spamHINTS

PI: Prof. Ross Anderson
Researcher: Richard Clayton

UNIVERSITY OF CAMBRIDGE
Computer Laboratory
Happily It’s Not The Same

• The sending of spam differs from the sending of legitimate email, not just in content but in the traffic patterns
• *Time* email is “9 to 5”, spam is 24 hours
• *Space* spam goes to many destinations or all to just one ISP (in a “dictionary” attack)
• *Size* spam is a constant size
• Virus/Worm traffic is like spam (but bigger)
ISP email handling

Smarthost

MX host

The Internet
Mail server log processing

• Running for over two years at Demon
• Blacklisting of smarthost much reduced 😊
• Has led to cost savings 😊
• Next obvious step is data sharing…
  – Demon tells you about your problems
  – You tell Demon about theirs
  – And you tell each other….
• …because some (a lot ?) is being missed 😞
spamHINTS research project
Challenges

• sFlow data is just packet headers
• No information about content of emails, and unlike the server log processing, no SMTP protocol content either
• LINX traffic flow rates very high (100Gb/s)
• sFlow data is sampled (1 in 8192 at present) so size estimation is complicated
Some good news

- LINX is major UK peering point, so likely to see traffic in two directions
- sFlow enabled on all switches so traffic likely to pass more than one sampling point
- Email only a small proportion of total traffic
- Spammers cannot move traffic to another port to avoid detection (always tcp/25)
Project plan

• Months 1-6
  – Data collation: start to collect & process data
  – Proselytize: create website & data access

• Months 6-9
  – Metrics: establish estimates of size & frequency

• Months 10-24
  – Heuristics: predict sources of spam and work with ISPs to validate accuracy
  – Proselytize: promote traffic analysis solution
Deliverables

• Months 1-6
  – List of sources of email
  – Website with appropriate access controls

• Months 6-9
  – Annotate lists with predicted size/frequency

• Months 10-24
  – Regular reports of sources of spam
  – Working code for deploying at other IXPs
Project support

**LINX** makes the sFlow data available plus infrastructure support

**Intel** providing about 50% of project funding

**NTL** “part and parcel” of long-term approach to their spam problem. Funding expected

**Other ISPs** talks ongoing with several

**DTI** may well chip in, provided industry does
Intellectual property

• Existing email log processing created for Demon Internet – but agreement to make it available under open source licence
• IP developed under project to be made available under open source licence (LINX will not contribute to project otherwise)
• Data to be made generally available (while meeting EU data protection requirements)
spamHINTS Summary

- Spam is not the same as legitimate email
- Processing traffic data at an Internet Exchange point is challenging but tractable
- Picking out spam traffic looks achievable
- Will provide ISPs with valuable data about their customers’ problems
What you can do!

• Approve use of LINX sFlow data
• £und the spamHINTS project!
  – you’ll get all the alpha/beta benefits ASAP!
• Start processing email server logs
• Share results of email server log processing
• Tell me (and thereby the DTI) that the Y1 deliverables will help your business
Richard Clayton
<rnc1@cl.cam.ac.uk>

www.spamhints.org