

Detecting email spam in sampled traffic data from LINX

Richard Clayton

EU Spam Symposium, Vienna, 24th May 2007

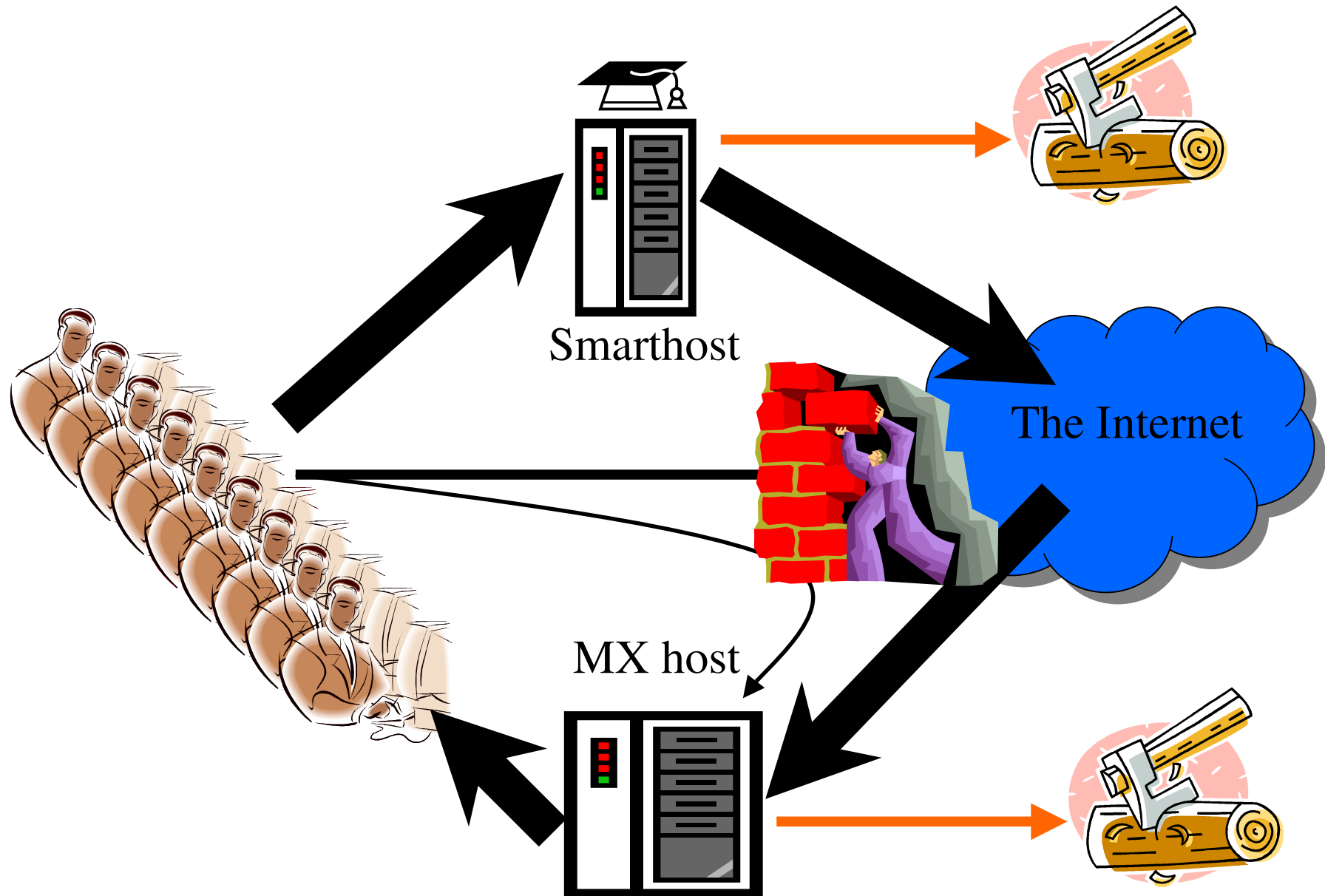


**UNIVERSITY OF
CAMBRIDGE**
Computer Laboratory



Demon

ISP email handling



Heuristics for log processing

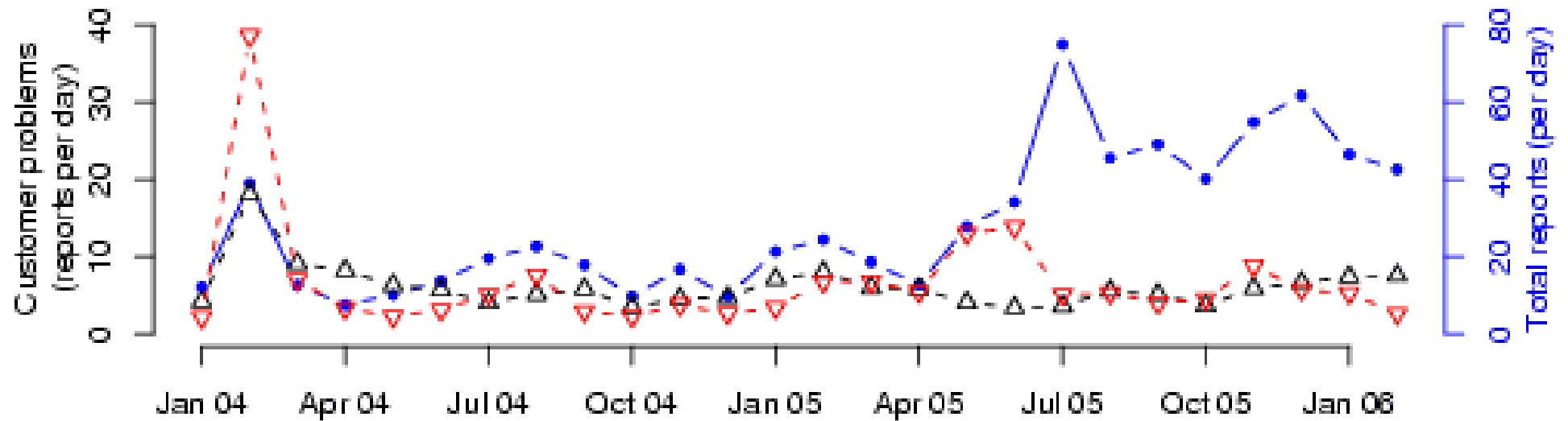
- Simple heuristics work really well
- Key measure is failures to deliver
 - addresses are old/constructed/blocked
- Multiple HELO lines very common in spam
- Look for outgoing email to the Internet
- Pay attention to spam filter results
 - but need to discount forwarding

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

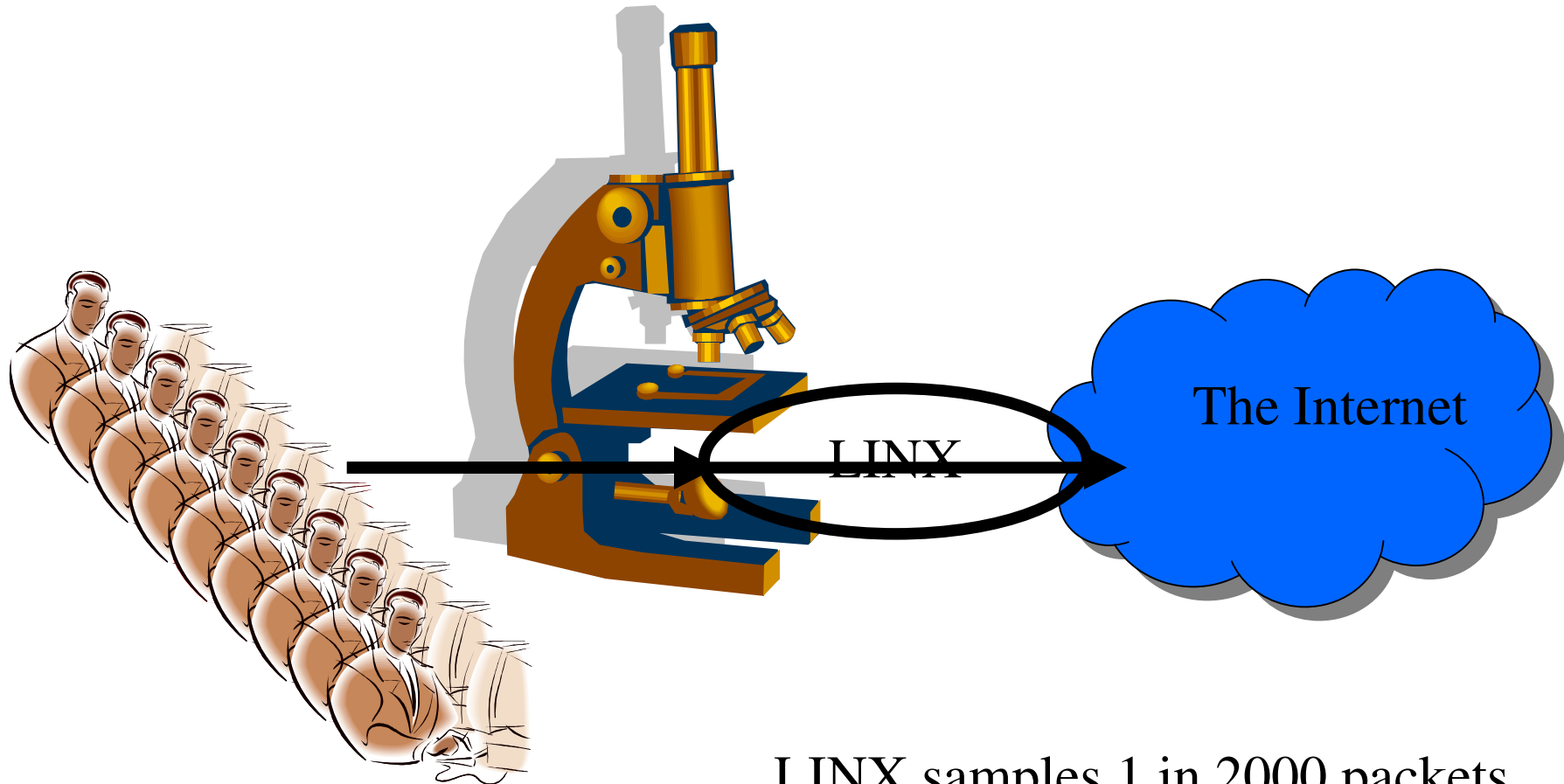
2007-05-19 10:59:15 4uzdcr@msa.hinet.net Size=2228
!!! peter@syzygia.com.tw
-> peter.y@seed.net.tw
-> peter.zr.kuo@foxconn.com
-> peter548@ms37.hinet.net
-> peter62514@yahoo.com.tw
-> peter740916@yahoo.com.tw
and 44 more valid destinations

Incoming reports (all sources)



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

spamHINTS research project

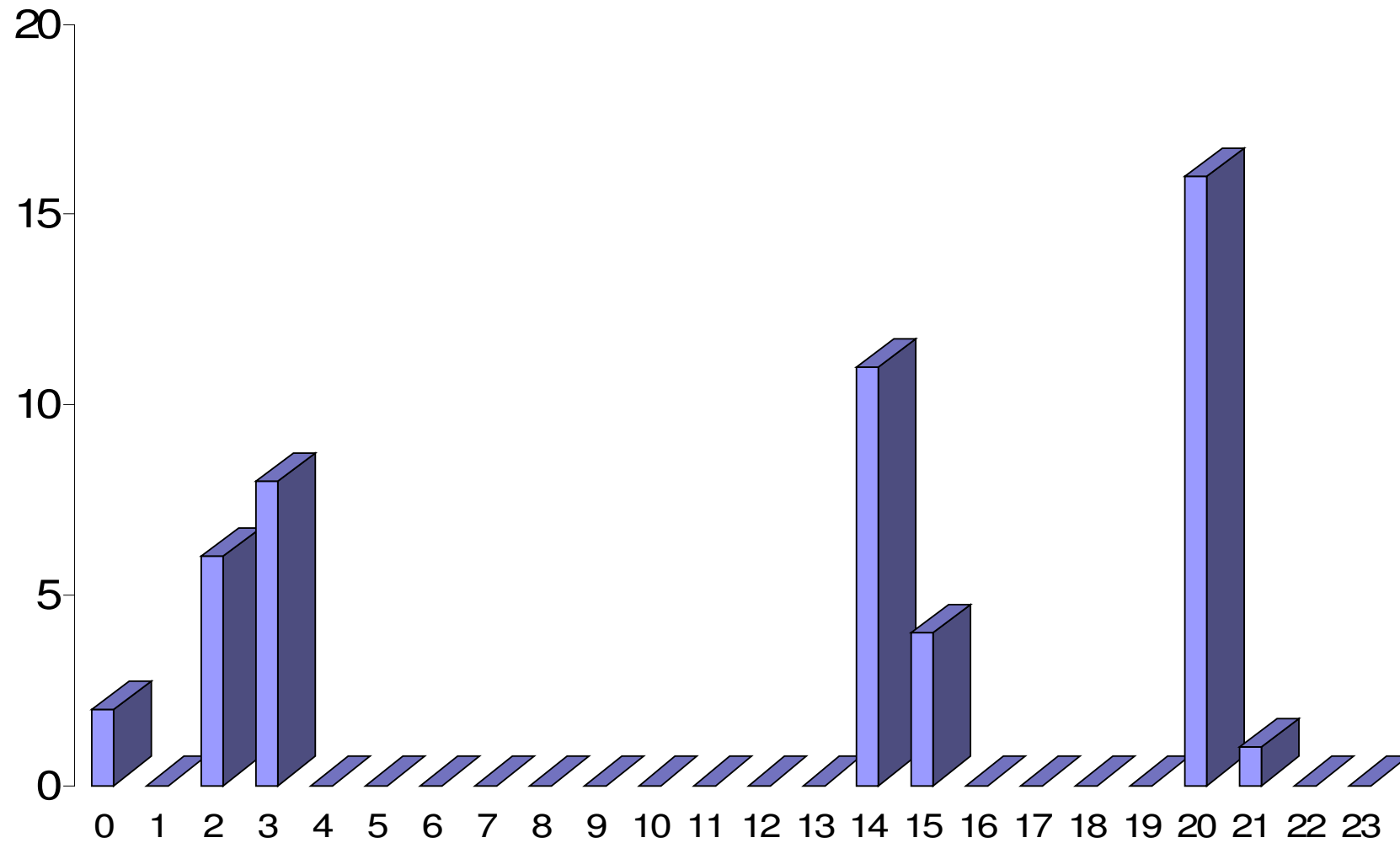


LINX samples 1 in 2000 packets
(using sFlow) and makes the port 25
traffic available for analysis...

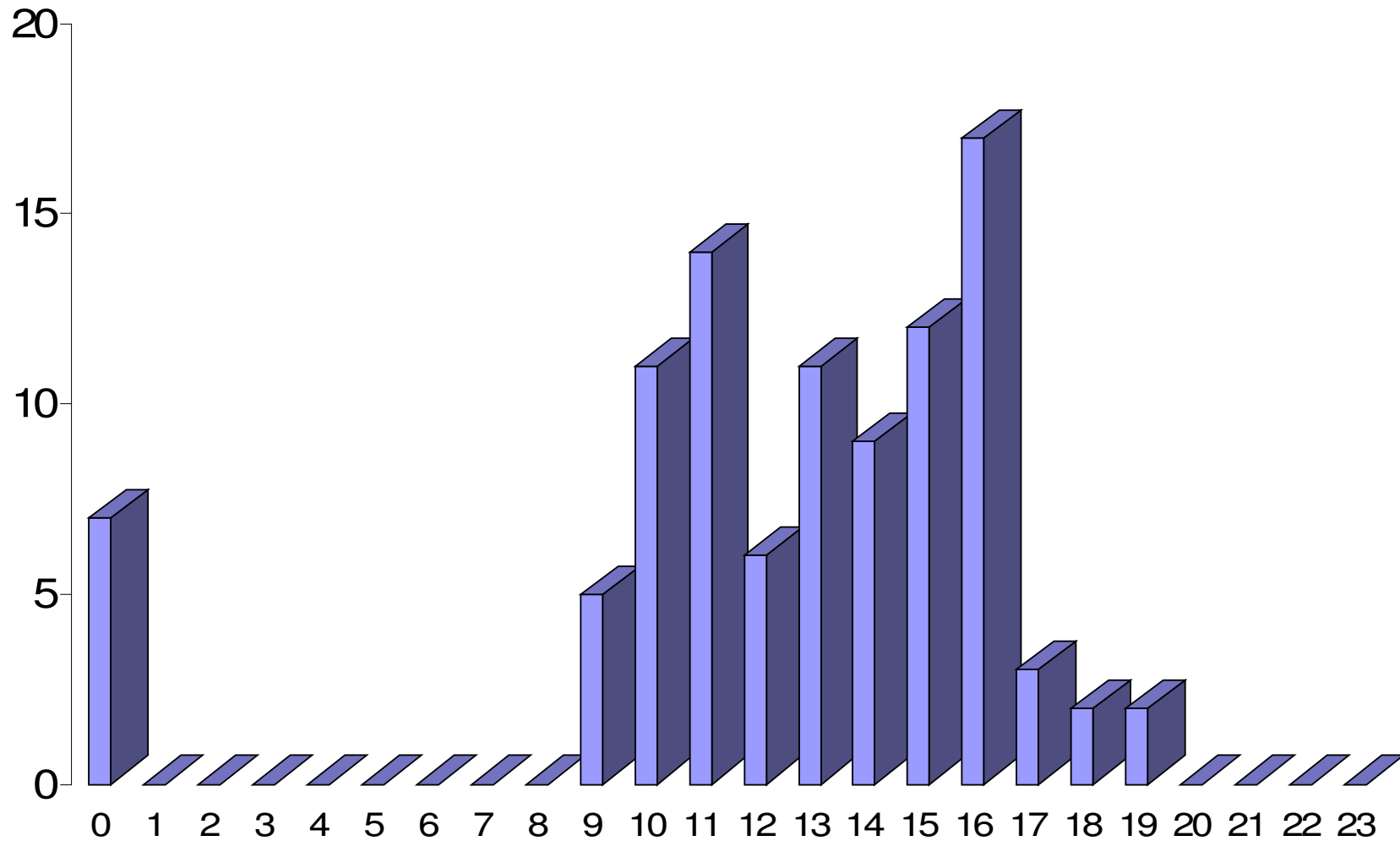
Basic idea

- Spam doesn't look like normal email, so by analysing the traffic patterns it will be detectable
- Big benefits if this can be shown to work, only evasion technique is to look more like normal email and send less
- Running this at the LINX permits amortisation of costs (and development) across the whole industry

Known “open server”

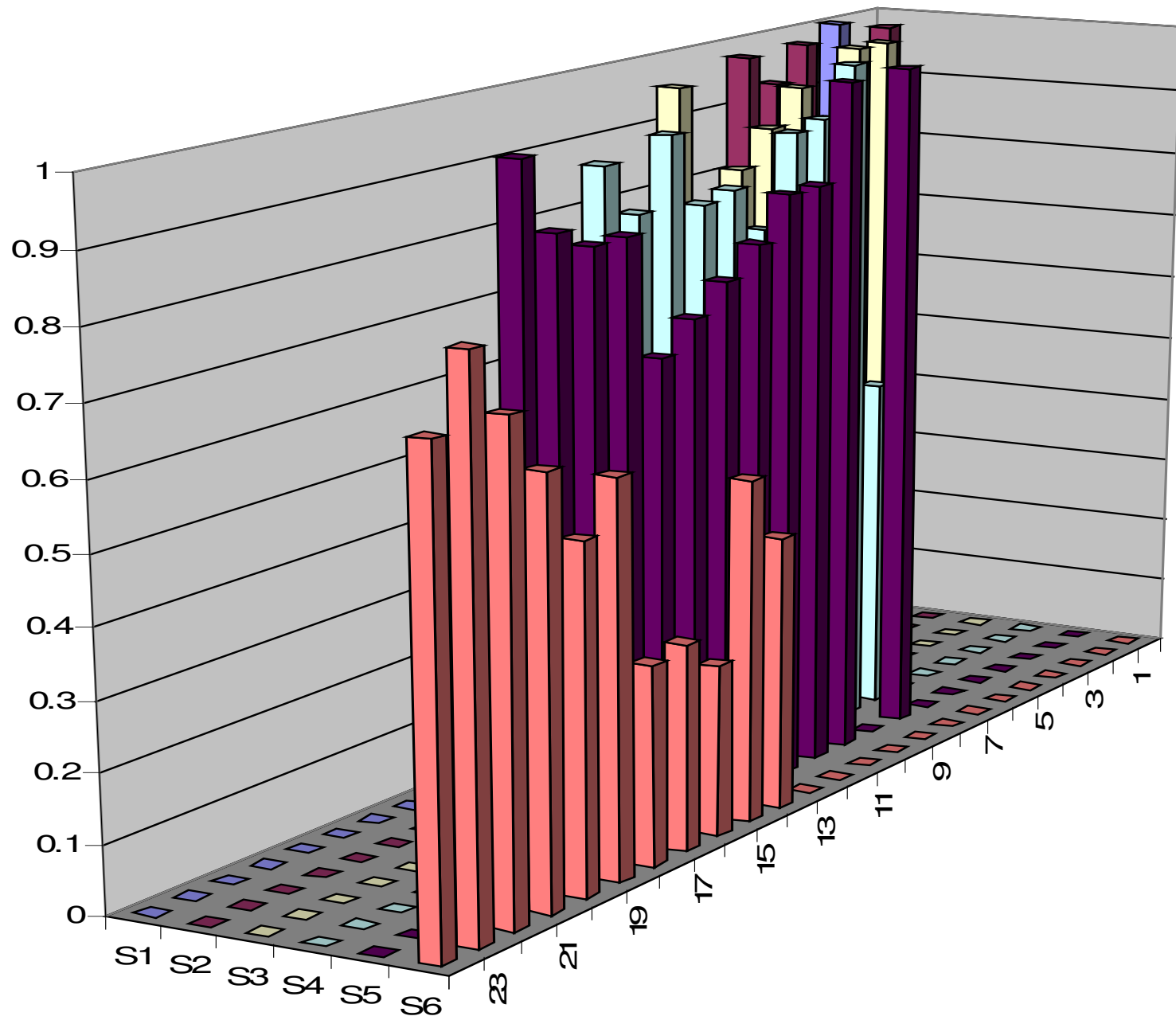


Another known “open server”

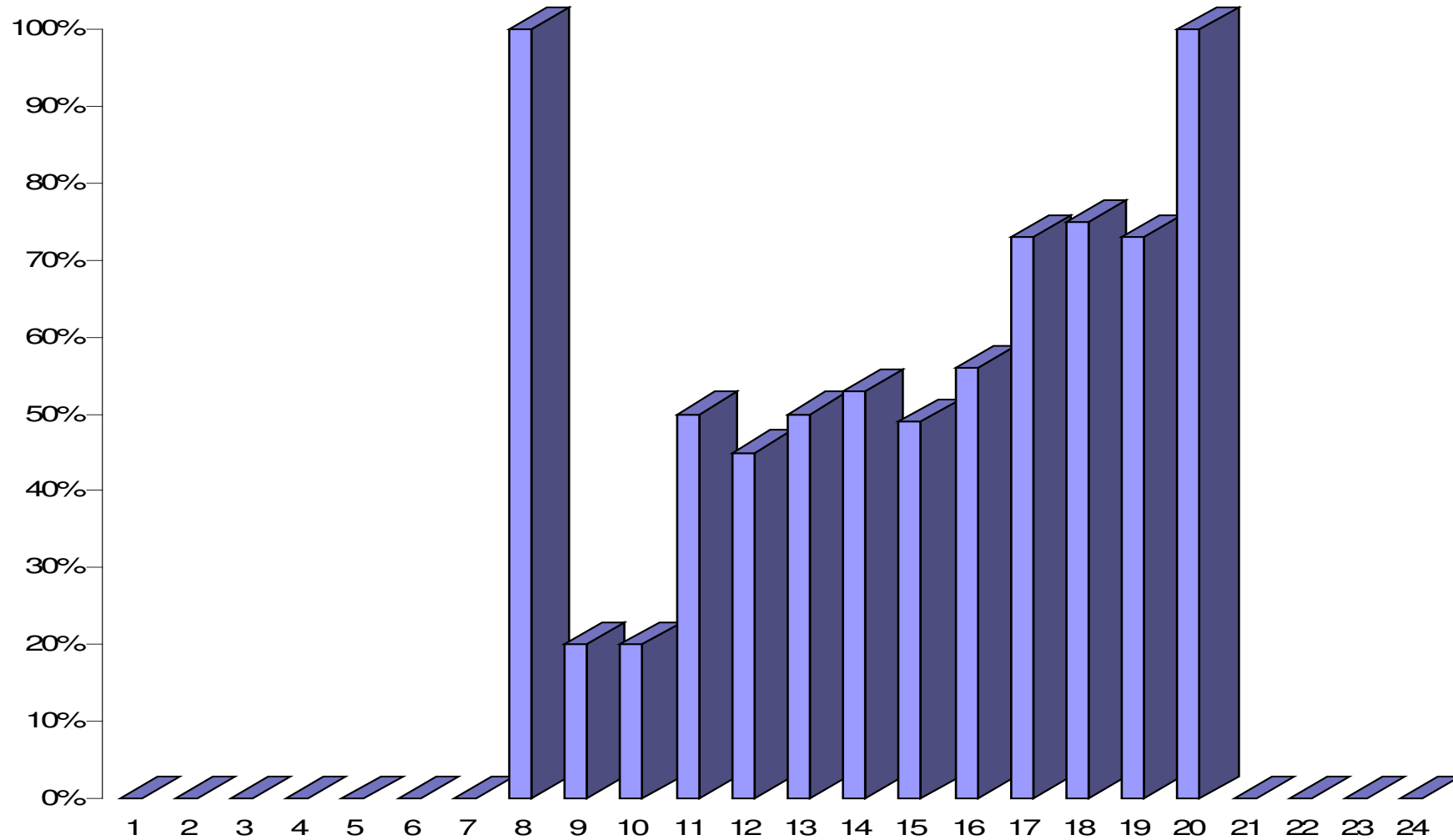


Look for excessive variation

- 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...



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



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



Demon