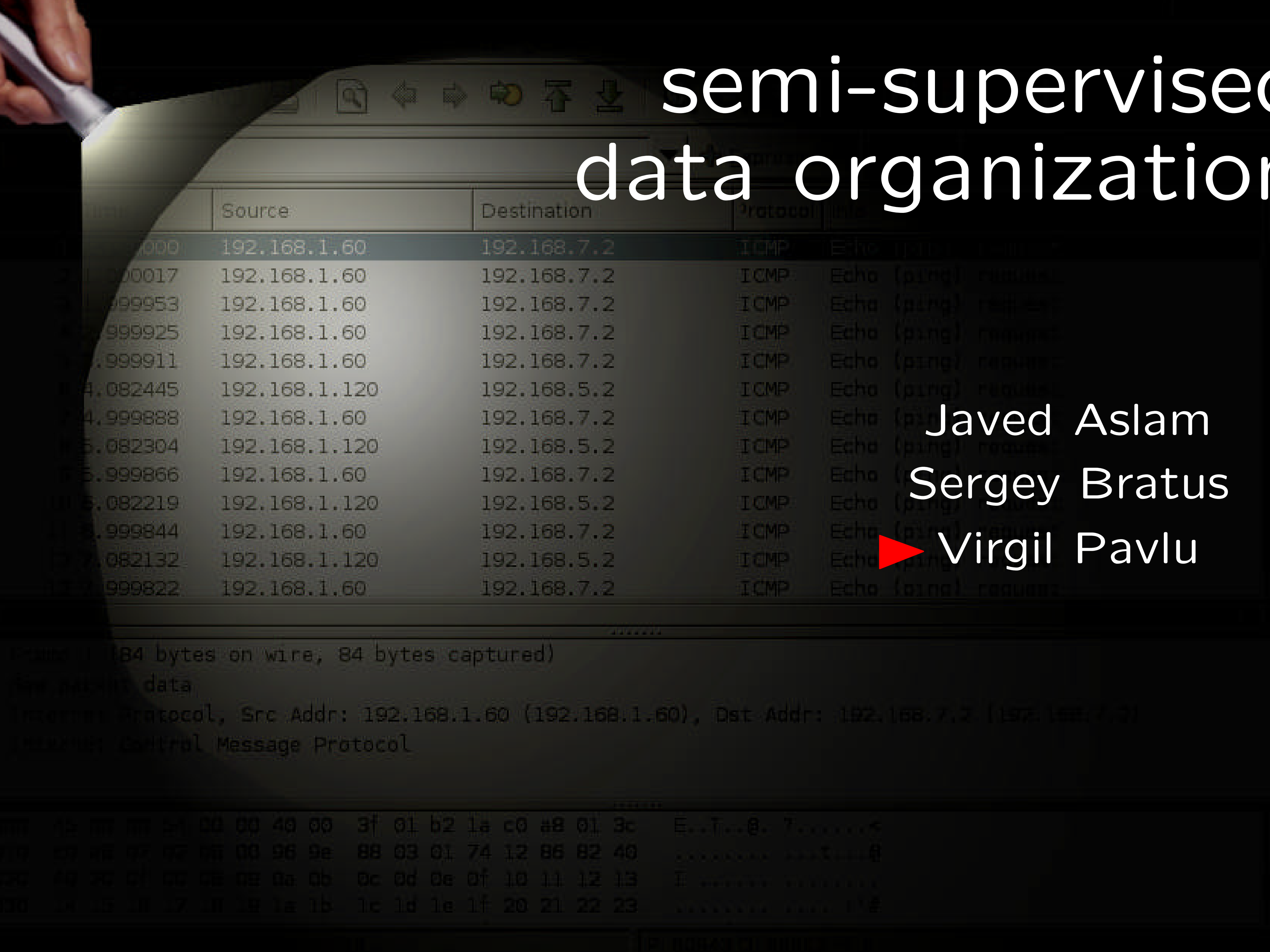
semi-supervised data organization

Javed Aslam

Sergey Bratus

▶ Virgil Pavlu

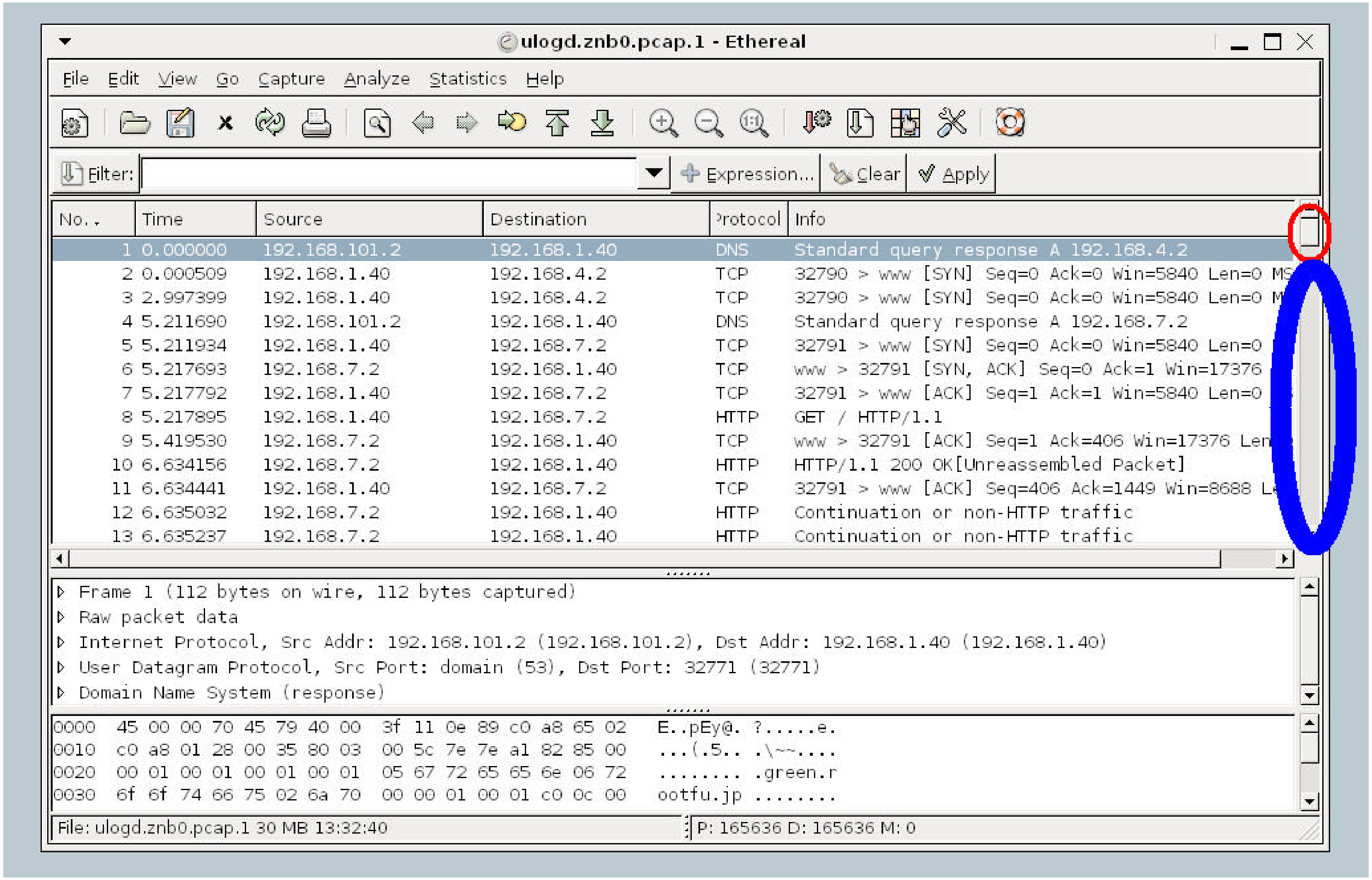


```
         Source           Destination      Protocol  Info
   000   192.168.1.60     192.168.7.2      ICMP      Echo
 00017   192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
 99953   192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
999925   192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
.999911  192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
4.082445 192.168.1.120    192.168.5.2      ICMP      Echo (ping) request
4.999888 192.168.1.60     192.168.7.2      ICMP      Echo
5.082304 192.168.1.120    192.168.5.2      ICMP      Echo (ping) request
5.999866 192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
6.082219 192.168.1.120    192.168.5.2      ICMP      Echo (ping) request
.999844  192.168.1.60     192.168.7.2      ICMP      Echo (ping) request
.082132  192.168.1.120    192.168.5.2      ICMP      Echo
999822   192.168.1.60     192.168.7.2      ICMP      Echo (ping) request

Frame 1 (84 bytes on wire, 84 bytes captured)
Raw packet data
Internet Protocol, Src Addr: 192.168.1.60 (192.168.1.60), Dst Addr: 192.168.7.2 (192.168.7.2)
Internet Control Message Protocol

3f 01 b2 1a c0 a8 01 3c   E..T..@.?.......<
88 03 01 74 12 86 82 40   ...........t...@
0c 0d 0e 0f 10 11 12 13
1c 1d 1e 1f 20 21 22 23
```

motivation : too much data

- log analysis :
  - lots of data
  - hard to search/visualize
  - very few labeled records
- but
  - easy to cluster/classify
  - interesting clusters have high density linkage
  - lots of similar records

motivation: lots of log data



No.	Time	Source	Destination	Protocol	Info
1	0.000000	192.168.101.2	192.168.1.40	DNS	Standard query response A 192.168.4.2
2	0.000509	192.168.1.40	192.168.4.2	TCP	32790 > www [SYN] Seq=0 Ack=0 Win=5840 Len=0 MS
3	2.997399	192.168.1.40	192.168.4.2	TCP	32790 > www [SYN] Seq=0 Ack=0 Win=5840 Len=0 M
4	5.211690	192.168.101.2	192.168.1.40	DNS	Standard query response A 192.168.7.2
5	5.211934	192.168.1.40	192.168.7.2	TCP	32791 > www [SYN] Seq=0 Ack=0 Win=5840 Len=0
6	5.217693	192.168.7.2	192.168.1.40	TCP	www > 32791 [SYN, ACK] Seq=0 Ack=1 Win=17376
7	5.217792	192.168.1.40	192.168.7.2	TCP	32791 > www [ACK] Seq=1 Ack=1 Win=5840 Len=0
8	5.217895	192.168.1.40	192.168.7.2	HTTP	GET / HTTP/1.1
9	5.419530	192.168.7.2	192.168.1.40	TCP	www > 32791 [ACK] Seq=1 Ack=406 Win=17376 Len
10	6.634156	192.168.7.2	192.168.1.40	HTTP	HTTP/1.1 200 OK[Unreassembled Packet]
11	6.634441	192.168.1.40	192.168.7.2	TCP	32791 > www [ACK] Seq=406 Ack=1449 Win=8688 Le
12	6.635032	192.168.7.2	192.168.1.40	HTTP	Continuation or non-HTTP traffic
13	6.635237	192.168.7.2	192.168.1.40	HTTP	Continuation or non-HTTP traffic

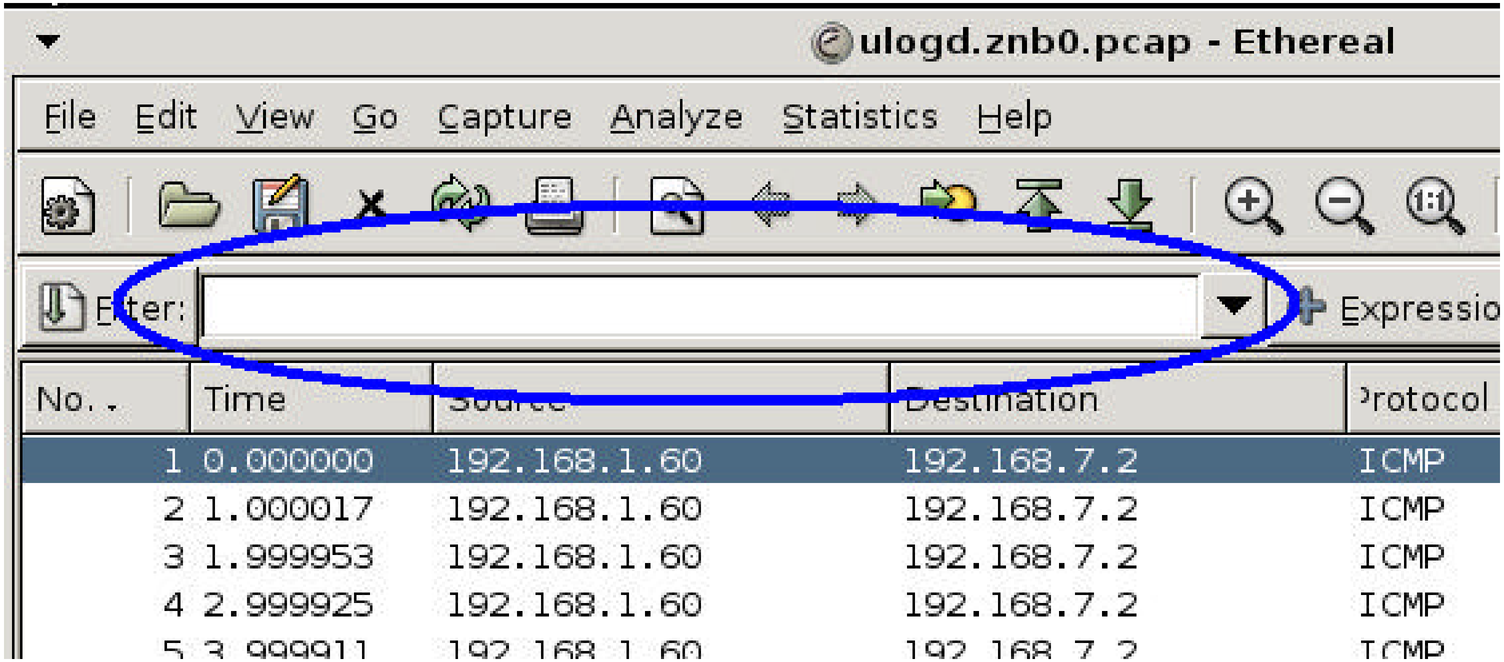
- ▷ Frame 1 (112 bytes on wire, 112 bytes captured)
- ▷ Raw packet data
- ▷ Internet Protocol, Src Addr: 192.168.101.2 (192.168.101.2), Dst Addr: 192.168.1.40 (192.168.1.40)
- ▷ User Datagram Protocol, Src Port: domain (53), Dst Port: 32771 (32771)
- ▷ Domain Name System (response)

```
0000  45 00 00 70 45 79 40 00  3f 11 0e 89 c0 a8 65 02   E..pEy@. ?.....e.
0010  c0 a8 01 28 00 35 80 03  00 5c 7e 7e a1 82 85 00   ...(.5.. .\~~....
0020  00 01 00 01 00 01 00 01  05 67 72 65 65 6e 06 72   ........ .green.r
0030  6f 6f 74 66 75 02 6a 70  00 00 01 00 01 c0 0c 00   ootfu.jp ........
```

File: ulogd.znb0.pcap.1 30 MB 13:32:40

P: 165636 D: 165636 M: 0

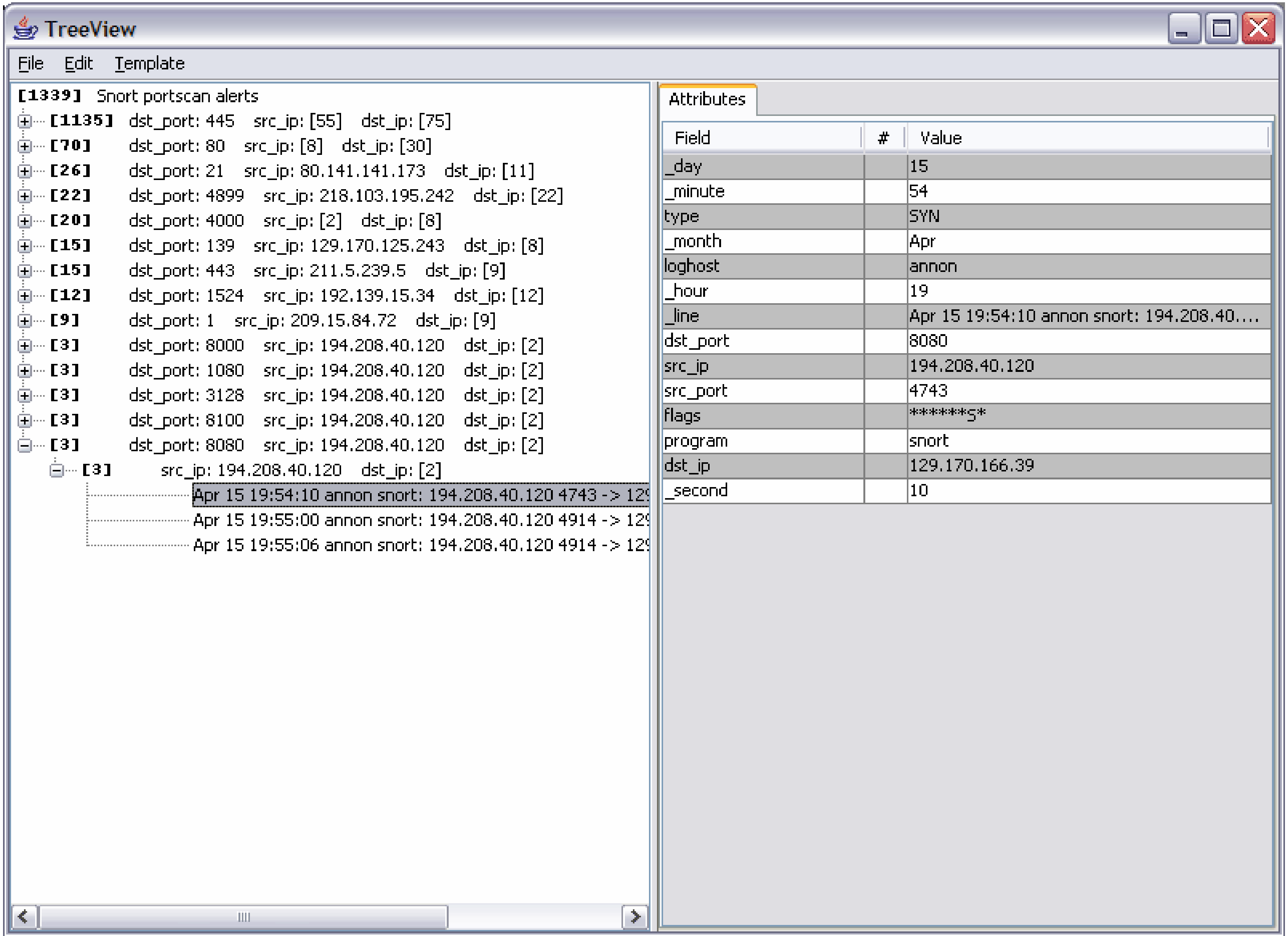
Ethereal : Filter



No. .	Time	Source	Destination	Protocol
1	0.000000	192.168.1.60	192.168.7.2	ICMP
2	1.000017	192.168.1.60	192.168.7.2	ICMP
3	1.999953	192.168.1.60	192.168.7.2	ICMP
4	2.999925	192.168.1.60	192.168.7.2	ICMP
5	3.999911	192.168.1.60	192.168.7.2	ICMP

- Filter in Ethereal :
  - if we would know what to look for....
  - similar records do not necessarily match boolean logic
  - filters get too long
  - one at a time, loses big picture
    - difficult browsing

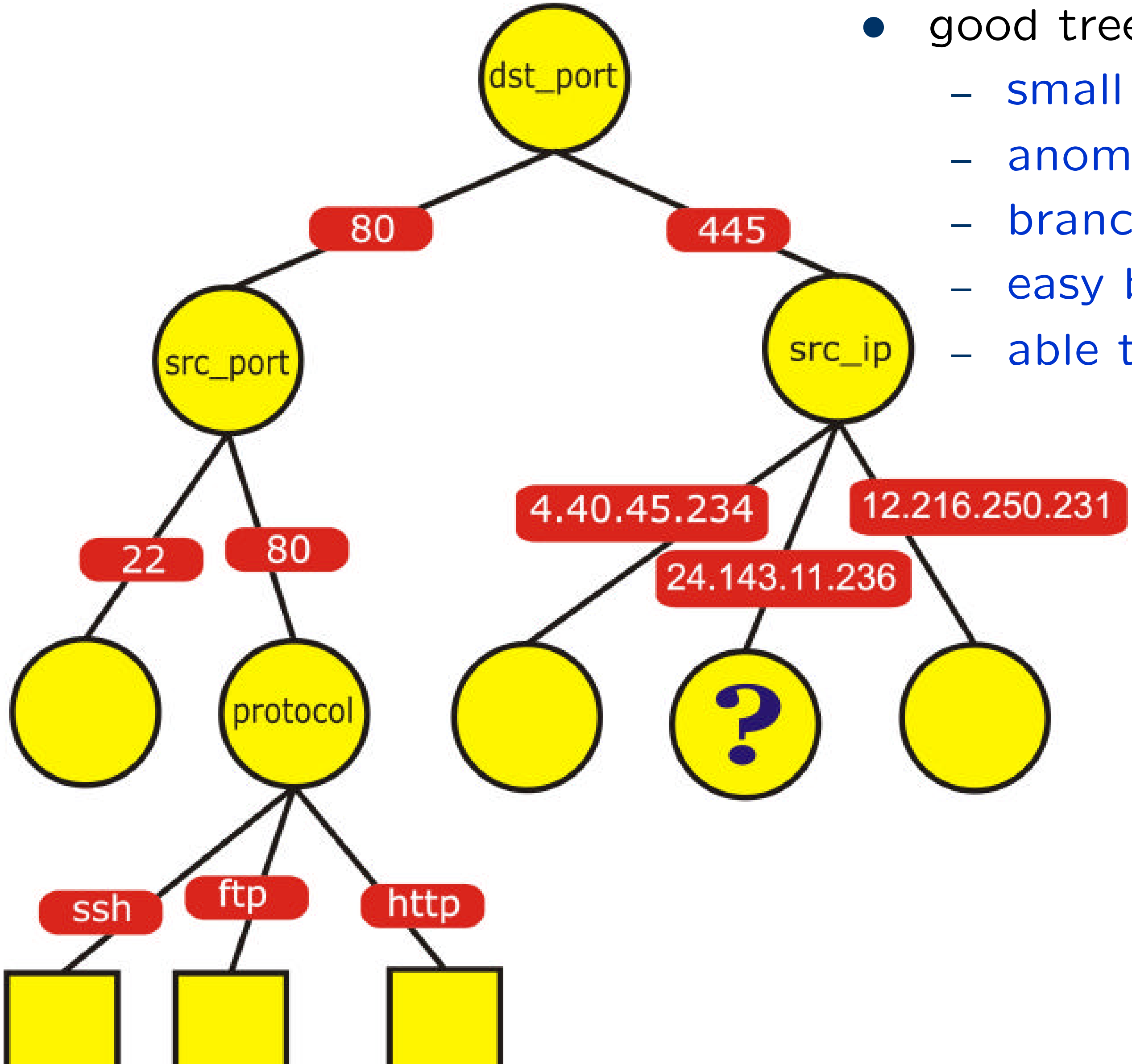
TreeView



```
[1339]  Snort portscan alerts
  [1135]  dst_port: 445   src_ip: [55]   dst_ip: [75]
  [70]    dst_port: 80   src_ip: [8]   dst_ip: [30]
  [26]    dst_port: 21   src_ip: 80.141.141.173   dst_ip: [11]
  [22]    dst_port: 4899   src_ip: 218.103.195.242   dst_ip: [22]
  [20]    dst_port: 4000   src_ip: [2]   dst_ip: [8]
  [15]    dst_port: 139   src_ip: 129.170.125.243   dst_ip: [8]
  [15]    dst_port: 443   src_ip: 211.5.239.5   dst_ip: [9]
  [12]    dst_port: 1524   src_ip: 192.139.15.34   dst_ip: [12]
  [9]     dst_port: 1   src_ip: 209.15.84.72   dst_ip: [9]
  [3]     dst_port: 8000   src_ip: 194.208.40.120   dst_ip: [2]
  [3]     dst_port: 1080   src_ip: 194.208.40.120   dst_ip: [2]
  [3]     dst_port: 3128   src_ip: 194.208.40.120   dst_ip: [2]
  [3]     dst_port: 8100   src_ip: 194.208.40.120   dst_ip: [2]
  [3]     dst_port: 8080   src_ip: 194.208.40.120   dst_ip: [2]
      [3]     src_ip: 194.208.40.120   dst_ip: [2]
              Apr 15 19:54:10 annon snort: 194.208.40.120 4743 -> 12
              Apr 15 19:55:00 annon snort: 194.208.40.120 4914 -> 12
              Apr 15 19:55:06 annon snort: 194.208.40.120 4914 -> 12
```

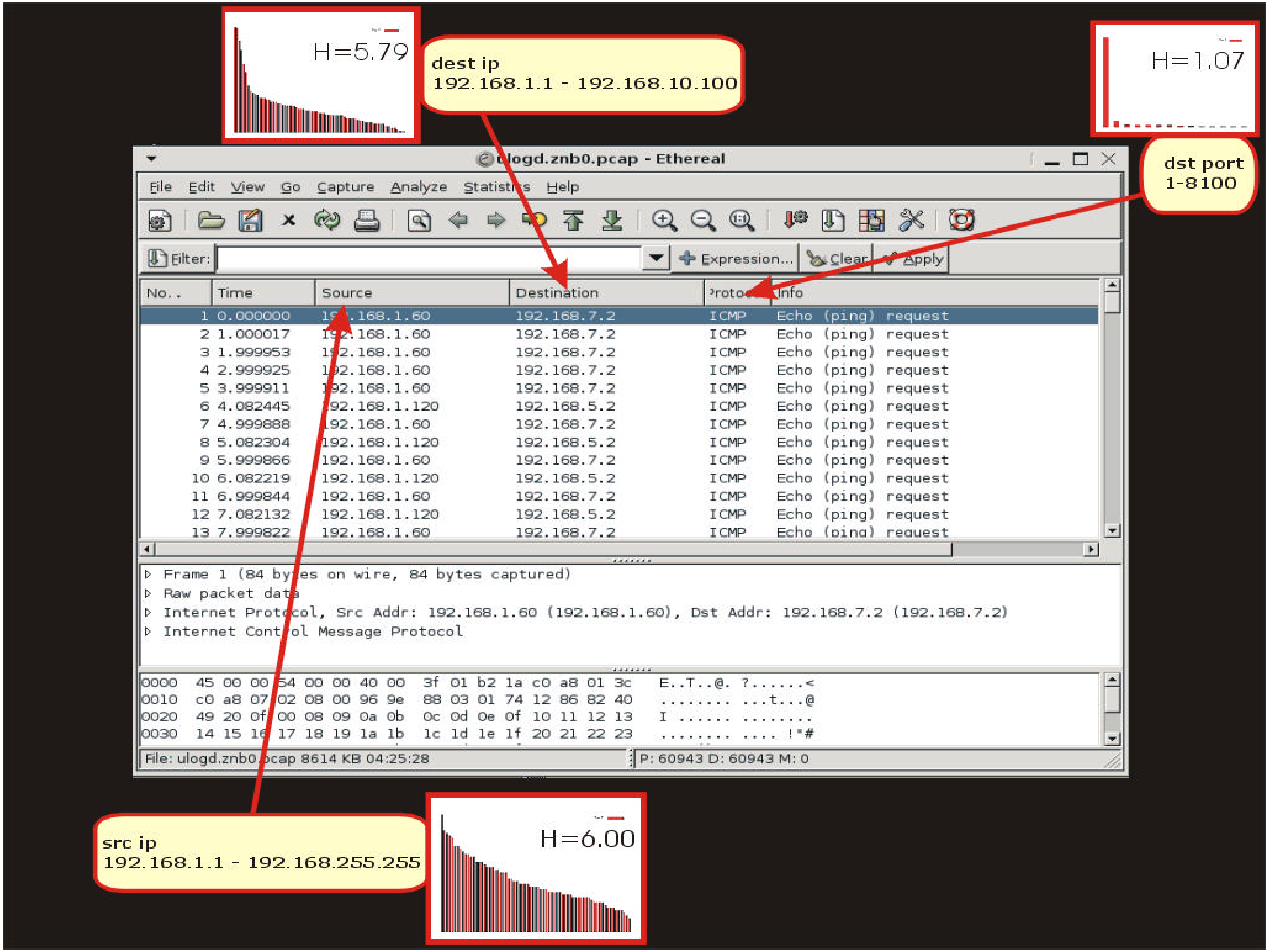
Field	#	Value
_day		15
_minute		54
type		SYN
_month		Apr
loghost		annon
_hour		19
_line		Apr 15 19:54:10 annon snort: 194.208.40....
dst_port		8080
src_ip		194.208.40.120
src_port		4743
flags		******S*
program		snort
dst_ip		129.170.166.39
_second		10

data organization : trees



- good tree :
  - small branching factor
  - anomalies grouped together
  - branches are different
  - easy browsing
  - able to use feedback

from table to tree



H=5.79
H=1.07
H=6.00
File Edit View Go Capture Analyze Statistics Help
Filter:
Expression... Clear Apply

dest ip
192.168.1.1 - 192.168.10.100

dst port
1-8100

src ip
192.168.1.1 - 192.168.255.255

ulogd.znb0.pcap - Ethereal

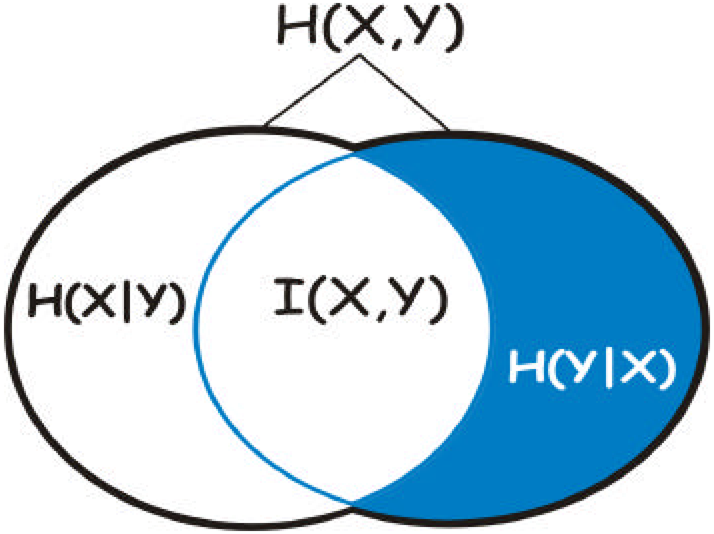
No. .	Time	Source	Destination	Protocol	Info
1	0.000000	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
2	1.000017	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
3	1.999953	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
4	2.999925	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
5	3.999911	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
6	4.082445	192.168.1.120	192.168.5.2	ICMP	Echo (ping) request
7	4.999888	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
8	5.082304	192.168.1.120	192.168.5.2	ICMP	Echo (ping) request
9	5.999866	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
10	6.082219	192.168.1.120	192.168.5.2	ICMP	Echo (ping) request
11	6.999844	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request
12	7.082132	192.168.1.120	192.168.5.2	ICMP	Echo (ping) request
13	7.999822	192.168.1.60	192.168.7.2	ICMP	Echo (ping) request

```
▷ Frame 1 (84 bytes on wire, 84 bytes captured)
▷ Raw packet data
▷ Internet Protocol, Src Addr: 192.168.1.60 (192.168.1.60), Dst Addr: 192.168.7.2 (192.168.7.2)
▷ Internet Control Message Protocol
```

```
0000  45 00 00 54 00 00 40 00  3f 01 b2 1a c0 a8 01 3c   E..T..@. ?......<
0010  c0 a8 07 02 08 00 96 9e  88 03 01 74 12 86 82 40   ........ ...t...@
0020  49 20 0f 00 08 09 0a 0b  0c 0d 0e 0f 10 11 12 13   I ...... ........
0030  14 15 16 17 18 19 1a 1b  1c 1d 1e 1f 20 21 22 23   ........ .... !"#
```

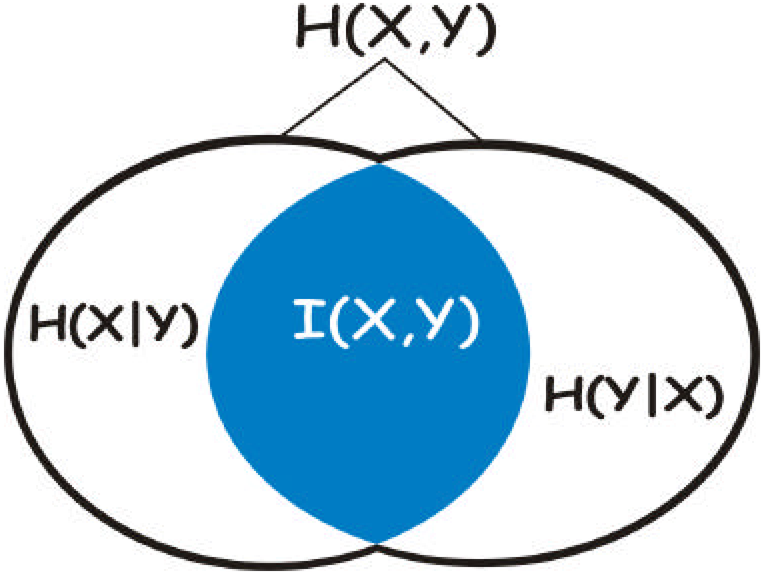
File: ulogd.znb0.pcap 8614 KB 04:25:28 | P: 60943 D: 60943 M: 0

information theory



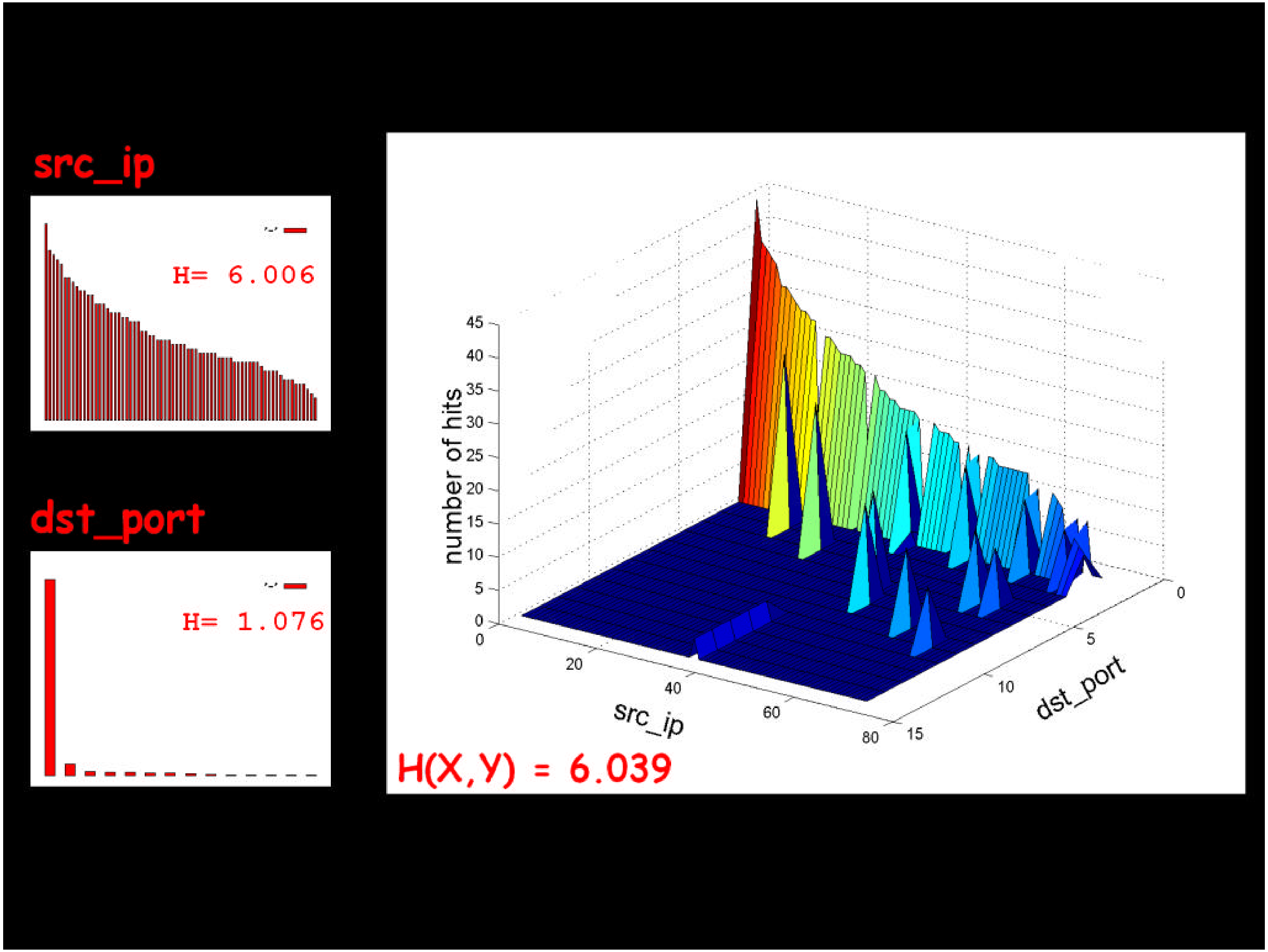
$$\begin{aligned} H(X,Y) &= H(X) + H(Y) - I(X,Y) = \\ &= H(X) + H(Y|X) \end{aligned}$$

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) = \\ &= H(Y) - I(X,Y) \end{aligned}$$

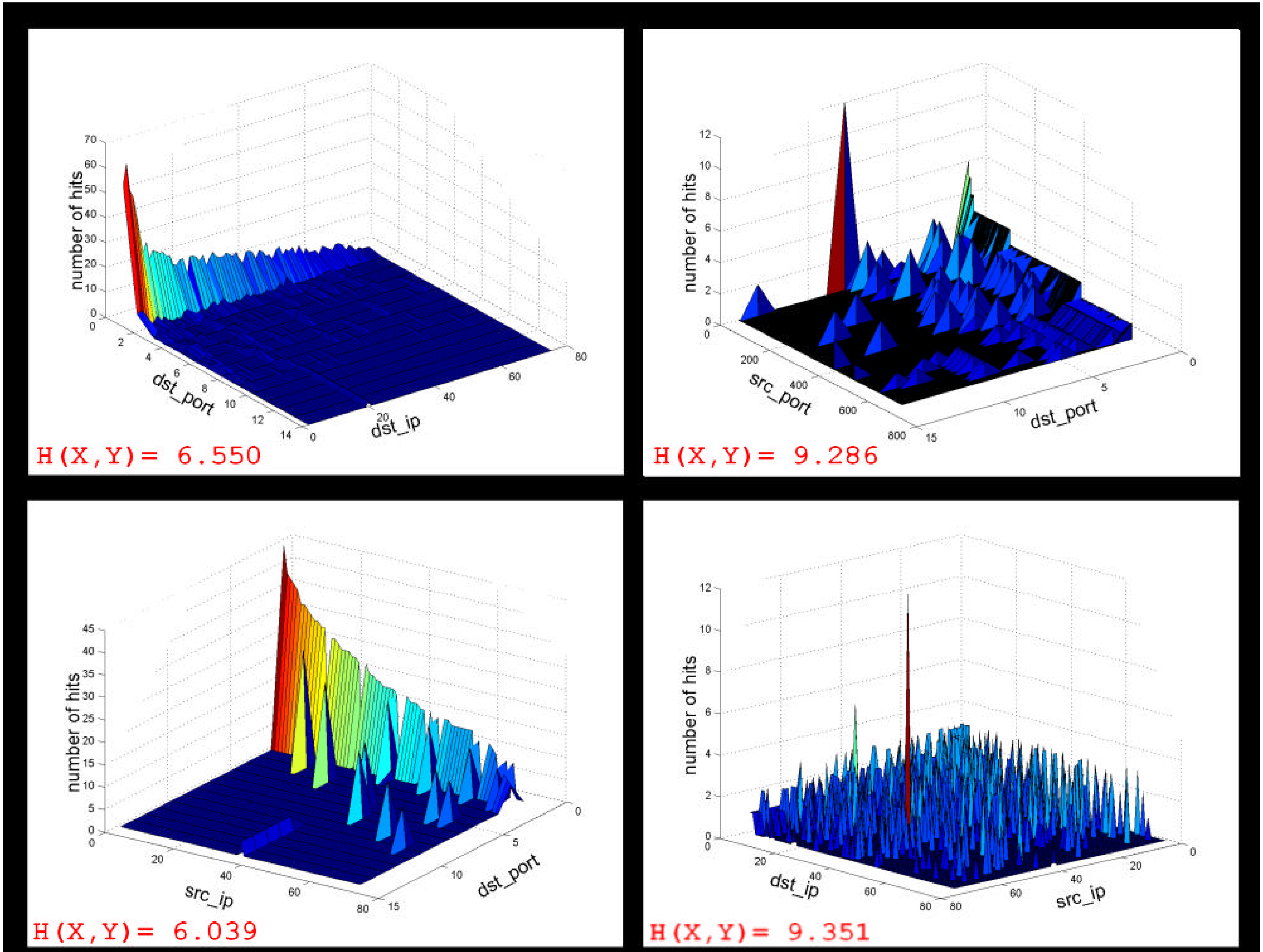


$$\begin{aligned} I(X,Y) &= H(X,Y) - H(X|Y) - H(Y|X) = \\ &= H(X) - H(Y|X) \end{aligned}$$

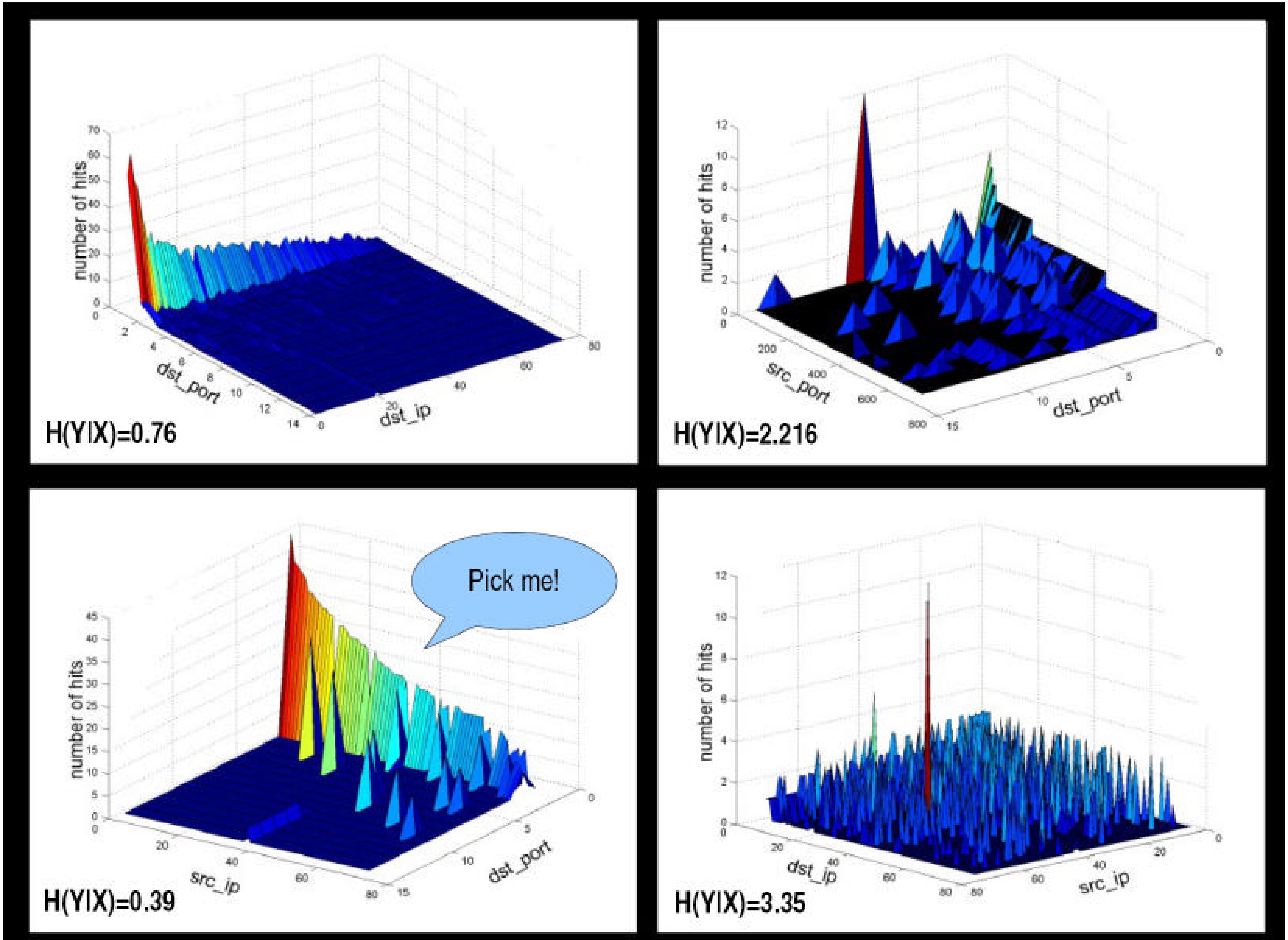
information theory



information theory



information theory



Jensen-Shannon divergence

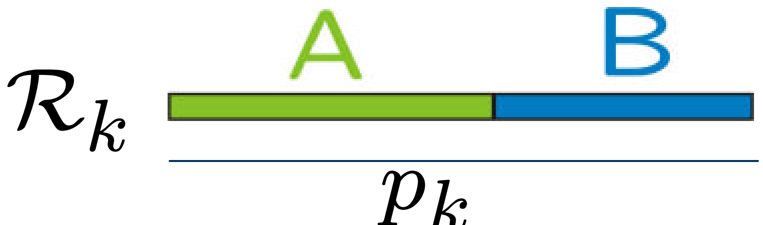
$$JS(D_1, \ldots, D_n) = H(\sum_{k=1}^{n} p_k D_k) - \sum_{k=1}^{n} p_k H(D_k)$$

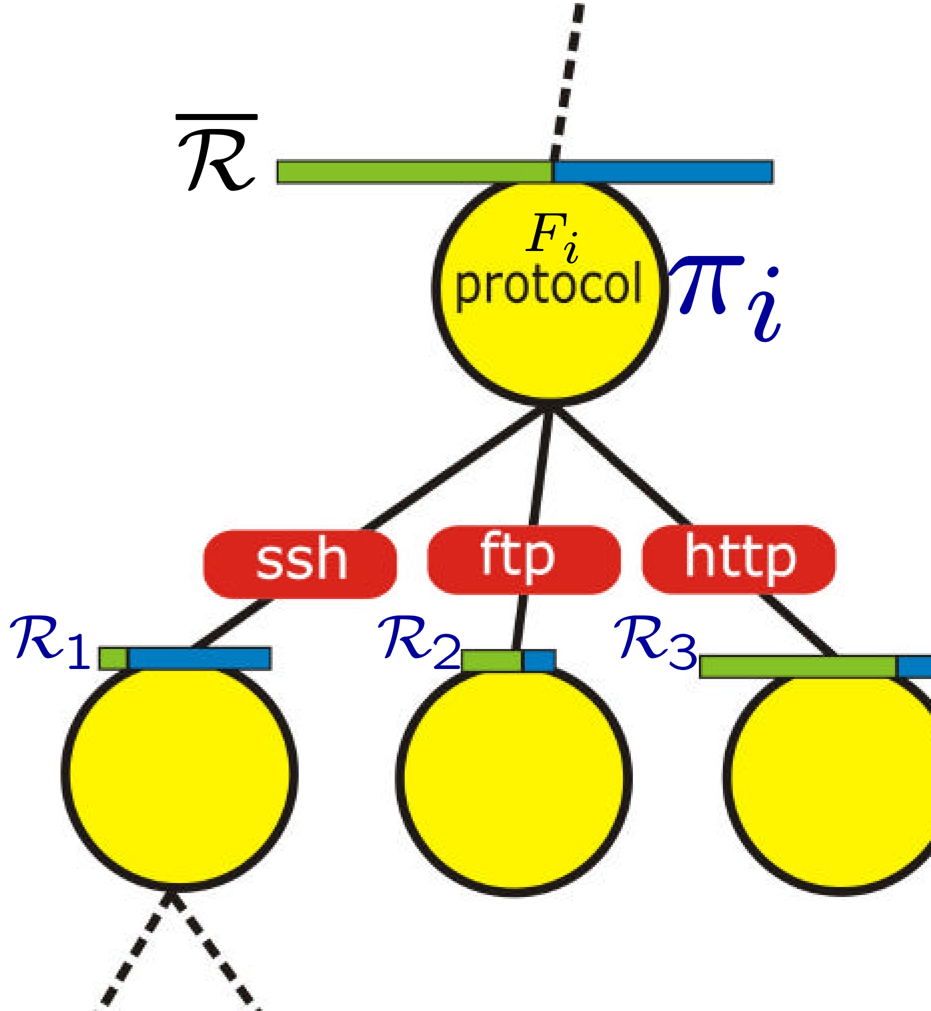
- measures (dis)similarity between several distributions
- almost a distance
- represents information reduction from encoding the distributions together rather than separately
- zero iff all distributions are identical
- better analytical properties than relative entropy

JS divergence for tree

$$JS_{\pi_i}(\mathcal{R}_1,\ldots,\mathcal{R}_{|F_i|}) = H\Big(\sum_{k=1}^{|F_i|} p_k^i \mathcal{R}_k\Big) - \sum_{k=1}^{|F_i|} p_k^i H(\mathcal{R}_k)$$

- $\overline{\mathcal{R}} = \mathcal{R}_1 \cup \ldots \cup \mathcal{R}_{|F_i|}$
- measures dissimmilarities between tree branches
- bar on top of each node indicates the number of records and their class labels(unknown)

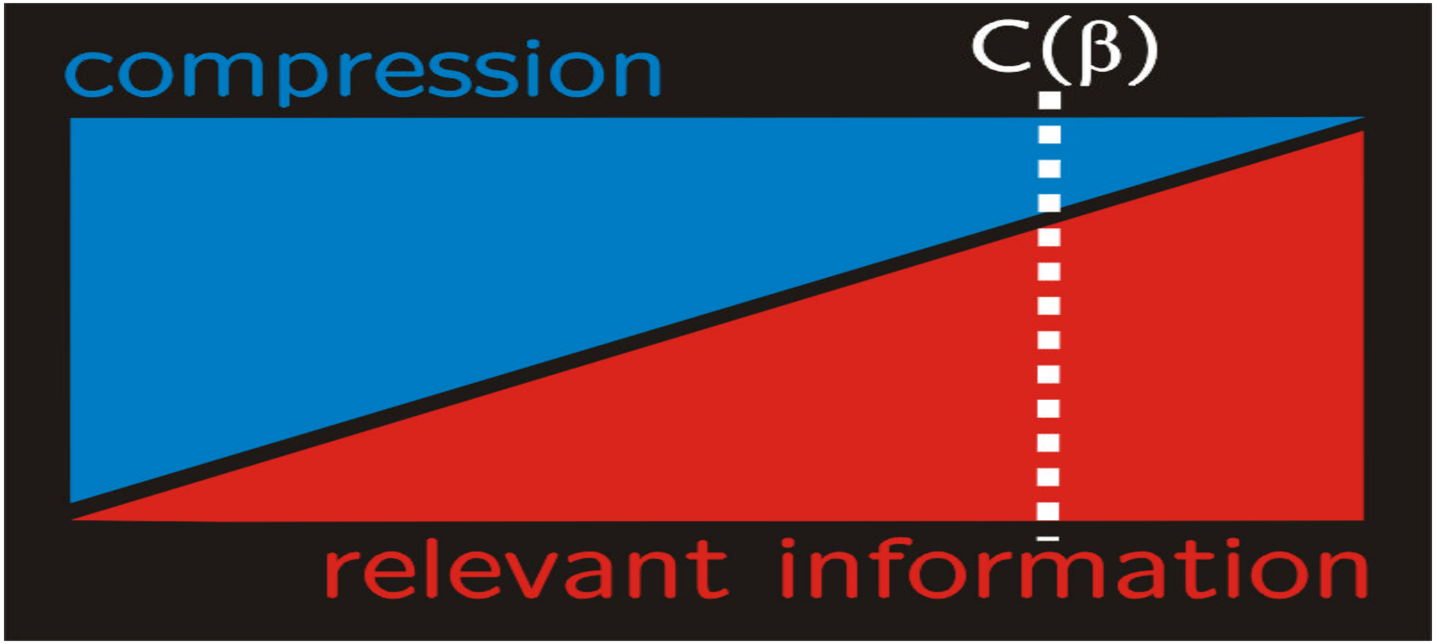




information bottleneck

- [Tishby, Pereira, Bialek]
  - X is the set of objects to be clustered/compressed
  - Y = relevant feature(s)
  - find cluster C to achieve

$$\underset{C}{argmin}\, I(X;C) - \beta I(Y;C)$$



information bottleneck and JS

- information bottleneck formula

$$\underset{C}{argmin}\, I(X;C) - \beta I(Y;C)$$

- applied with JS divergence

$$F_{(1)} = \underset{\{F_i|H(F_i)\neq 0\}}{argmin} H(F_i) - \beta \cdot JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}),$$

semi-supervised

- L=set of labels provided by the user
  - Only a tiny percentage of records will be marked either way.
  - Not all copies of identical records (or very similar) records will be marked

- all quantities of interest become conditionals of L:

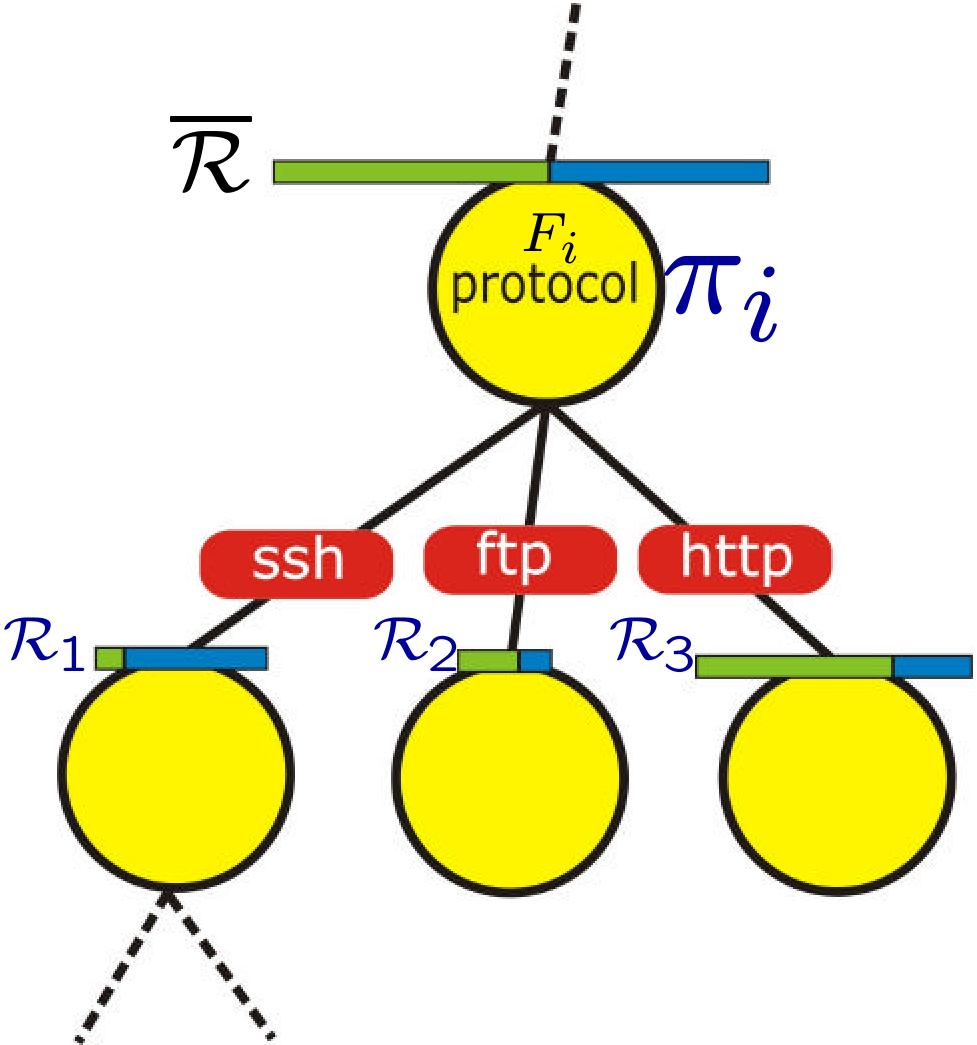
$$H(F|L) = H(F) - I(F;L)$$

$$\begin{aligned} JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}|L) &= JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}) \\ &- (I(\overline{\mathcal{R}}; L) - \sum_{k=1}^{|F_i|} p_k^i I(\mathcal{R}_k; L)) \end{aligned}$$

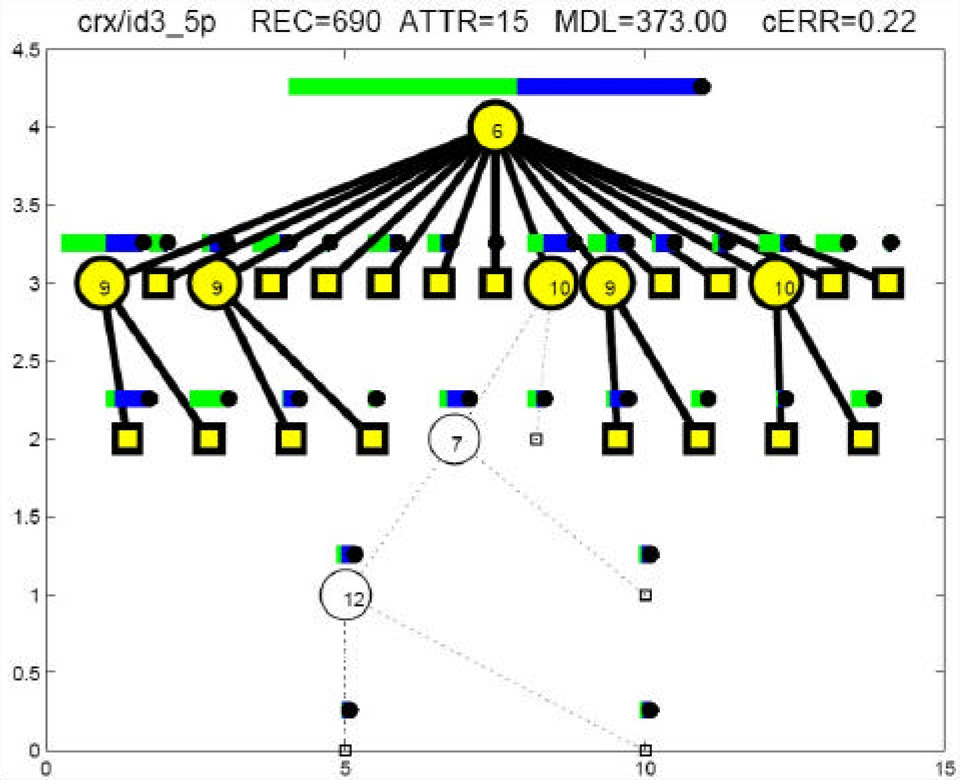
semi-supervised

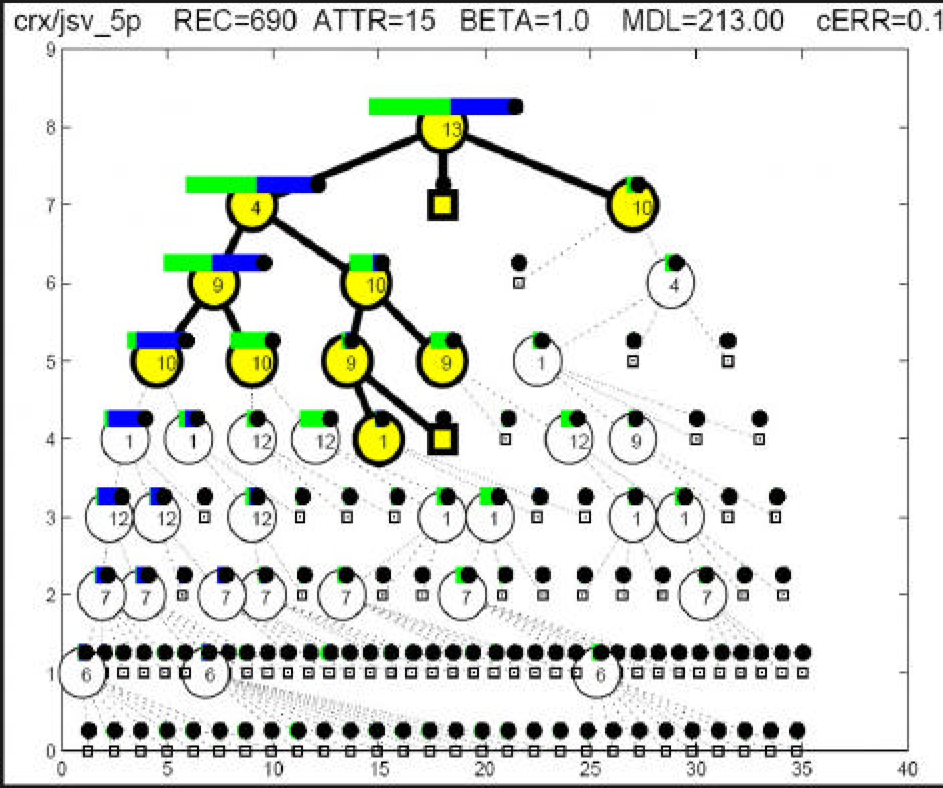
$$F^{(1)} = \underset{\{F_i|H(F_i)\neq 0\}}{argmin}\ H(F_i|L) - \beta \cdot JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}|L),$$

$$\overline{\mathcal{R}} = \mathcal{R}_1 \cup \ldots \cup \mathcal{R}_{|F_i|}$$

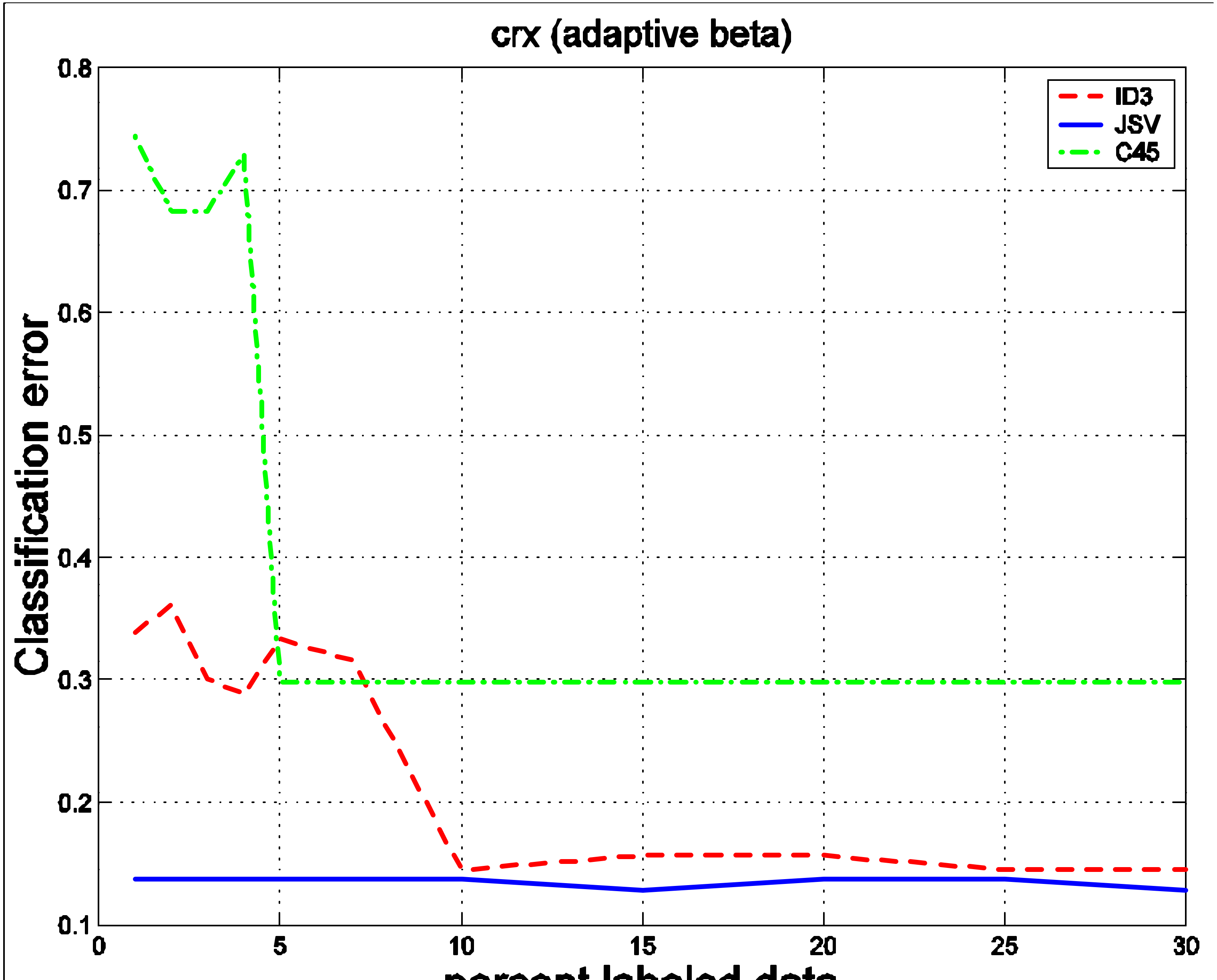


results


crx/id3_5p REC=690 ATTR=15 MDL=373.00 cERR=0.22
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
5
10
15
6
9
9
10
9
10
7
12

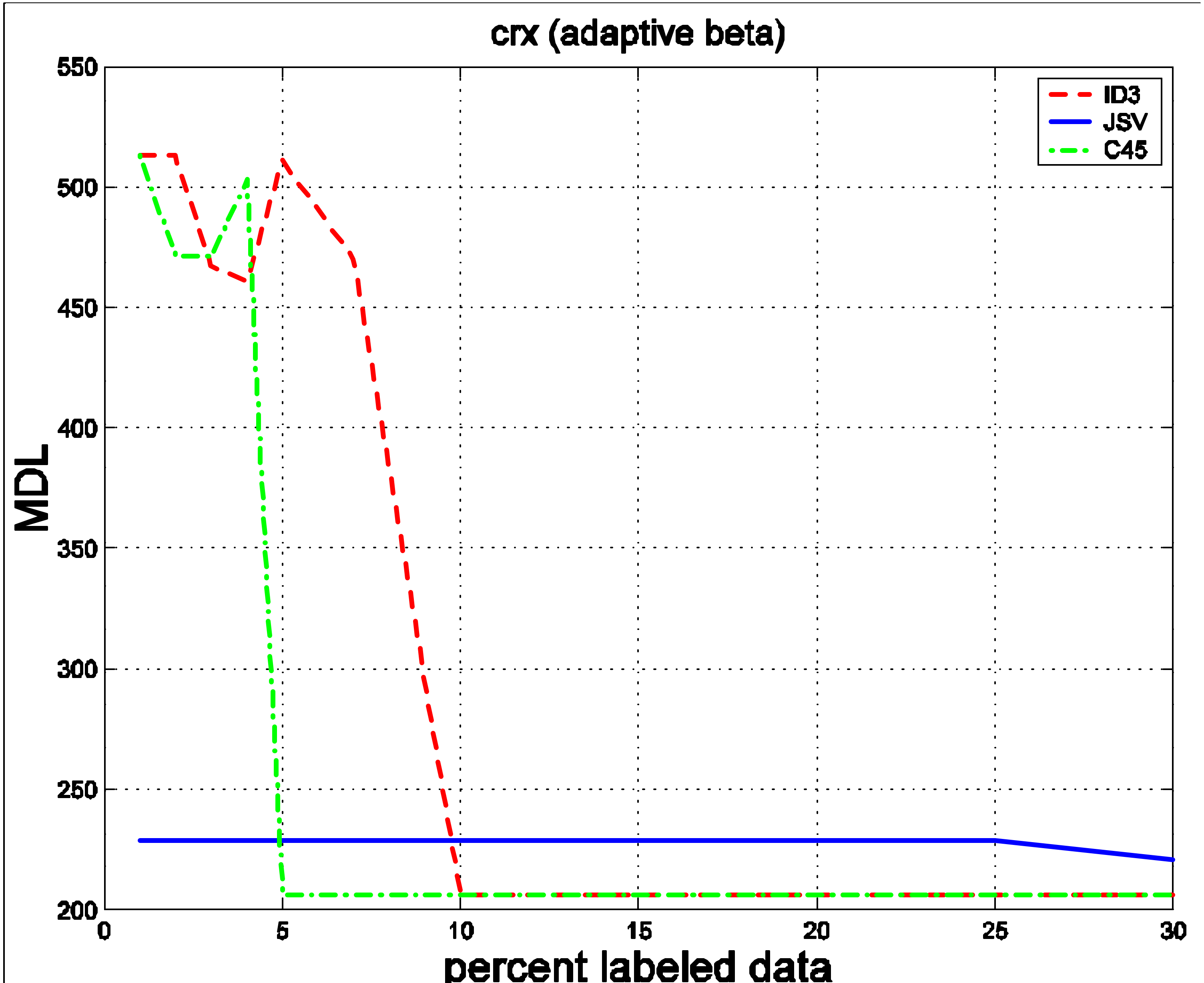

crx/jsv_5p REC=690 ATTR=15 BETA=1.0 MDL=213.00 cERR=0.1
9
8
7
6
5
4
3
2
1
0
0
5
10
15
20
25
30
35
40
13
4
10
9
10
4
10
10
9
9
1
1
1
12
12
1
12
9
12
12
12
1
1
1
1
7
7
7
7
7
7
7
6
6
6

results



crx (adaptive beta)
Classification error
percent labeled data
ID3
JSV
C45
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30

results


crx (adaptive beta)
MDL
550
500
450
400
350
300
250
200
0
5
10
15
20
25
30
percent labeled data
ID3
JSV
C45

results

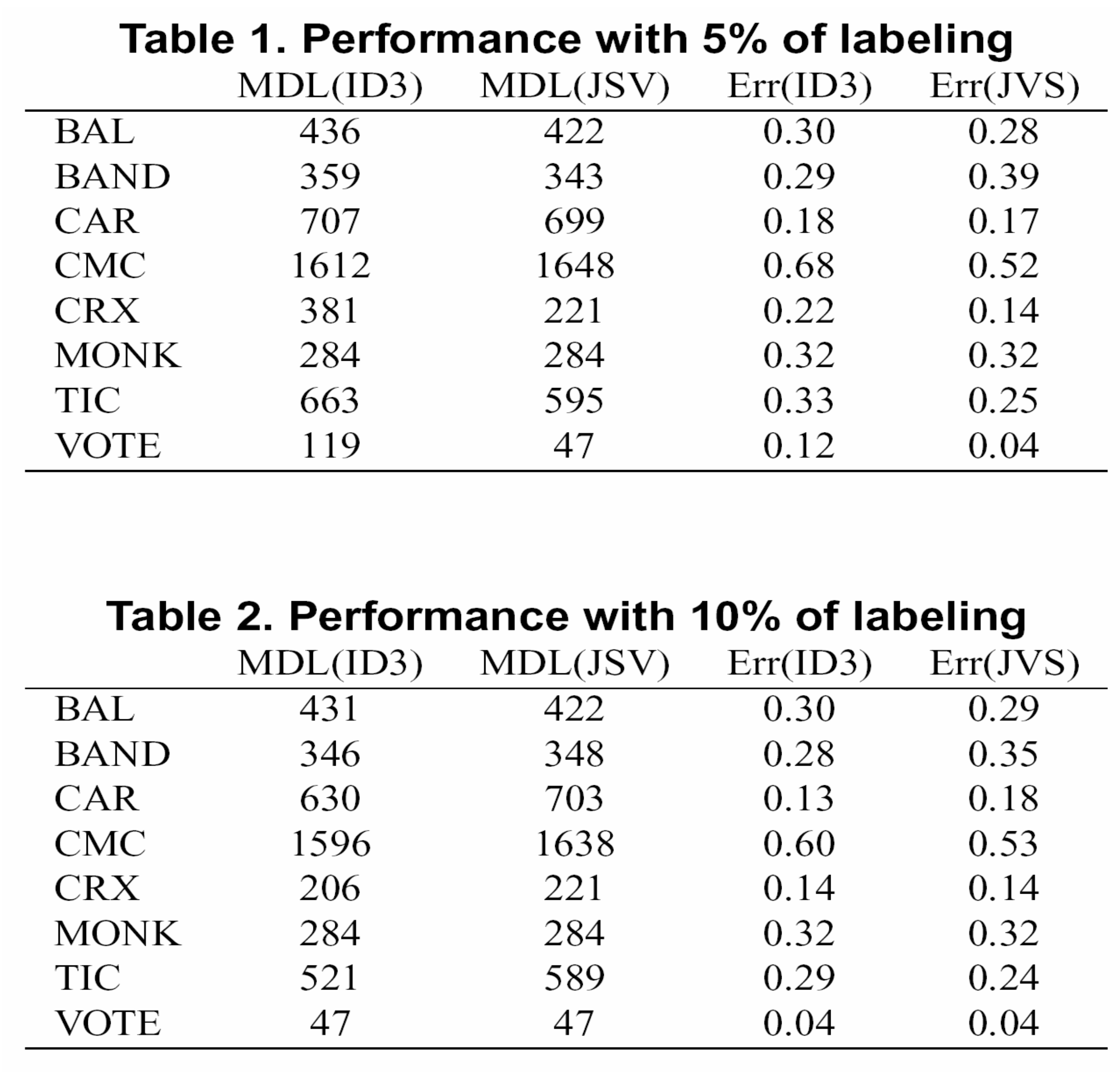


Table 1. Performance with 5% of labeling

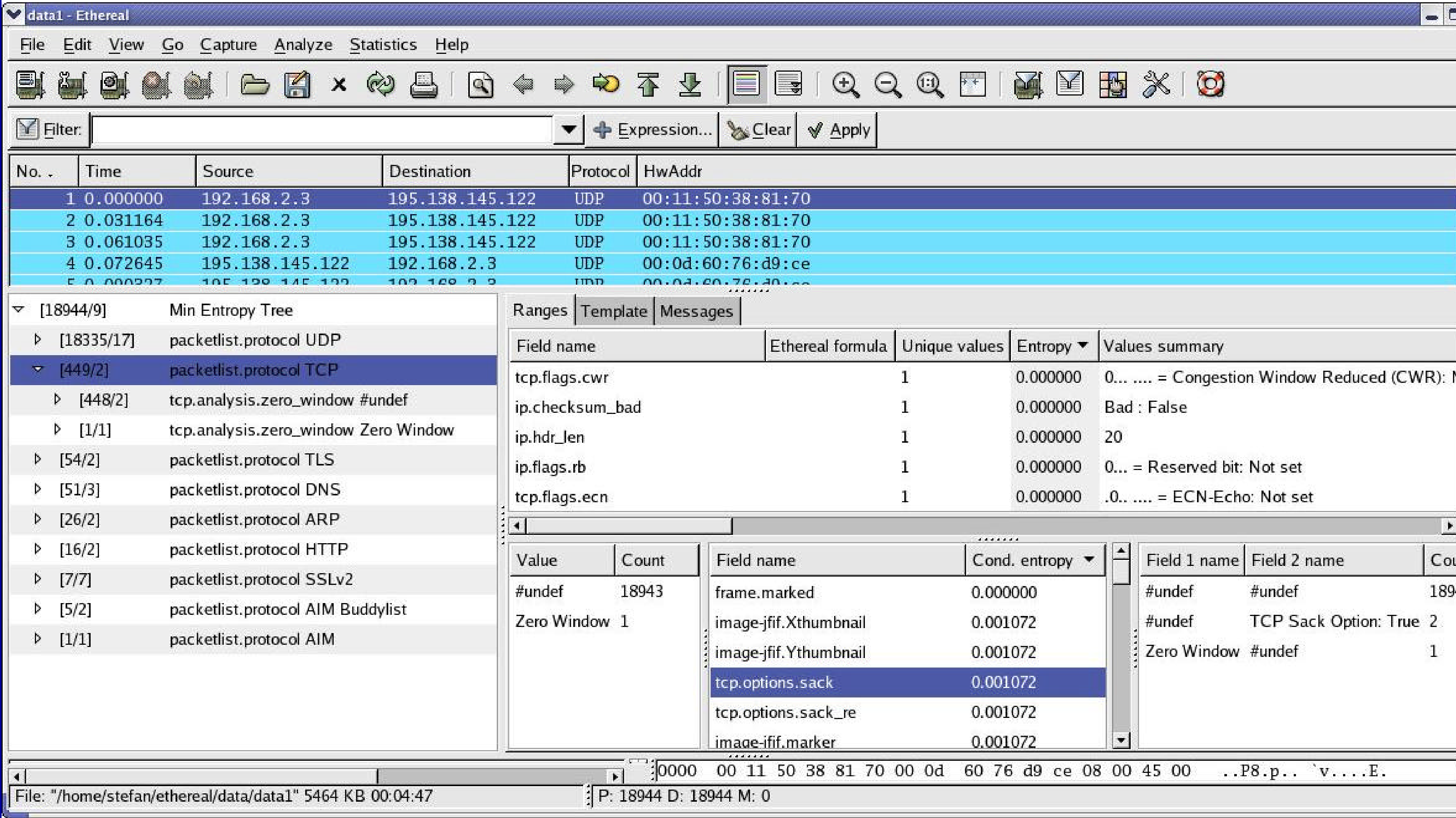
	MDL(ID3)	MDL(JSV)	Err(ID3)	Err(JVS)
BAL	436	422	0.30	0.28
BAND	359	343	0.29	0.39
CAR	707	699	0.18	0.17
CMC	1612	1648	0.68	0.52
CRX	381	221	0.22	0.14
MONK	284	284	0.32	0.32
TIC	663	595	0.33	0.25
VOTE	119	47	0.12	0.04

Table 2. Performance with 10% of labeling

	MDL(ID3)	MDL(JSV)	Err(ID3)	Err(JVS)
BAL	431	422	0.30	0.29
BAND	346	348	0.28	0.35
CAR	630	703	0.13	0.18
CMC	1596	1638	0.60	0.53
CRX	206	221	0.14	0.14
MONK	284	284	0.32	0.32
TIC	521	589	0.29	0.24
VOTE	47	47	0.04	0.04

application: ethereal plugin

- kerf.cs.dartmouth.edu



No.	Time	Source	Destination	Protocol	HwAddr
1	0.000000	192.168.2.3	195.138.145.122	UDP	00:11:50:38:81:70
2	0.031164	192.168.2.3	195.138.145.122	UDP	00:11:50:38:81:70
3	0.061035	192.168.2.3	195.138.145.122	UDP	00:11:50:38:81:70
4	0.072645	195.138.145.122	192.168.2.3	UDP	00:0d:60:76:d9:ce
5	0.090327	195.138.145.122	192.168.2.3	UDP	00:0d:60:76:d9:ce

[18944/9]	Min Entropy Tree
[18335/17]	packetlist.protocol UDP
[449/2]	packetlist.protocol TCP
[448/2]	tcp.analysis.zero_window #undef
[1/1]	tcp.analysis.zero_window Zero Window
[54/2]	packetlist.protocol TLS
[51/3]	packetlist.protocol DNS
[26/2]	packetlist.protocol ARP
[16/2]	packetlist.protocol HTTP
[7/7]	packetlist.protocol SSLv2
[5/2]	packetlist.protocol AIM Buddylist
[1/1]	packetlist.protocol AIM

Field name	Ethereal formula	Unique values	Entropy	Values summary
tcp.flags.cwr		1	0.000000	0... .... = Congestion Window Reduced (CWR): N
ip.checksum_bad		1	0.000000	Bad : False
ip.hdr_len		1	0.000000	20
ip.flags.rb		1	0.000000	0... = Reserved bit: Not set
tcp.flags.ecn		1	0.000000	.0.. .... = ECN-Echo: Not set

Value	Count
#undef	18943
Zero Window	1

Field name	Cond. entropy
frame.marked	0.000000
image-jfif.Xthumbnail	0.001072
image-jfif.Ythumbnail	0.001072
tcp.options.sack	0.001072
tcp.options.sack_re	0.001072
image-jfif.marker	0.001072

Field 1 name	Field 2 name	Cou
#undef	#undef	189
#undef	TCP Sack Option: True	2
Zero Window	#undef	1

```
0000  00 11 50 38 81 70 00 0d  60 76 d9 ce 08 00 45 00   ..P8.p.. `v....E.
```

File: "/home/stefan/ethereal/data/data1" 5464 KB 00:04:47

P: 18944 D: 18944 M: 0

Thank You