Using Traffic Analysis to Detect Email Spam

Richard Clayton

TERENA, Lyngby, 22nd May 2007

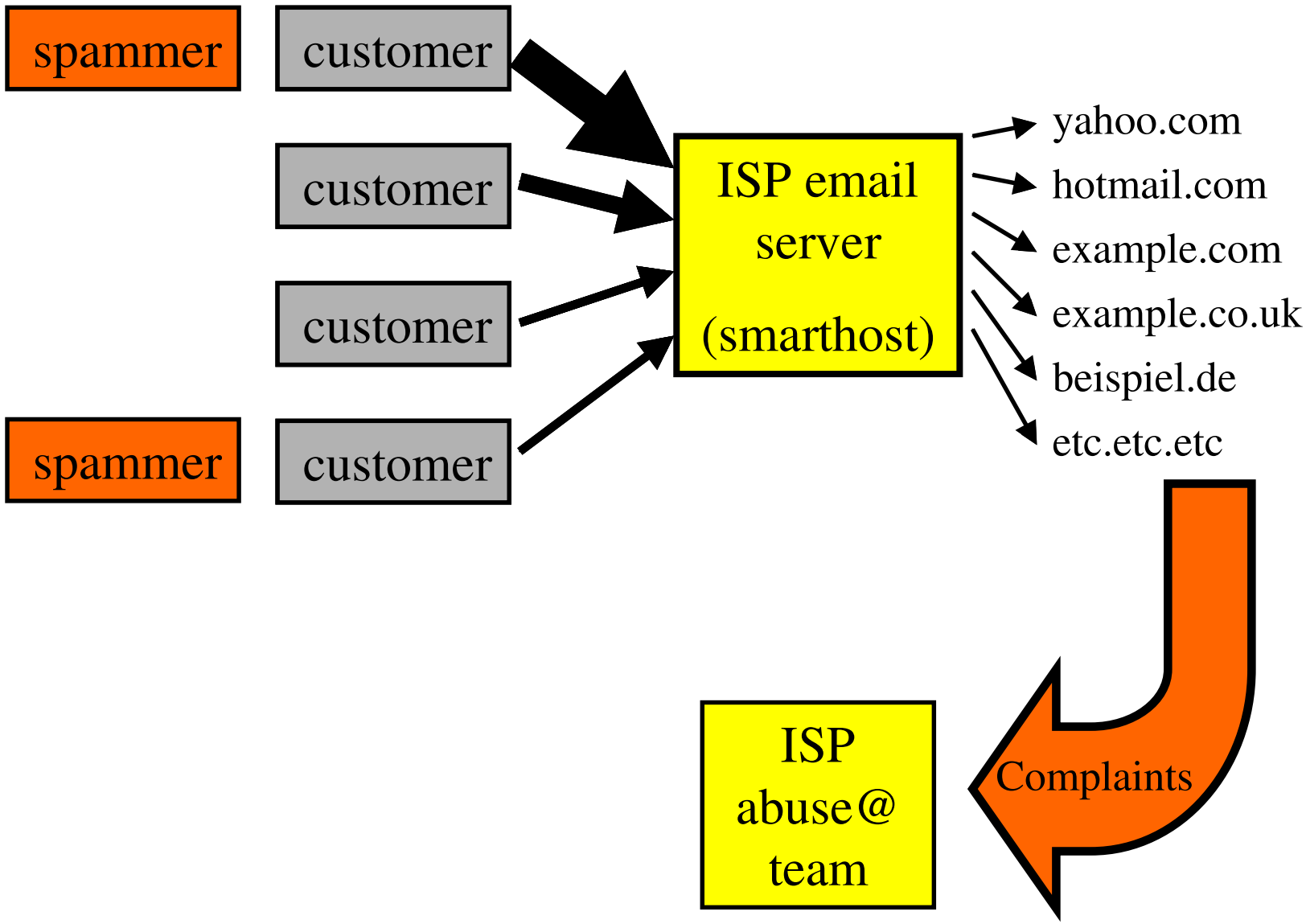
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Summary

- Log processing for customers
- Log processing for non-customers
- Looking at sampled sFlow data

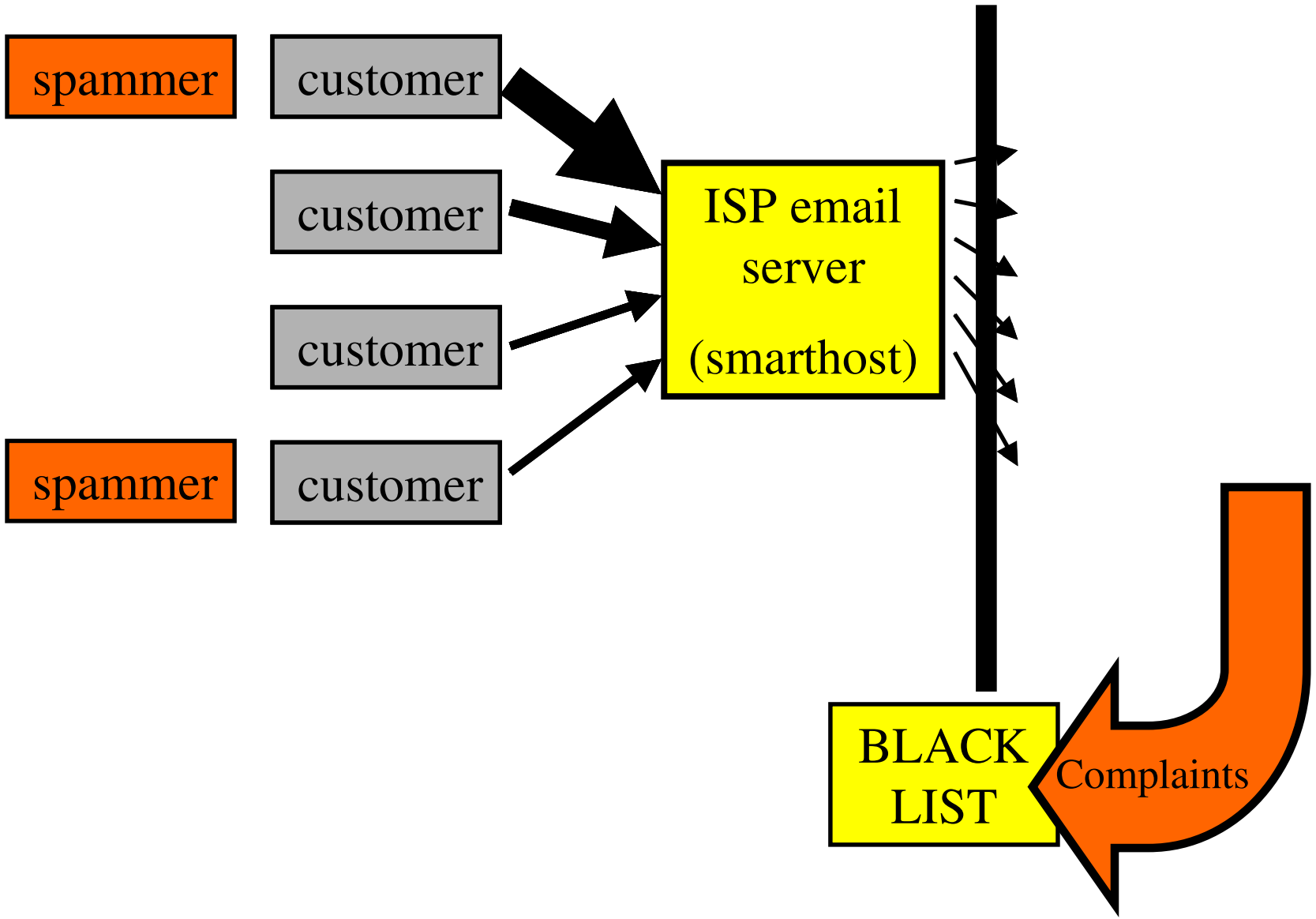
What problems do ISPs have?

- Insecure customers
  - very few real spammers sending directly !
- Botnets
  - compromised end-user machines
- SOCKS proxies &c
  - misconfiguration
- SMTP AUTH
  - Exchange “admin” accounts + *many others*


spammer
customer
customer
customer
spammer
customer
ISP email server (smarthost)
yahoo.com
hotmail.com
example.com
example.co.uk
beispiel.de
etc.etc.etc
ISP abuse@ team
Complaints

ISP's **R**eal **P**roblem

- Blacklisting of IP ranges & smarthosts
  - **`listme@listme.dsbl.org`**
- Rapid action necessary to ensure continued service to all other customers
- But reports may go to the blacklist and not to the ISP (or will lack essential details)


spammer
customer
customer
customer
spammer
customer
ISP email server (smarthost)
BLACK LIST
Complaints

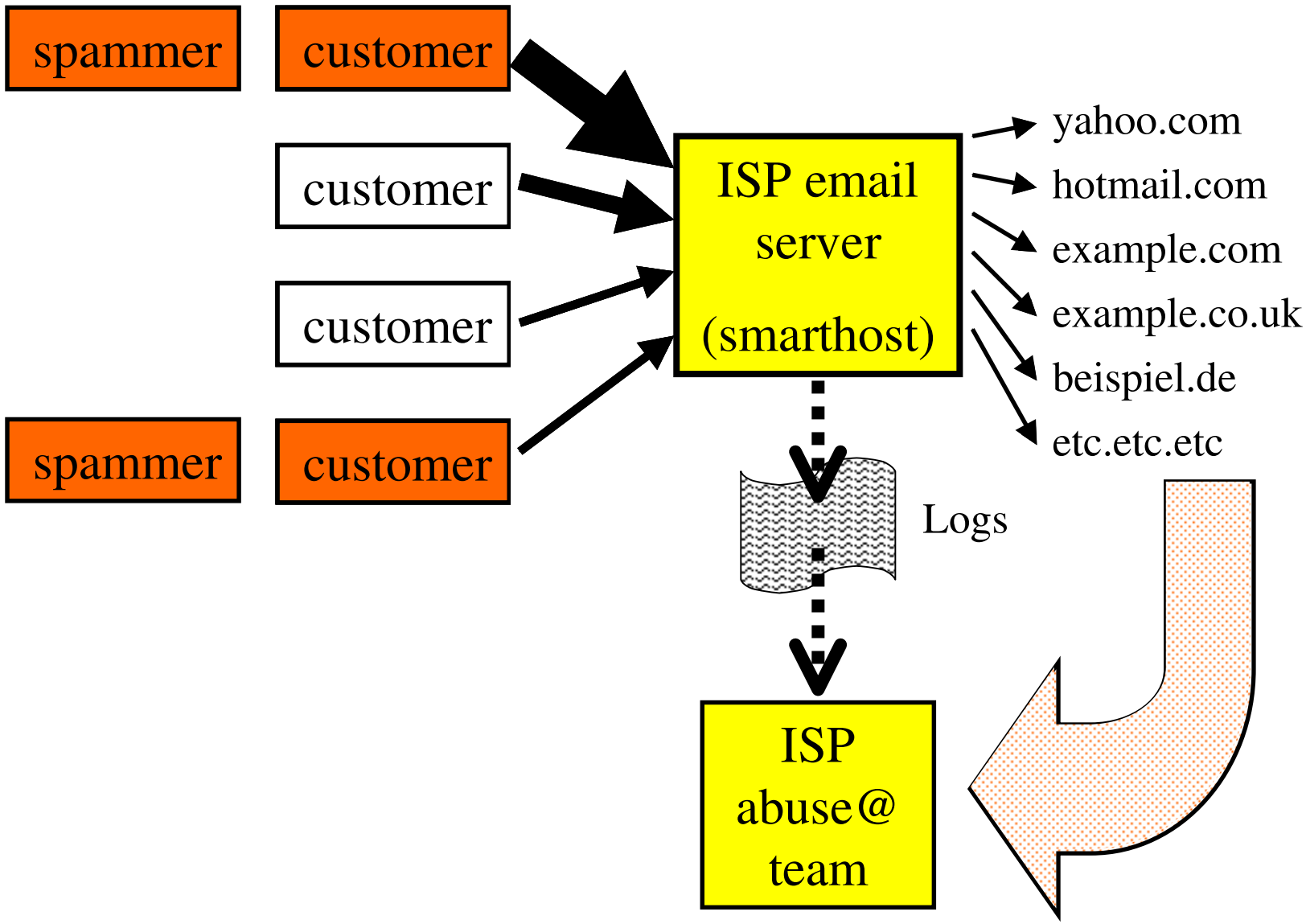
Spotting outgoing spam

- Expensive to examine outgoing content
- Legal/contractual issues with blocking
  - "false positives" could cost you customers
- Volume is not a good indicator of spam
  - many customers with occasional mailshots
  - daily limits only suitable for consumers
- "Incorrect" sender doesn't indicate spam
  - many customers with multiple domains

Key insight

- Lots of spam is to ancient email addresses
- Lots of spam is to invented addresses
- Lots of spam is blocked by remote filters

- Can process server logs to pick out this information. Spam has delivery failures whereas legitimate email mainly works


spammer
customer
customer
customer
spammer
customer
ISP email server (smarthost)
yahoo.com
hotmail.com
example.com
example.co.uk
beispiel.de
etc.etc.etc
Logs
ISP abuse@ team

Log processing heuristics

- **Report “too many” failures to deliver**
  - more than 20 works pretty well
- Ignore “bounces” !
  - have null “< >” return path, these often fail
  - detect rejection daemons without < > paths
- Ignore “mailing lists”
  - most destinations work, only some fail (10%)
  - more than one mailing list is a spam indicator!

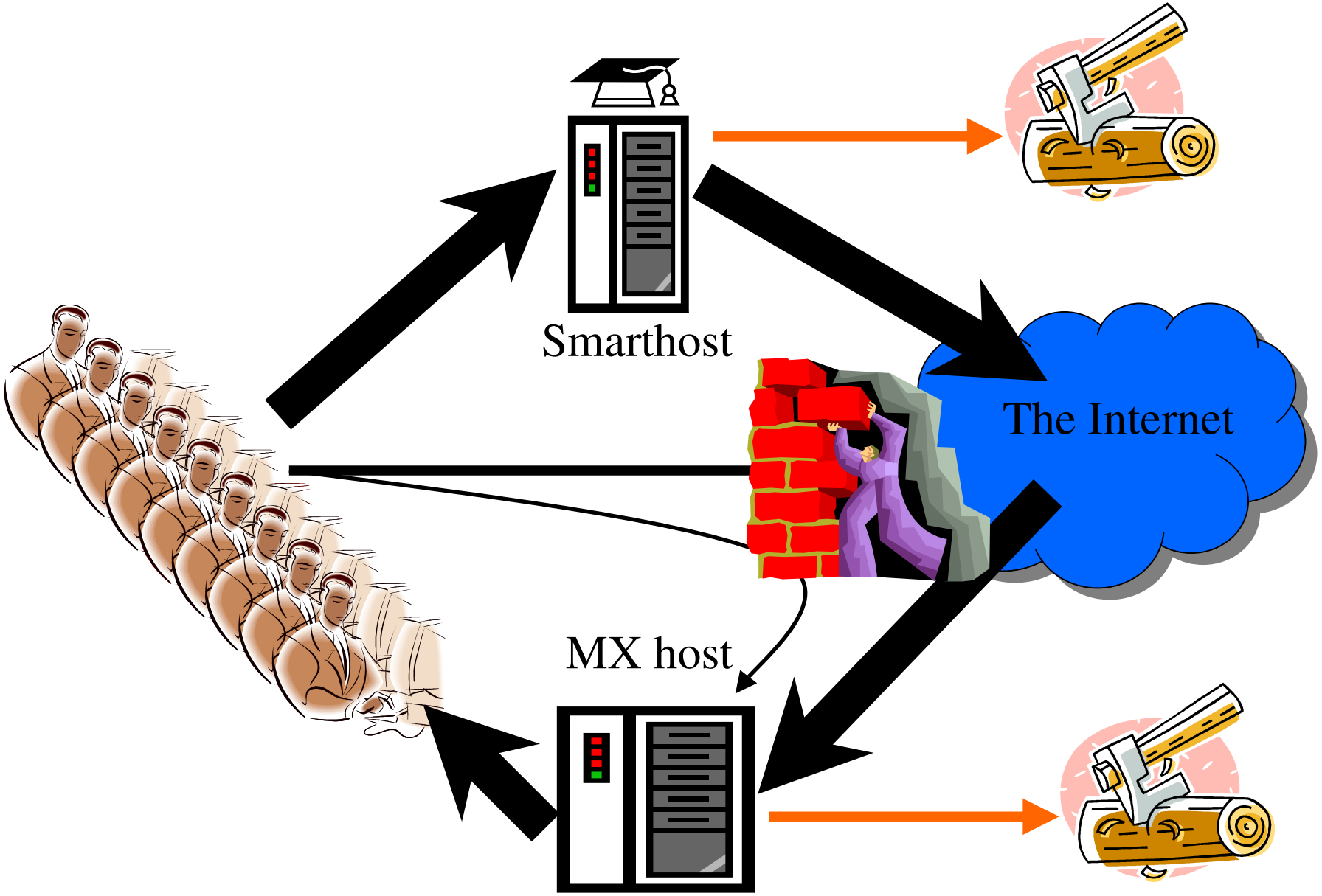
Bonus! also detects viruses

- Common for mass mailing “worms” to use address book (mainly valid addresses)
- But remote sites may reject malware

AND VERY USEFUL!

- Virus authors don’t know how to say HELO
- **So virus infections are also detected**

ISP email handling



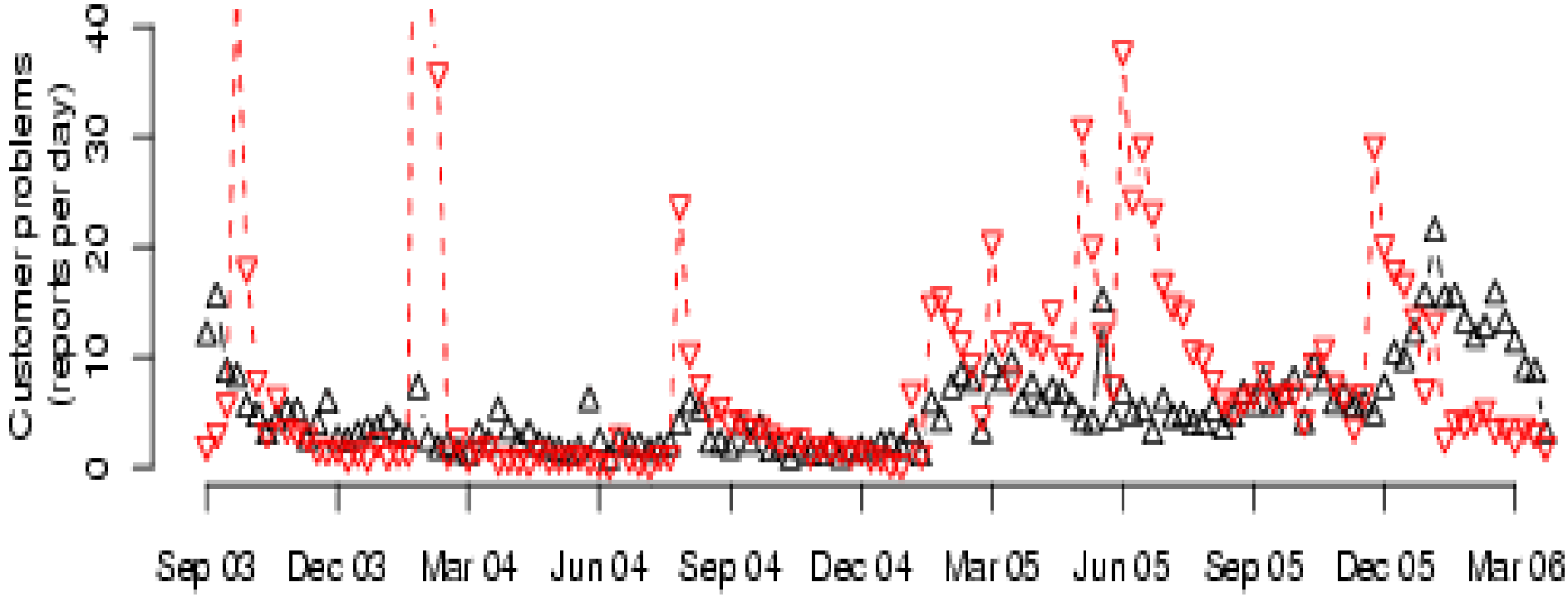
Heuristics for incoming email

- Simple heuristics on failures work really well
  - just as for smarthost
- Multiple HELO lines very common
  - often match MAIL FROM (to mislead)
  - may match RCPT TO (? authenticator ?)
- Look for outgoing email to the Internet
- Pay attention to spam filter results
  - but need to discount forwarding

```
2007-05-19 10:47:15 vzjwcqk0n@msa.hinet.net          Size=2199
                    !!! 0930456496@yahoo.com
                    !!! 09365874588@fdf.sdfads
                    !!! 0939155631@yahoo.com.yw
                     -> 0931244221@fetnet.net
                     -> 0932132625@pchome.com.tw
2007-05-19 10:50:22 985eubg@msa.hinet.net            Size=2206
                    !!! cy-i88222@ms.cy.edw.tw
                    !!! cynthia0421@1111.com.tw
                     -> cy.tung@msa.hinet.net
                     -> cy3219@hotmail.com
                     -> cy_chiang@hotmail.com
                     -> cyc.aa508@msa.hinet.net
                   and 31 more valid destinations
2007-05-19 10:59:15 4uzdcr@msa.hinet.net             Size=2228
                    !!! peter@syzygia.com.tw
                     -> peter.y@seed.net.tw
                     -> peter.zr.kuo@foxconn.com
                     -> peter548@ms37.hinet.net
                     -> peter62514@yahoo.com.tw
                     -> peter740916@yahoo.com.tw
                     and 44 more valid destinations
```

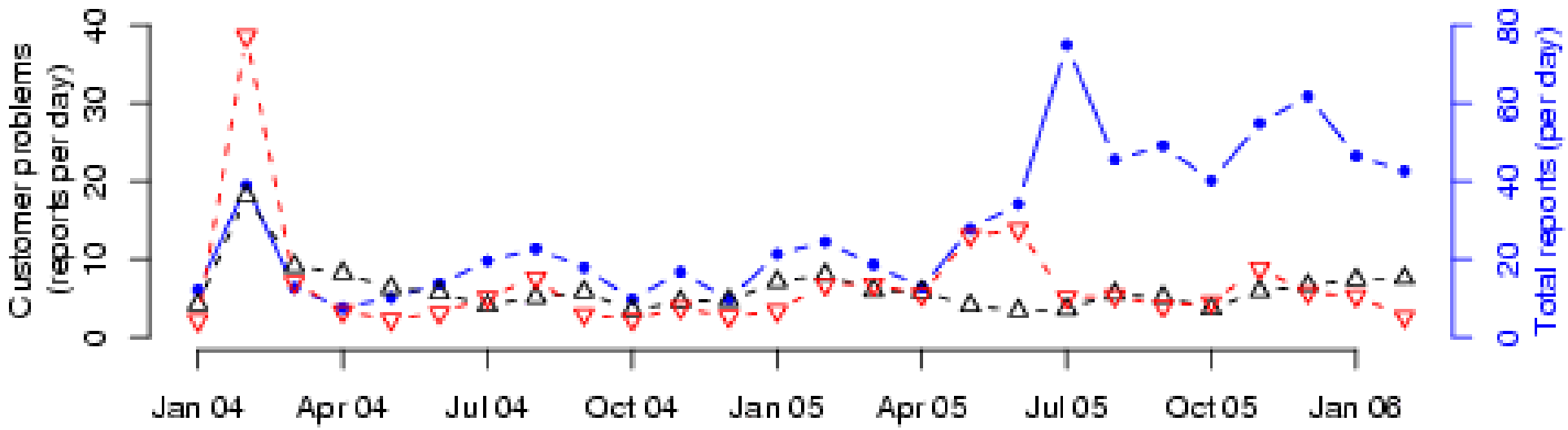
```
HELO = lrhnow.usa.net

2007-05-19 23:11:22 kwntefsqhi@usa.net                  Size= 8339
                      -> ken@example1.demon.co.uk

HELO = lkrw.hotmail.com

2007-05-19 23:11:24 zmjkuzzs@hotmail.com                Size=11340
                      -> ken@example2.demon.co.uk

HELO = pshw.netscape.net

2007-05-19 23:14:52 dscceljzmy@netscape.net             Size= 6122
                      -> steve.xf@example3.demon.co.uk

HELO = zmgp.cs.com

2007-05-19 23:18:06 wmqjympdr@cs.com                    Size= 6925
                      -> kroll@example4.demon.co.uk
```

Email log processing @ demon



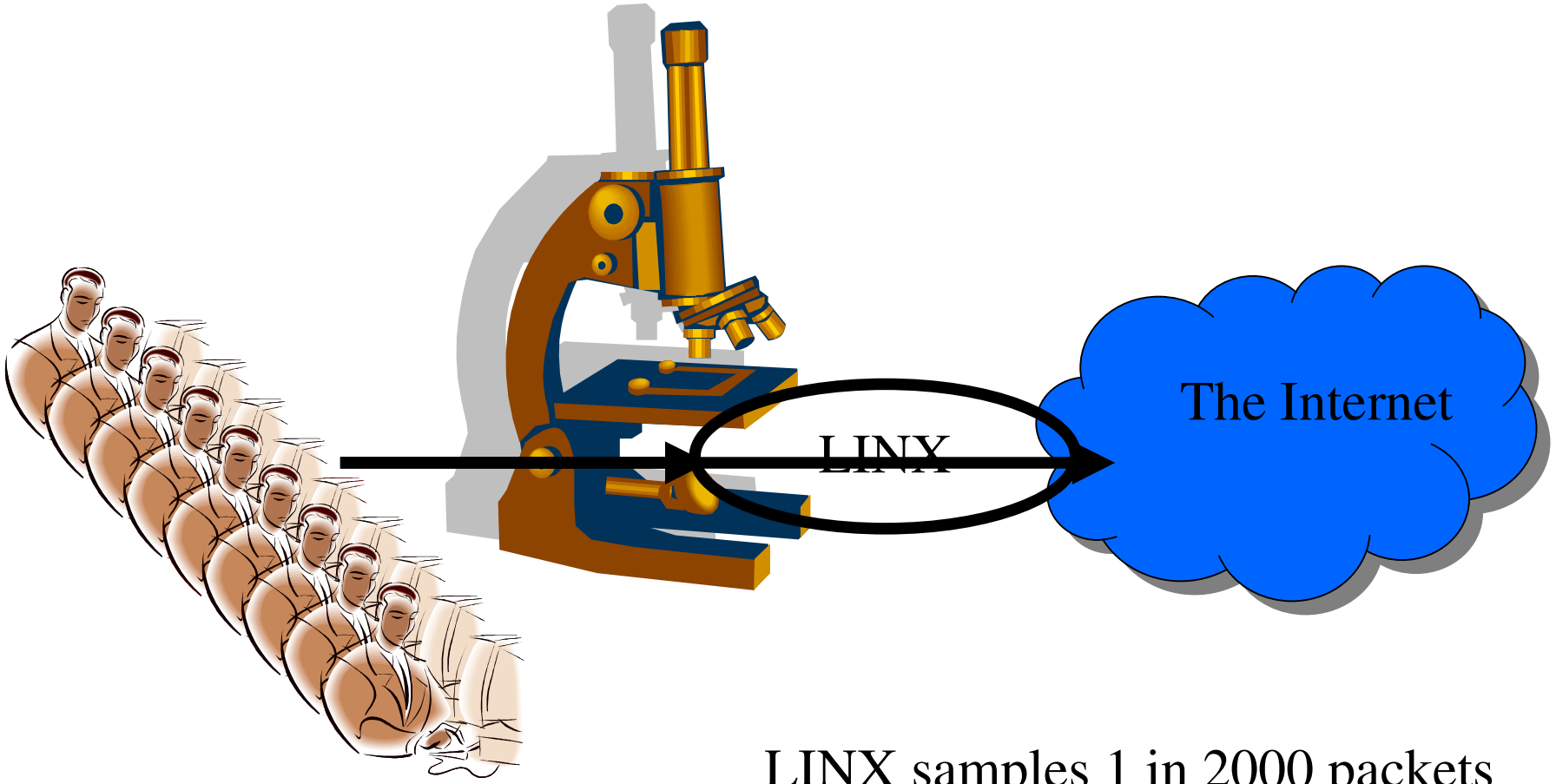
Detection of spam (black) and viruses (red)

Incoming reports (all sources)



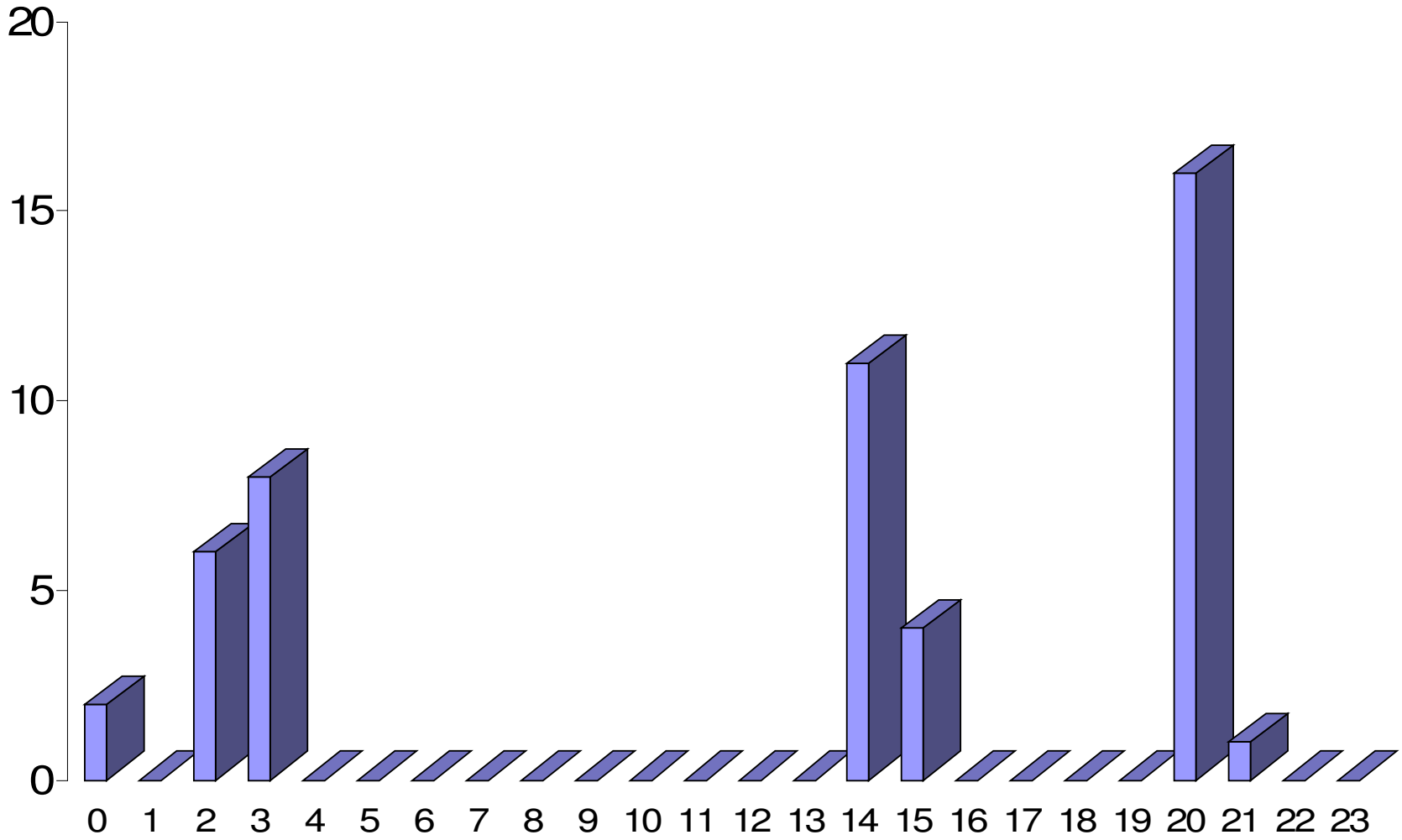
spam (black), viruses (red), reports (blue)

spamHINTS research project

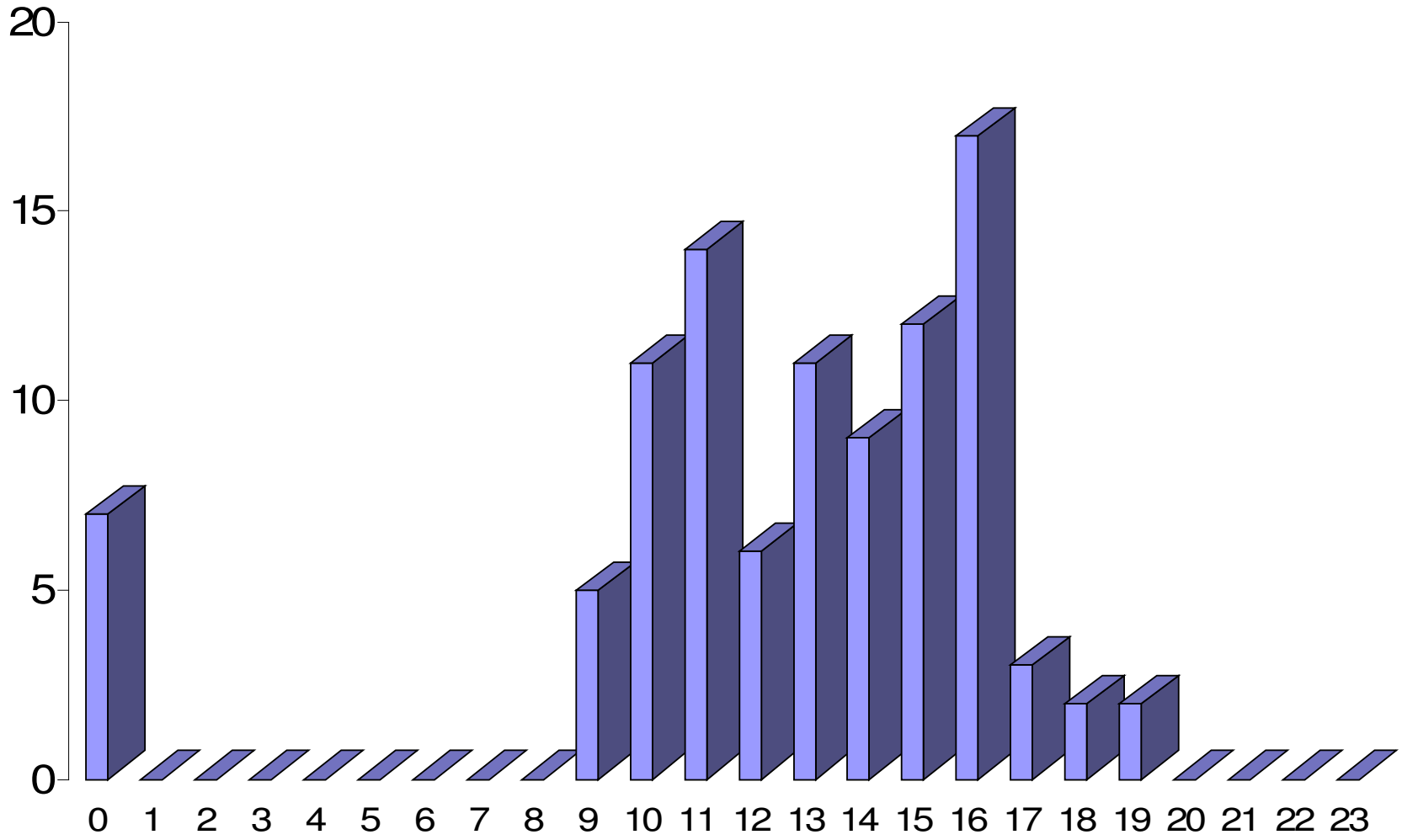


LINX samples 1 in 2000 packets (using sFlow) and makes the port 25 traffic available for analysis…

Known “open server”

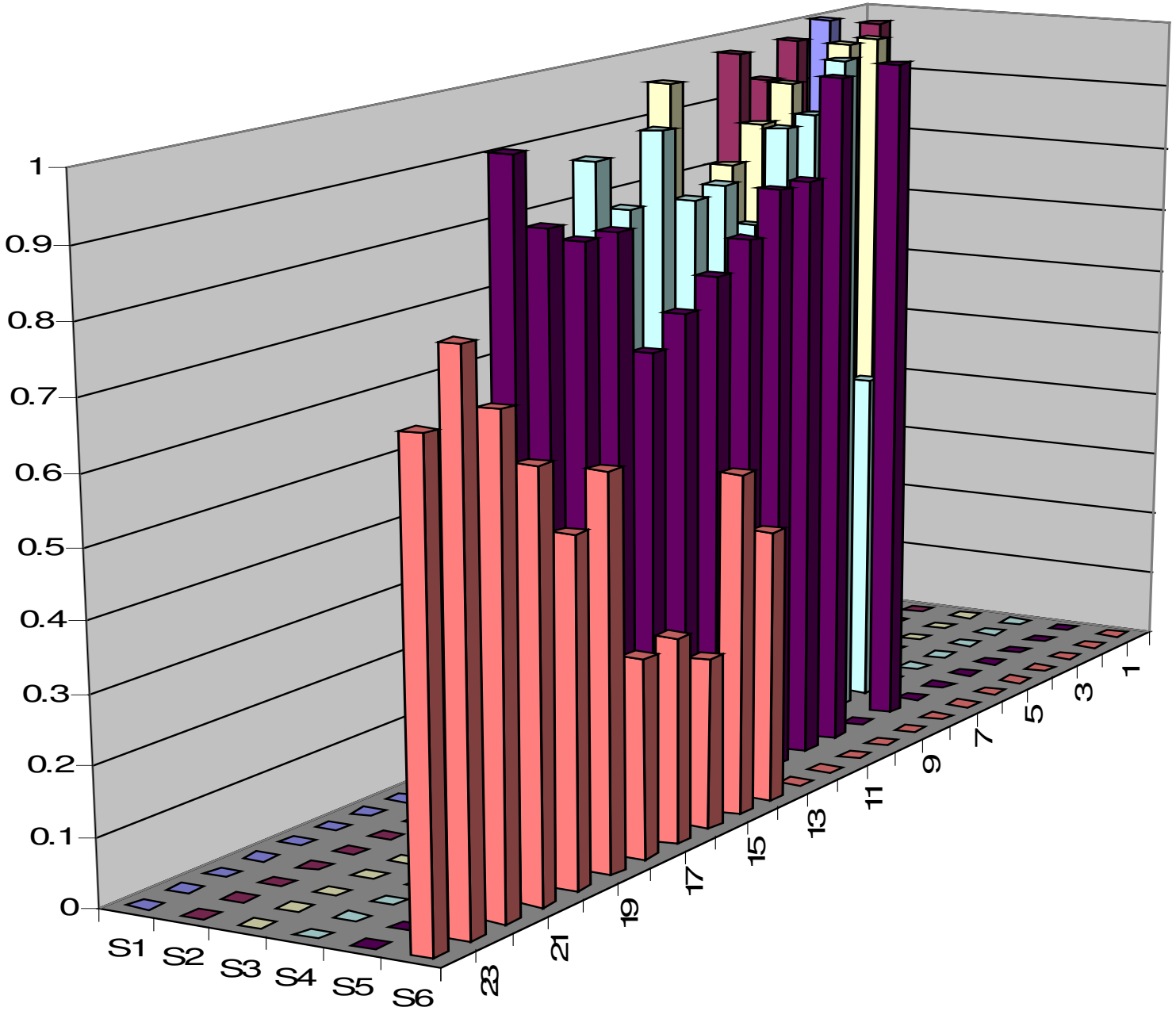


Another known “open server”

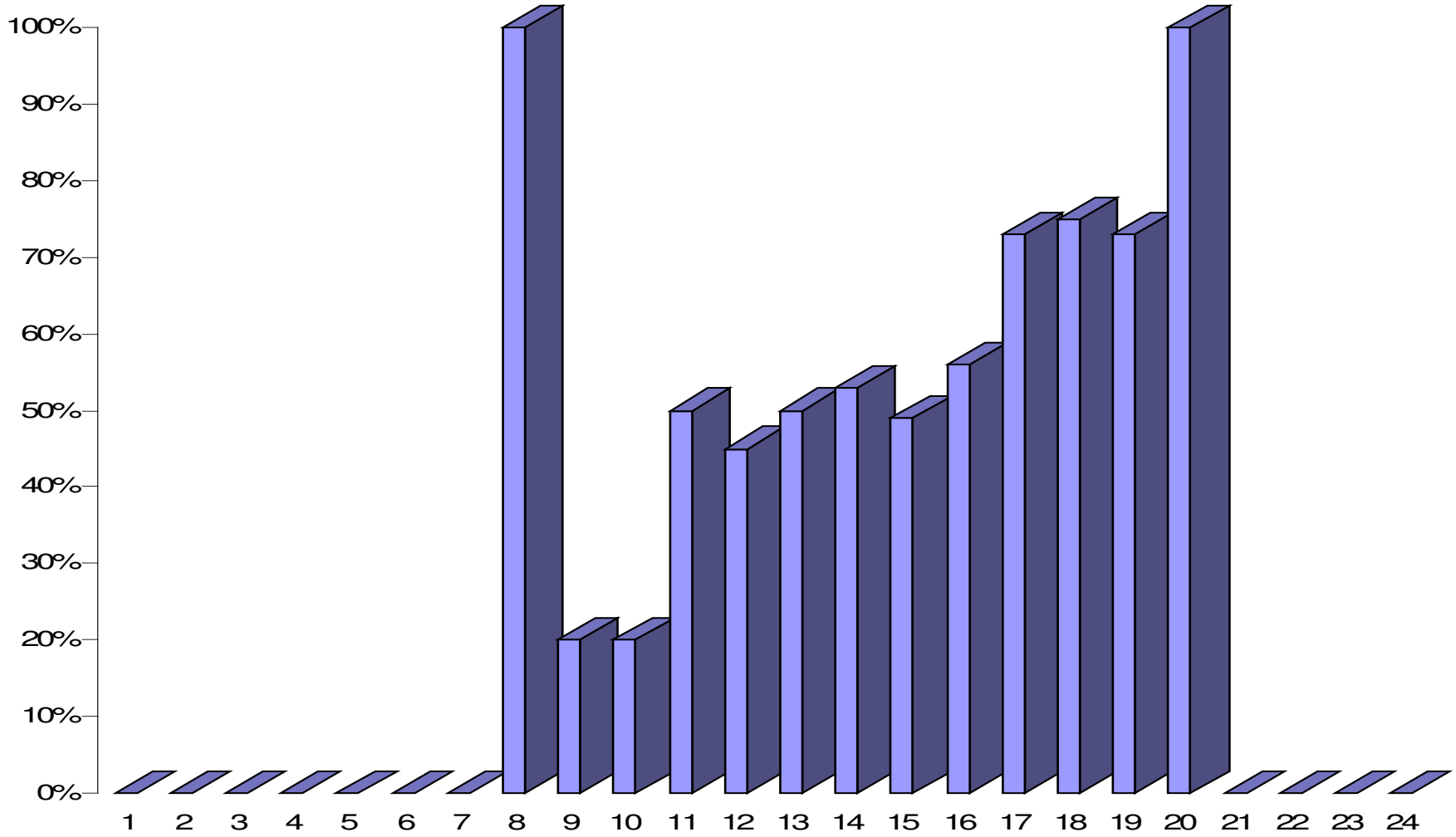


Look for excessive variation

- Look at number of hours active compared with number of four hour blocks active
- Use incoming email to Demon to pick out senders of spam and hence annotate them as good or bad…
- … did this for a large ISP, but problem is that “if it sends, it’s bad”. Nevertheless…


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
S1
S2
S3
S4
S5
S6
23
21
19
17
15
13
11
9
7
5
3
1

Spamminess vs hours of activity
for IPs active in 5 of 6 possible 4 hour periods



So work continues…

- sFlow data will always be useful to feed back ongoing activity to abuse teams
- Analysis may improve when both rings instrumented and when data available in real-time (so can compare historic data)
- Still to consider variations (and lack of variations) in destination as well as time

Summary

- Processing outgoing server logs **works well**
  - keeps smarthosts out of blacklists
- Processing incoming server logs **effective**
  - some sites may see little “looped back” traffic
- **Trying** to processing sampled sFlow data
  - sampling is making it a real challenge
  - more work needed on good distinguishers

http://www.cl.cam.ac.uk/~rnc1

CEAS papers: http://www.ceas.cc

2004: Stopping spam by extrusion detection
2005: Examining incoming server logs
2006: Early results from spamHINTS
2007: Email traffic: A qualitative snapshot

UNIVERSITY OF CAMBRIDGE
Computer Laboratory