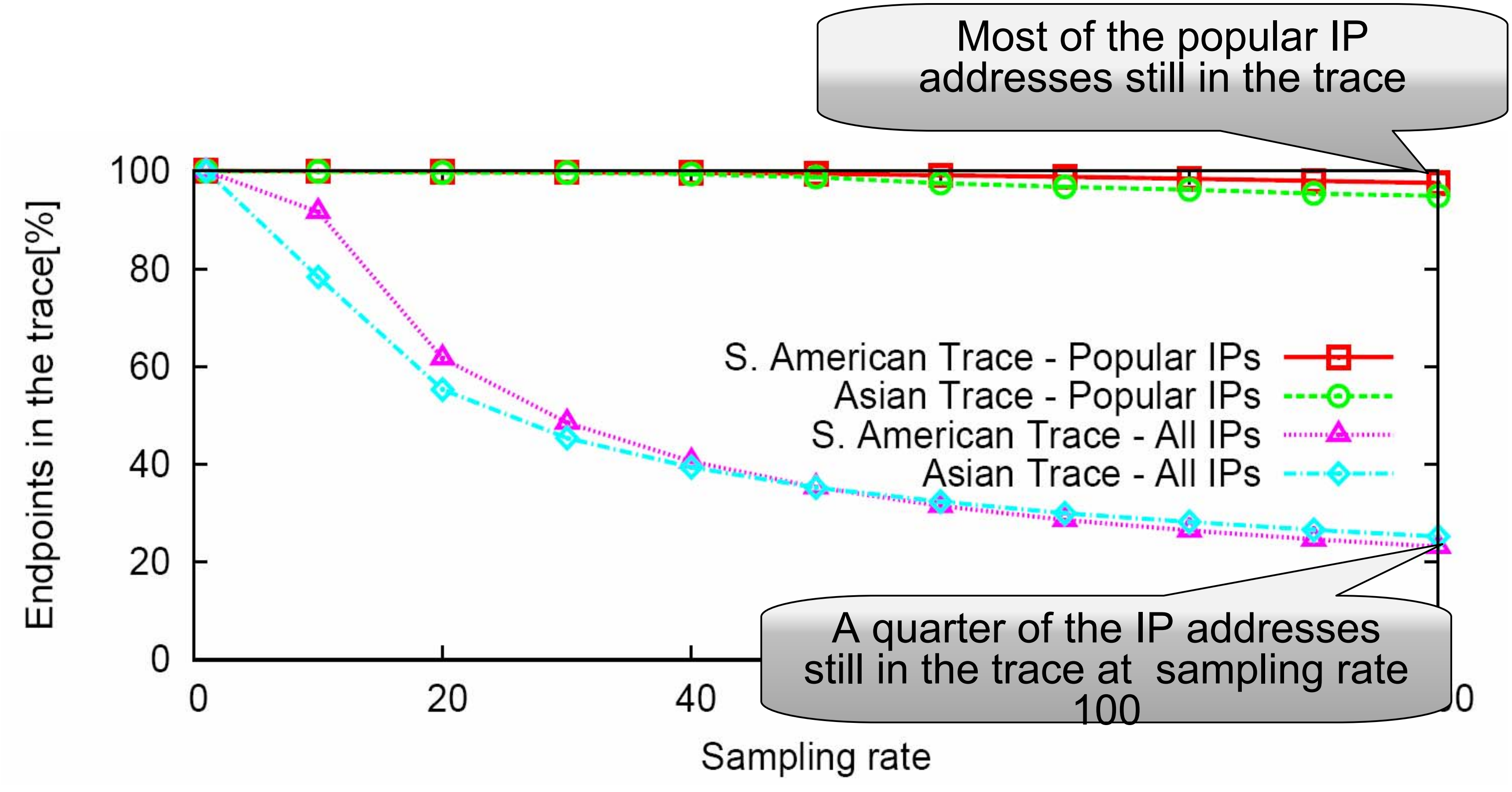
Working with Sampled Traffic

- Sampled data is considered to be poorer in information
- However ISPs consider scalable to gather only sampled data



Each packet has a 1/Sampling rate chance of being kept (Cisco Netflow)

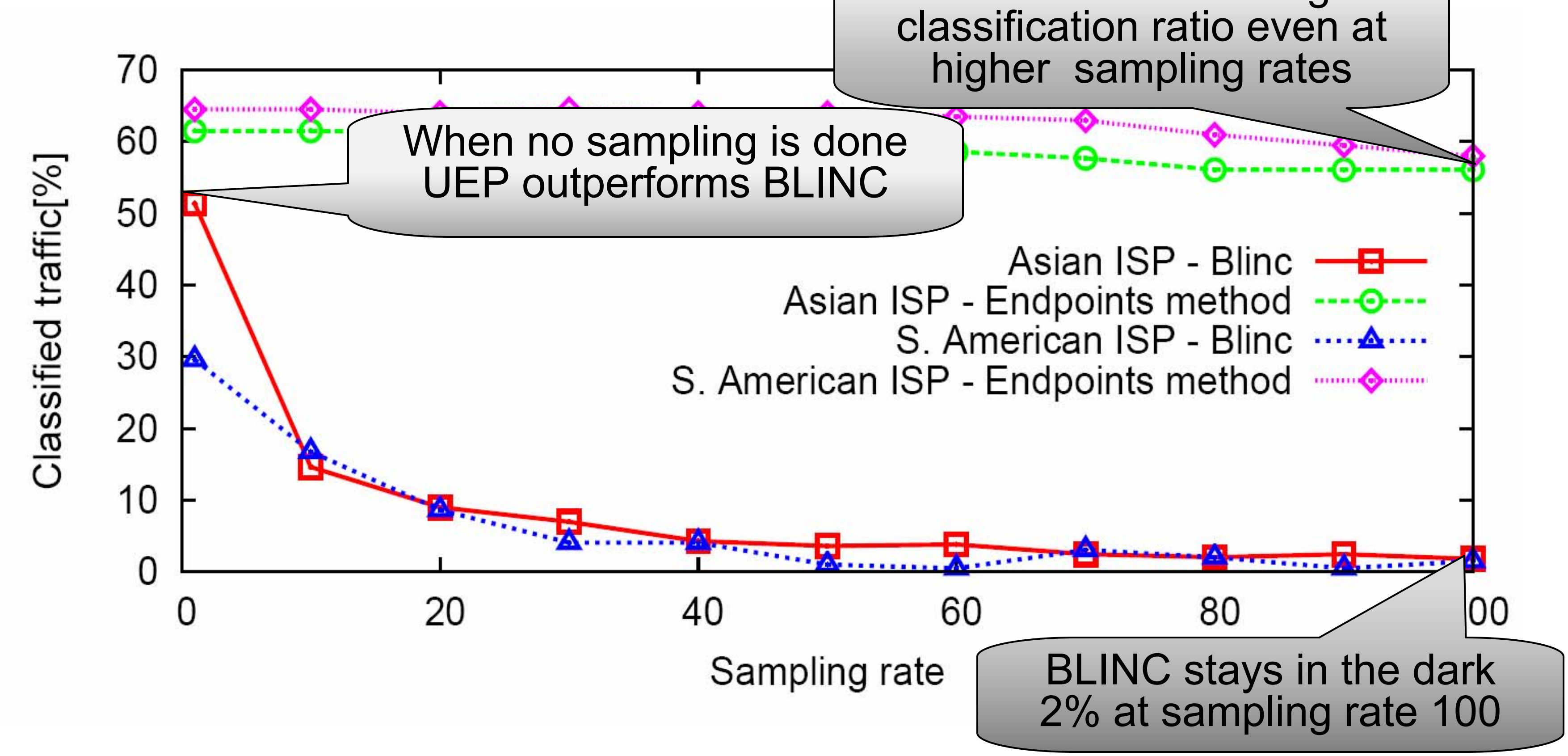
Working with Sampled Traffic



Most of the popular IP addresses still in the trace

A quarter of the IP addresses still in the trace at sampling rate 100

Working with Sampled Traffic



UEP maintains a large classification ratio even at higher sampling rates

When no sampling is done UEP outperforms BLINC

BLINC stays in the dark 2% at sampling rate 100

UEP retains high classification capabilities with sampled traffic

Endpoint Clustering

- Performed clustering of endpoints in order to cluster out common behavior

- Please see the paper for detailed results

Real strength:

We managed to achieve similar results both by using the trace and only by using UEP

Conclusions

- Key contribution:
  - Shift research focus from mining operational network traces to harnessing information that is already available on the web
- Our approach can:
  - Predict application and protocol usage trends in arbitrary networks
  - Dramatically outperform classification tools
  - Retain high classification capabilities when dealing with sampled data

Thanks

Ionut Trestian, Soups Ranjan,
Aleksandar Kuzmanovic, Antonio Nucci

http://ccr.sigcomm.org/online/?q=node/396