

Probabilistic Identification for Hard to classify Protocol

Elie Bursztein

PhD Student LSV ENS-CACHAN CNRS INRIA DGA

Introduction

- What is a hard to classify protocol ?

Payload

- Identification based on payload analysis

Classifier

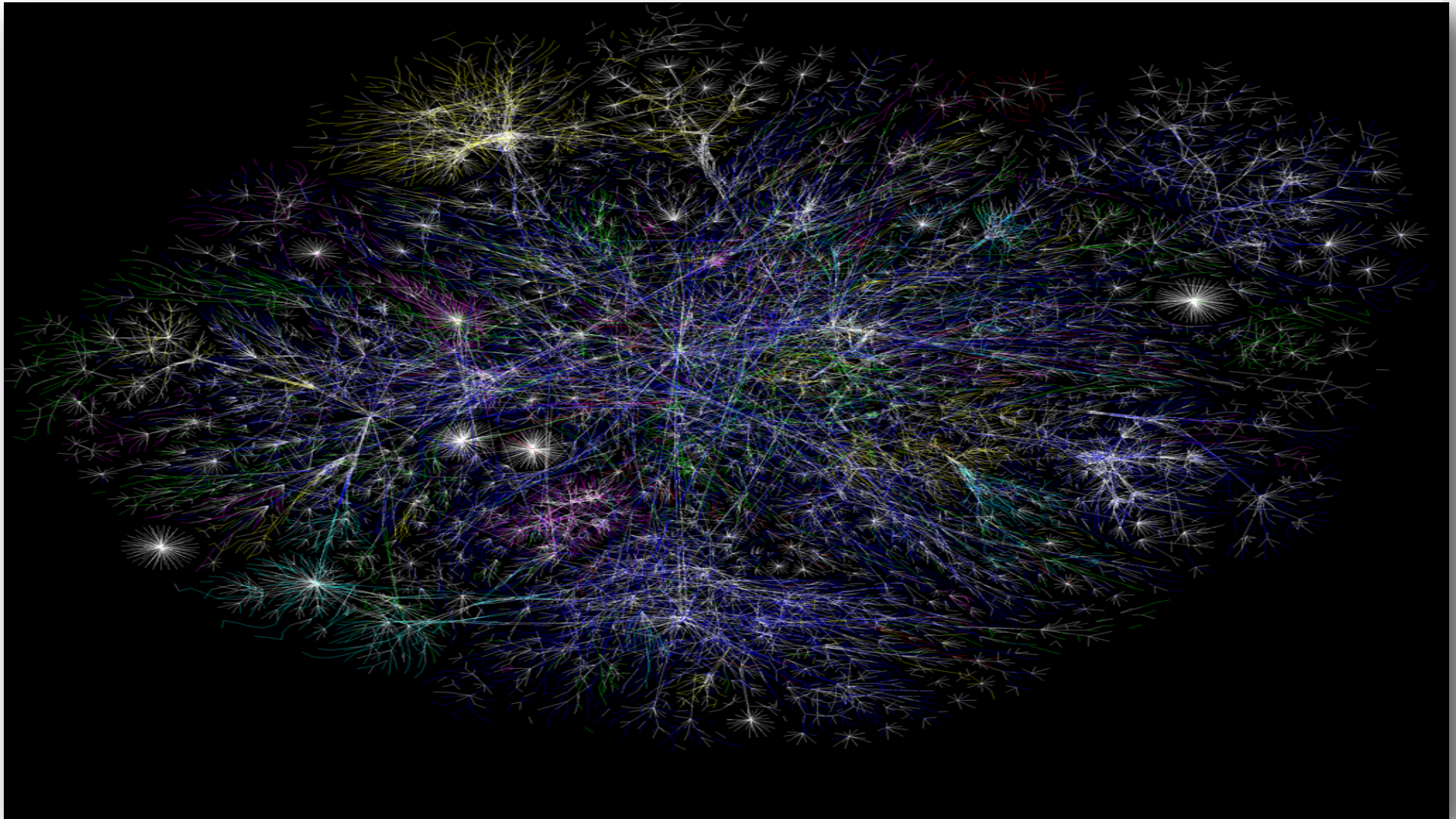
- Identification based on discriminator

Combining Techniques

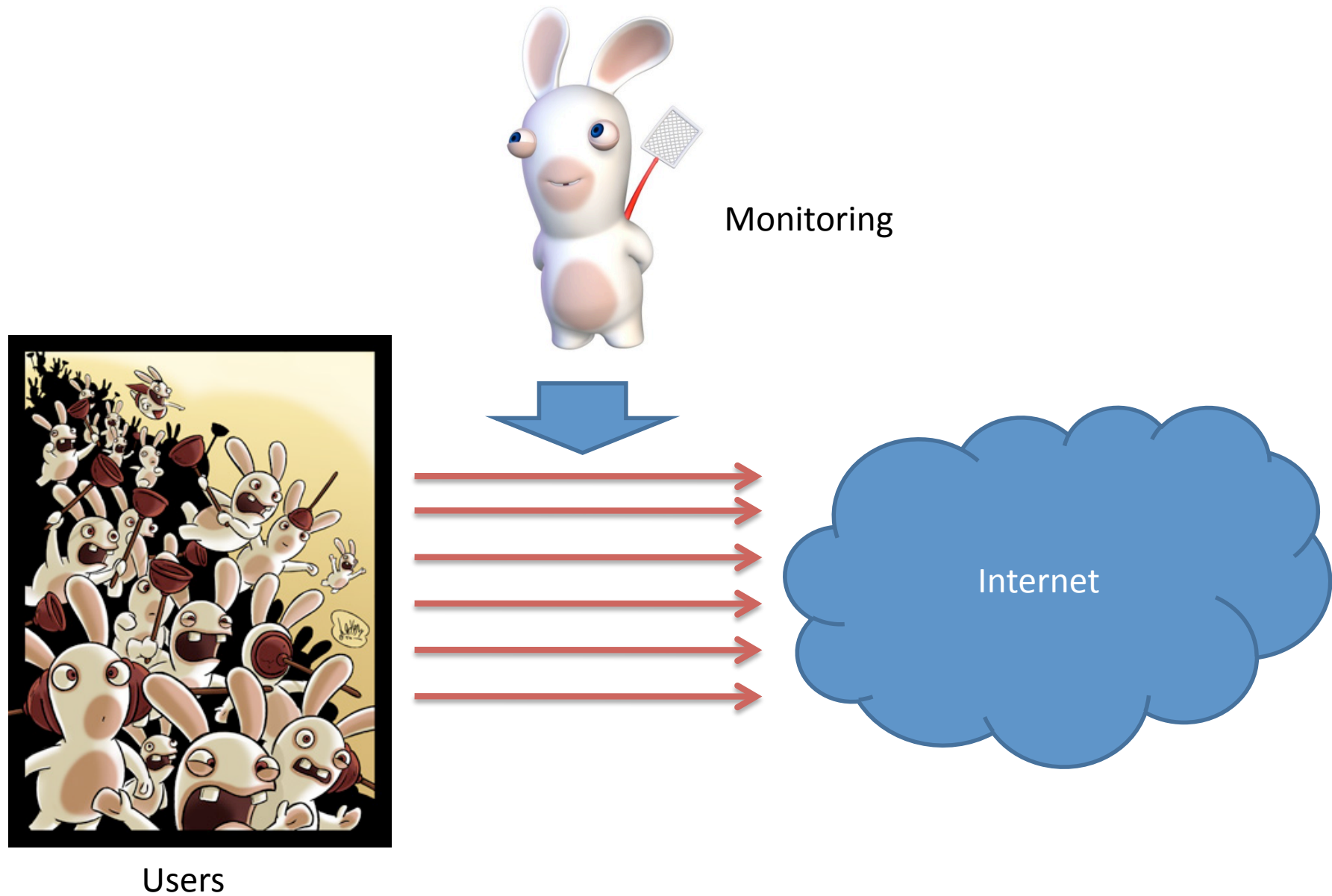
- Combining techniques for a better classification

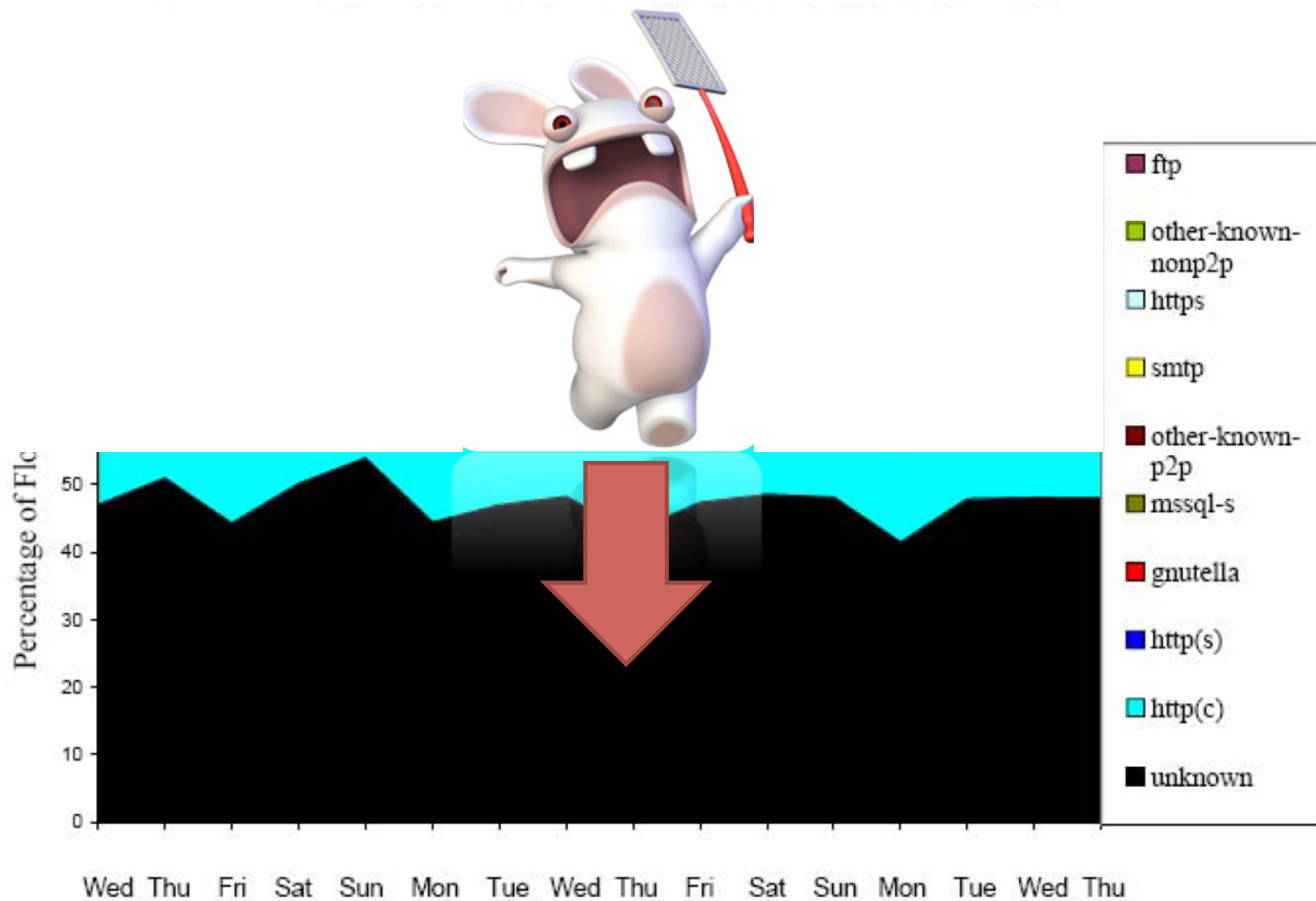
Conclusion

- What is next ?



Opte project



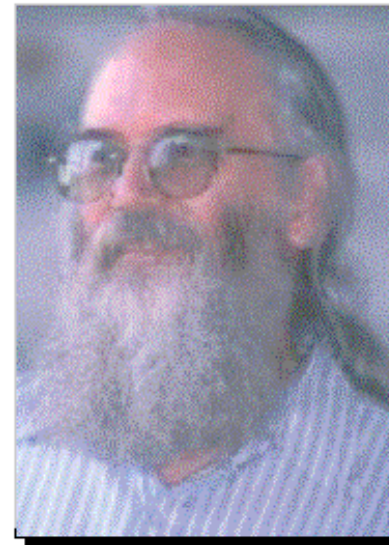


University of Calgary

Internet Assigned Numbers Authority (IANA)

- ❑ Founded in 1970
- ❑ Manage DNS root
- ❑ Coordinate IP and AS
- ❑ Assign a standard port to protocol

Jon Postel



Protocol	Port
HTTP	80
POP	110
FTP	21
SMTP	25
SMB/CISF	139

But what is the default port for Emule ???



Easy to classify

- ➡ Use fixed port
- ➡ Have an open specification
- ➡ Well know message type and size

Hard to classify

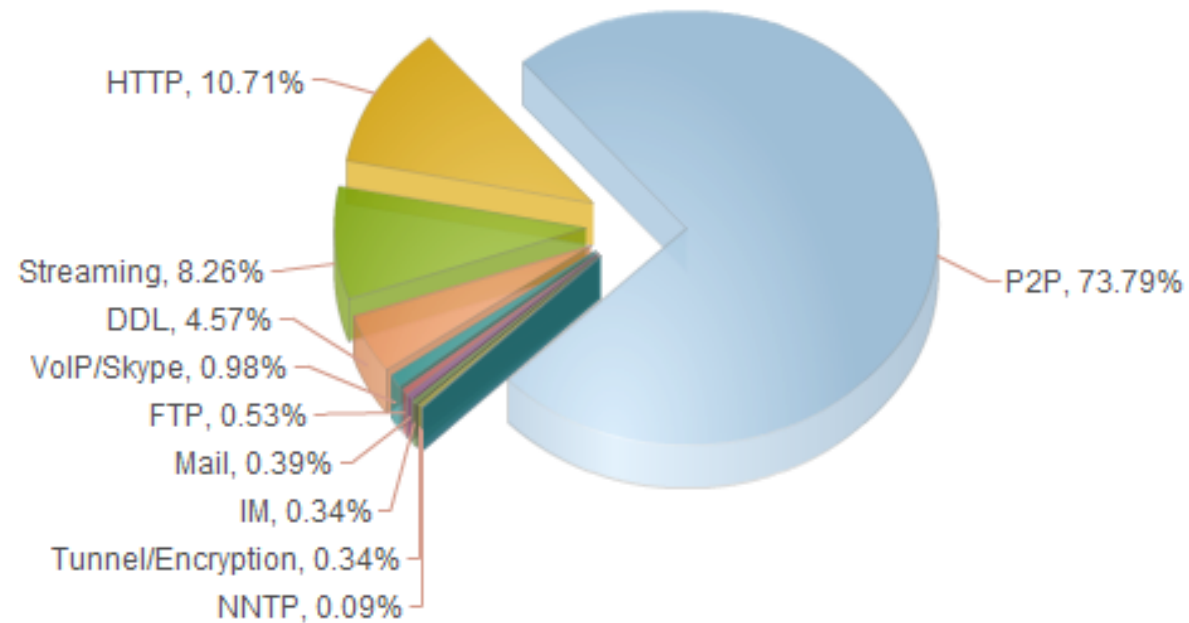
- ➡ Use dynamic port
- ➡ Hijhack well known port
- ➡ Introduce randomness in message
- ➡ Use cryptography to avoid payload analysis

Protocol Obfuscation

- ☐ Enable Protocol Obfuscation
- ☐ Allow obfuscated connections only (not recommended)
- ☐ Disable support for obfuscated connections

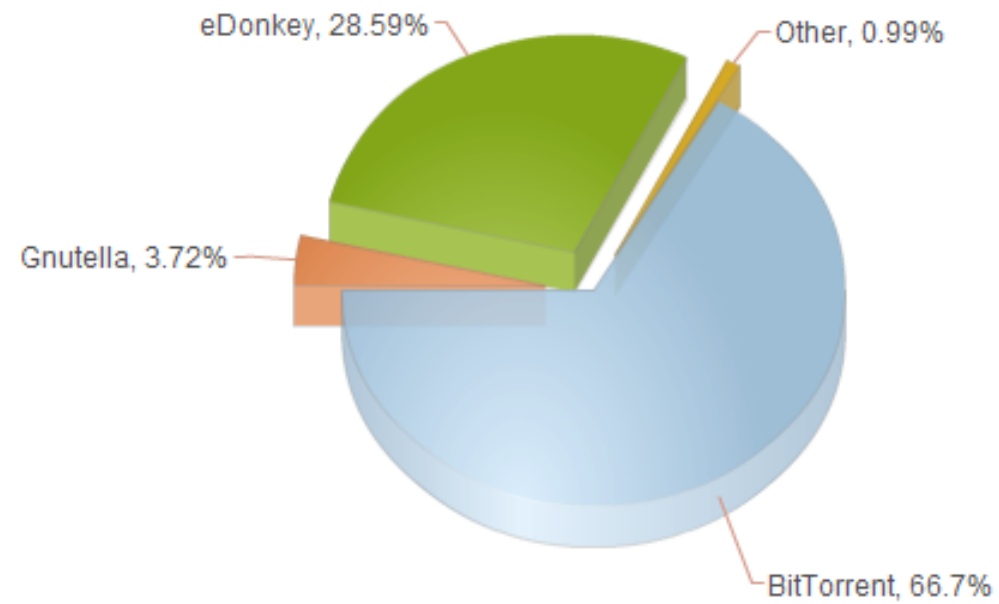
- ➡ Fighting coercion
- ➡ Preserving freedom of speech
- ➡ Bypass connectivity restriction
- ➡ Exchange illegal data

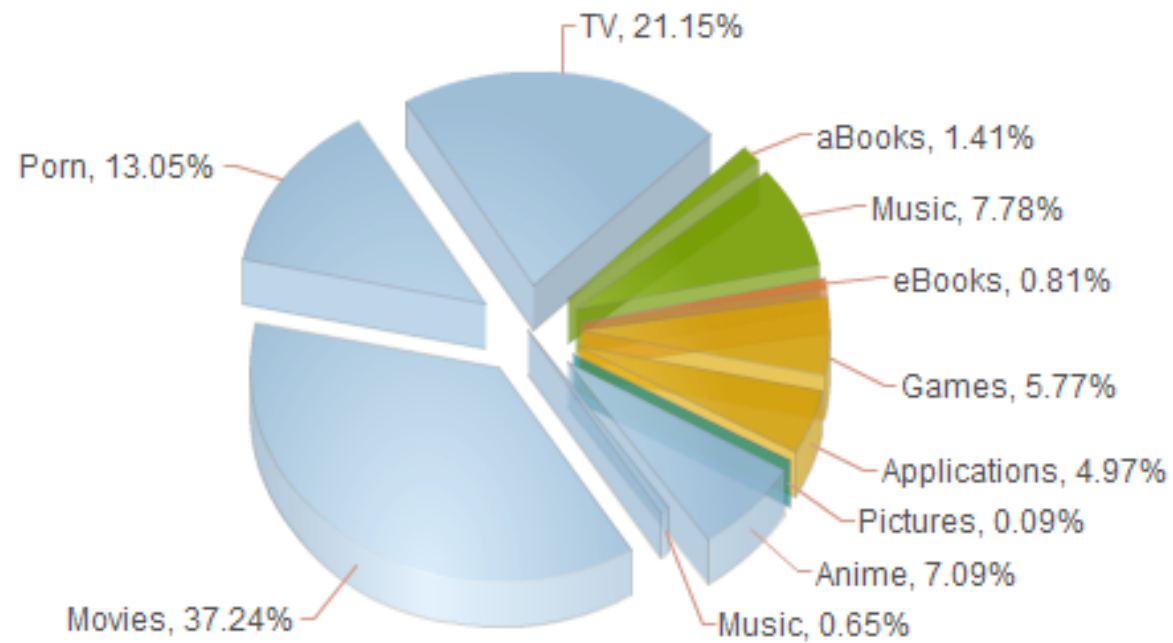
“This helps against situations where the eMule Protocol is unjustly discriminated or even completely blocked from a network by identifying its packets.”

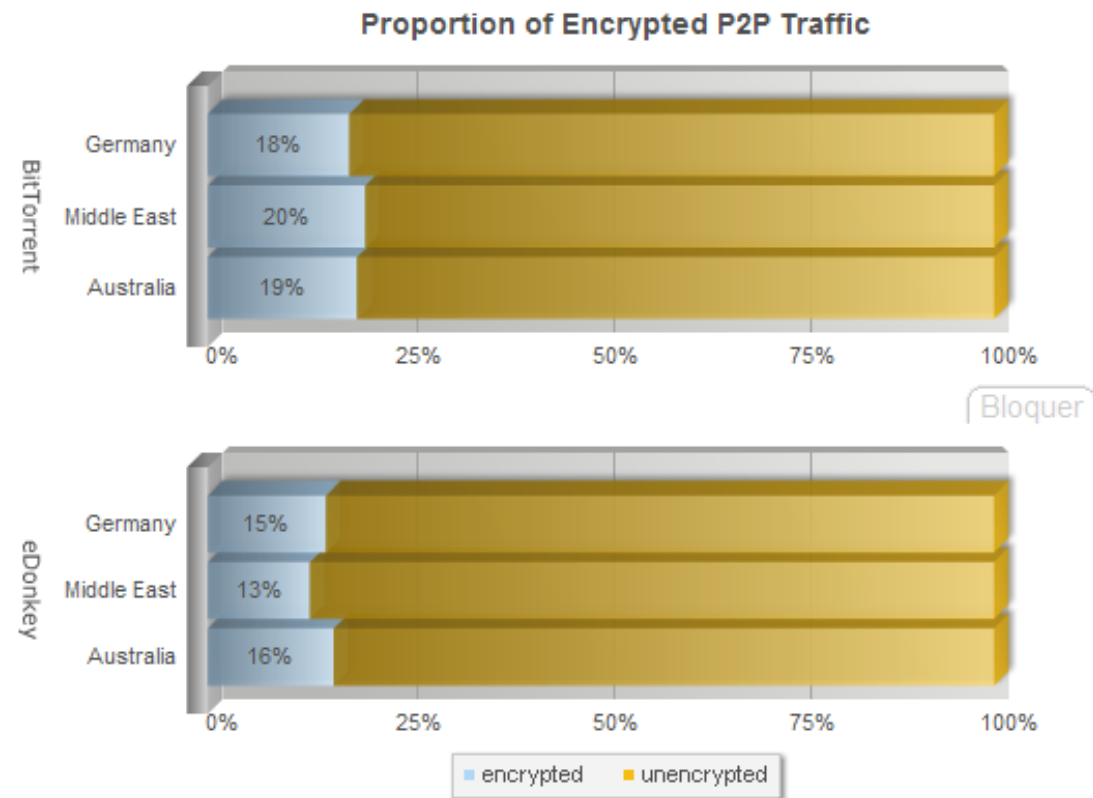


Ipoque

P2P Protocol Distribution by Volume
Germany, 2007







- Inspect the packet payload
- Known to be the most accurate method
- Payload not always available
 - Juridical reason
 - Encryption

Nice signature

```
pcmatch ssh m/^SSH-([\d]+)-/ i/protocol $1/ p|ssh|
```

A not so nice signature

```
pcmatch edonkey m`^[\xc5\xd4\xe3-\xe5].??.??.?([\x01\x02\x05\x14\x15\x16\x18\x19  
\x1a\x1b\x1c\x20\x21\x32\x33\x34\x35\x36\x38\x40\x41\x42\x43\x46\x47\x48\x49\x4a  
\x4b\x4c\x4d\x4e\x4f\x50\x51\x52\x53\x54\x55\x56\x57\x58[\x60\x81\x82\x90\x91\x9  
3\x96\x97\x98\x99\x9a\x9b\x9c\x9e\xa0\xa1\xa2\xa3\xa4]|\x59.....?[ -~  
]|\x96....$)` f/p2p/ i/eDonkey2000 or emule P2P filesharing generic signature/
```

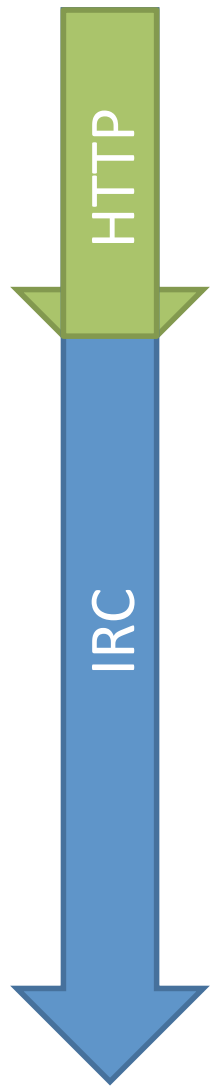
➡ Payload score use a *weighted moving average* formula:

Confidence in signature Position of the match

$$\mathbb{P} = \frac{D_{x_i} \times n + D_{x-1} \times (n - 1) + \dots + D_{x_1} \times (1)}{n + (n - 1) + \dots + 1}$$

Number of payload

➡ This formula take into account signature confidence



CONNECT irc.*****.org:6667 HTTP/1.0

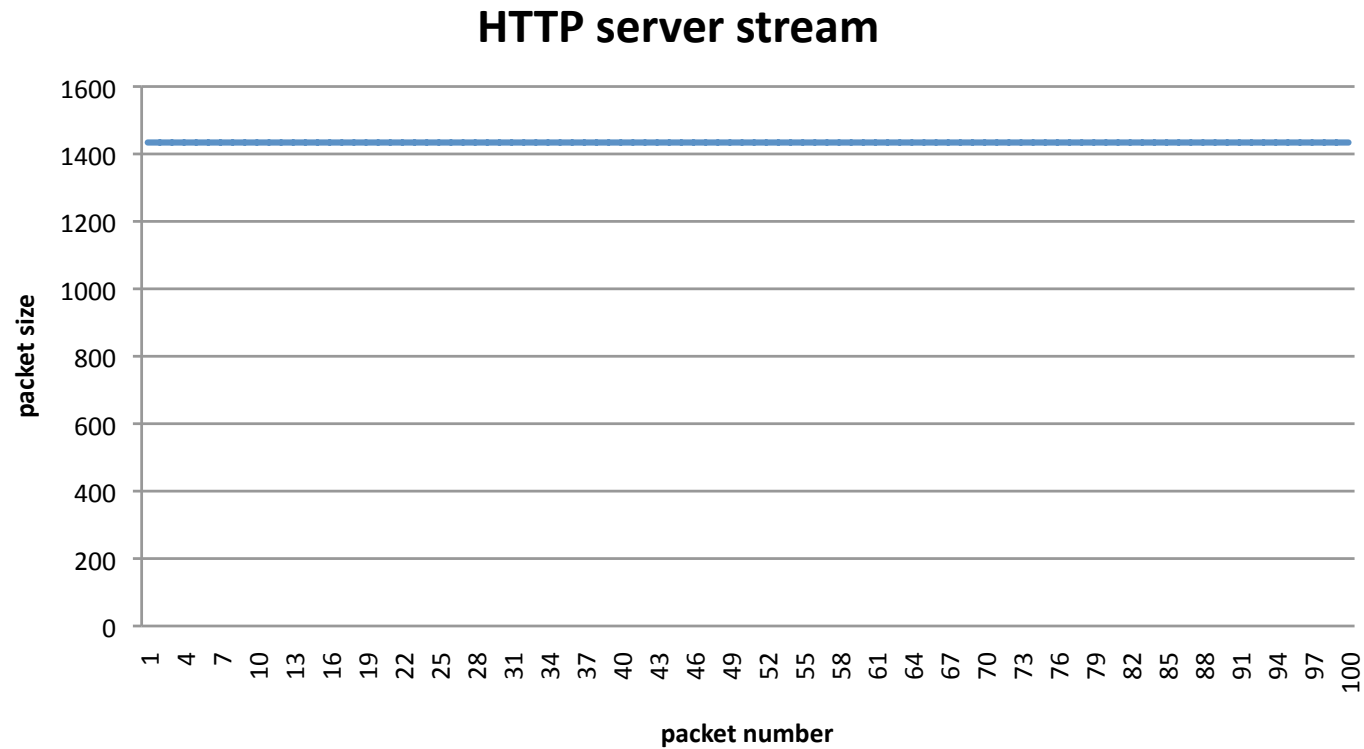
HTTP/1.0 200 Connection established

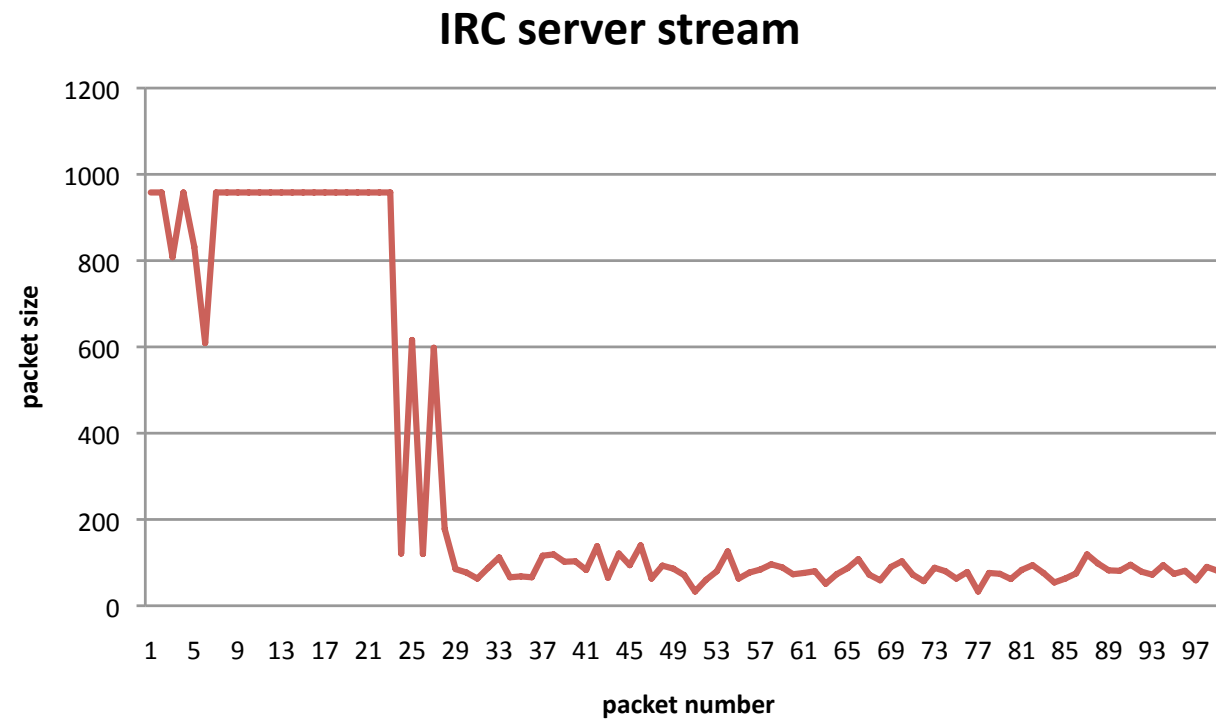
USER 0 0 0 ::

NICK ****

:irc.*****.org 001 **** :Welcome to the
****!0@xxx

- ➡ Analyze the shape of the session
- ➡ Use discriminators
 - ➡ Packet size
 - ➡ Packet delay
 - ➡ Packet Entropy



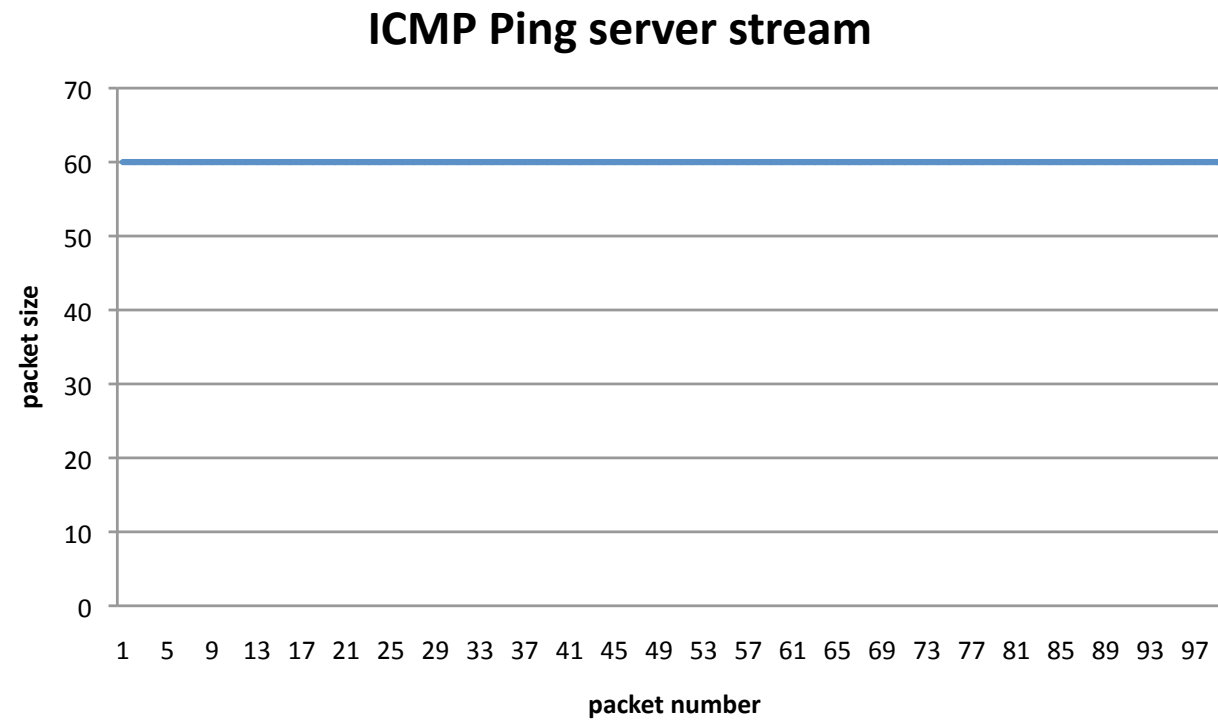


- Using 22 different set of profiles by protocol
 - A set for each packet from 1 to 10 for each stream
 - Clustering is done with the K-means algorithm
 - Lower and upper bound of each discriminator

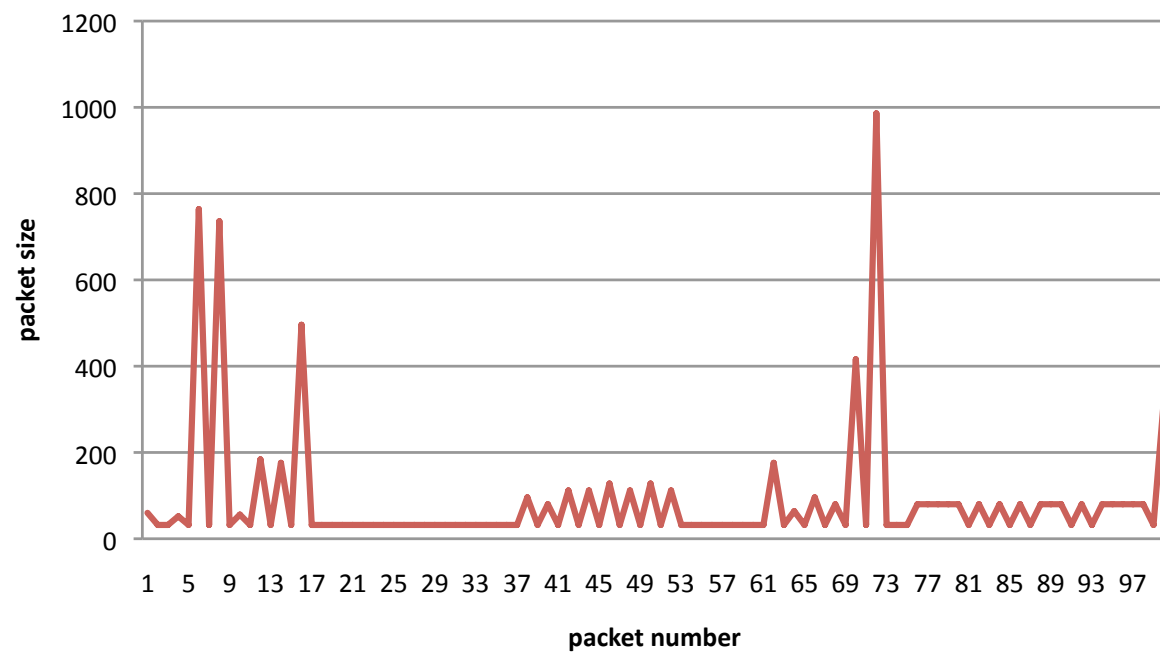
ICMP Ping profile

Ping:ICMP:2:2:64:64:995229:1004962:::7.54564:7.65728:

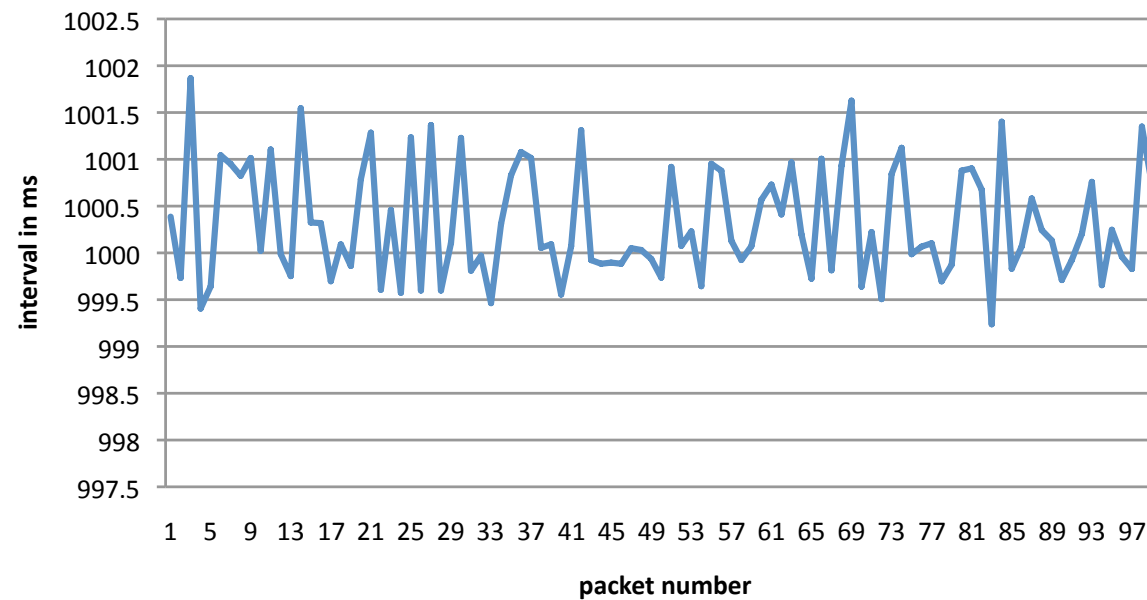
Packet size Interval between packets Packet Entropy

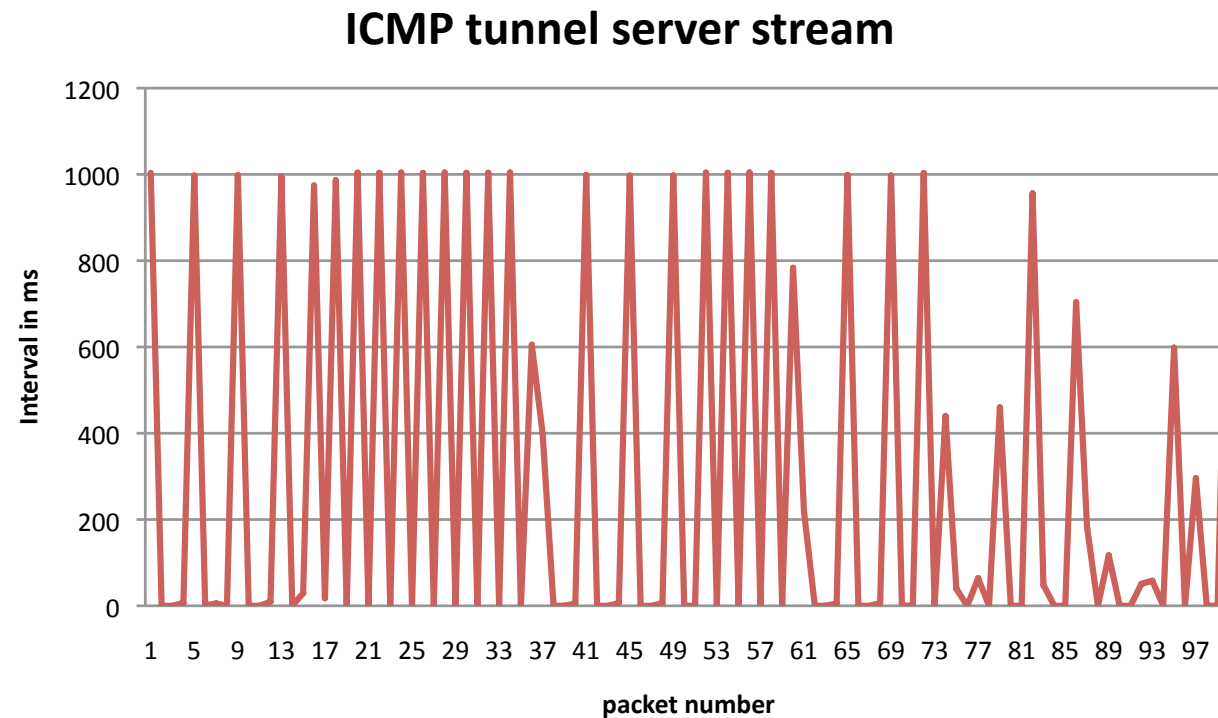


ICMP tunnel sever stream



ICMP ping server stream





Combining techniques to obtain a better identification

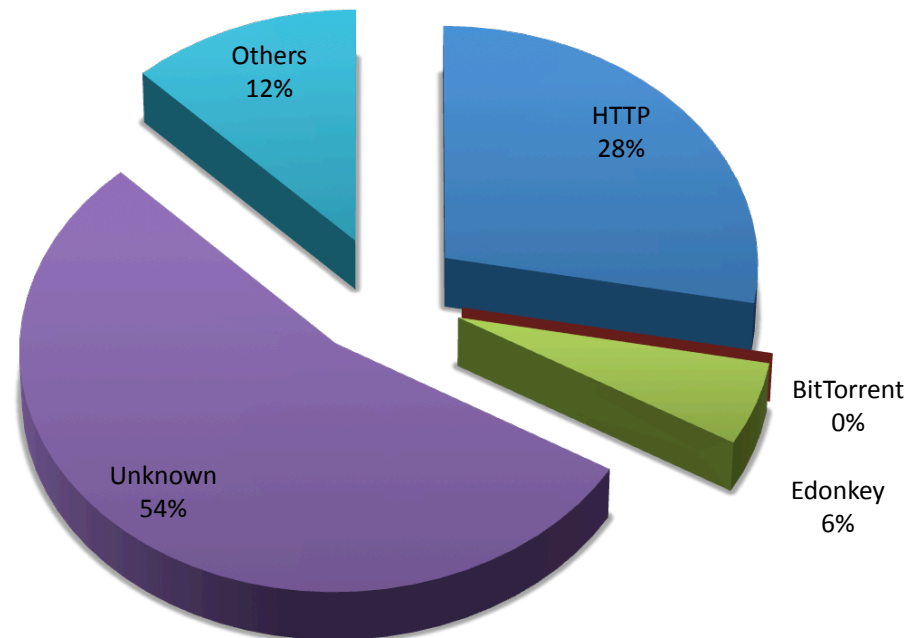
The diagram illustrates the combination of three different scores into a weighted average. At the top, three red labels with arrows point to the numerator of the formula: 'Port heuristic score' points to αH_x , 'Classifier score' points to βC_x , and 'Payload score' points to γS_x . Below the formula, three purple labels with arrows point to the denominator: 'port heuristic confidence' points to α , 'Classssifier confidence' points to β , and 'Payload confidence' points to γ .

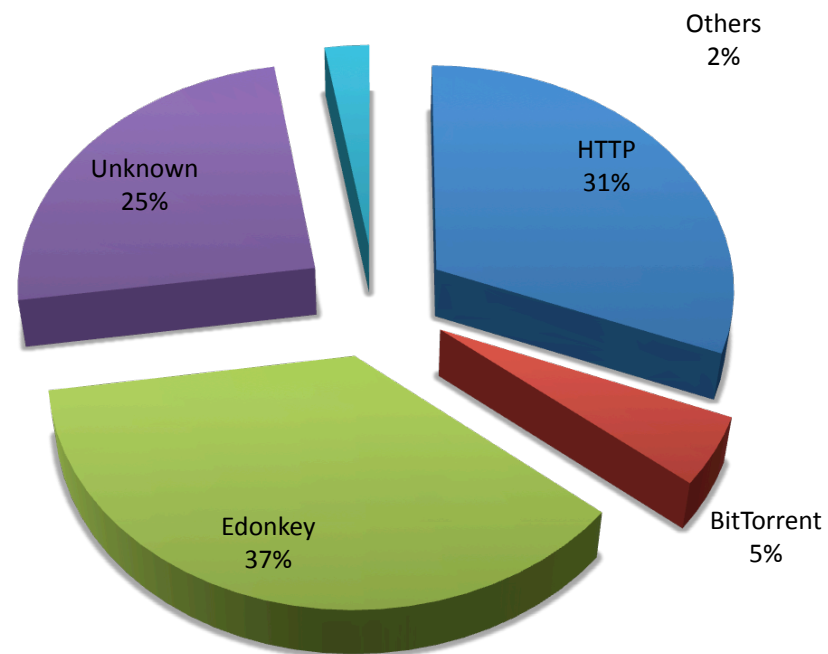
$$\mathbb{P}_x = \frac{\alpha H_x + \beta C_x + \gamma S_x}{\alpha + \beta + \gamma}$$

Implementation use the following confidence value: 1 5 10

```
(1)http (94%): x:1052 -> y.:2080 F:0/2 R::2/0 R:0/2 E:0/0 TCP CLOSED RST
(2)[Traffic] I:1 Kb/s (3pkt) O:37 Kb/s (20pkt) [Distance] C:local S:7
(3)[Protocol]:http (1.1) HyperText Transfer Protocol - RFC 2616
(4)[File] request:www.xxx.org/vip.html Ref:"http://xxx"
(5)[File] content: extension:.html family:text (X)HTML
(6)[File] request:www.xxx.org/hello.gif Ref:"http://xxx"
(7)[File] content: extension:.jpg family:image
(8)[Server] Apache httpd 2.0.52
(9)[Client] browser Internet Explorer 6.0 Windows XP
(10)[Client] proxy squid 2.5.STABLE4-20031106
(11)[Guessed protocol] http:94% Port:0% Class:100% Patt:100%
(12)[Guessed protocol] autodesk:9% Port:100% Class:n/a Patt:0%
```

$$\mathbb{P}_{http} = \frac{1 \times 0 + 5 \times 100 + 10 \times 100}{16} = 93.75$$





- More identification method
 - Number of connections ration (in/out)
 - Connection sequences
- When Random is too random

- ➡ Accurate classification is an increasing issue
- ➡ A lot need to be done to pace with evasion techniques



