

# spamHINTS

## Happily, It's Not The Same

This is a proposal for a research project into the use of “traffic analysis” to detect email “spam”. It will not be looking at the content of the email, but will be looking for the distinctive patterns, in time, in destination and in volume that distinguish the bulk sending of spam from legitimate email activity.

The project will build upon and extend existing (and very successful) work on analysing traffic data from email server log files. In particular it will determine how *sFlow* data from LINX, an Internet Exchange point, can be used to pick out the tell-tale patterns of spam and feed this information back to the originating ISPs.

Funding is being sought from several organisations – each taking a share – for just under £90K/annum; with a two year project envisaged.

### **Introduction**

Email spam is not just a technical problem, but a market failure compounded by regulatory deficiencies. However, at present, it is technology that is giving us most of the countermeasures against spam, with filtering being key to keeping spam out of our inboxes and hiding the worst of the problem.

Two (rather contrary) trends in filtering can be distinguished: highly personalised Bayesian filtering, and corporate out-sourcing of filtering decisions to specialist companies. It must be expected that both techniques will become far less effective over time, as spammers perfect methods of making spam look just like “real” email.

However, spam can never be quite “the same” as real email – and the area in which it will be hardest for the spammers to blend in will be in their communication patterns. They are doing things which normal senders of email do not do. They are sending in bulk with few replies coming back; they are sending 24 hours a day; and, they are sending with limited discrimination in their targets. Since they are not doing the same thing as legitimate senders, their activity can now be (and will continue to be) distinguishable by means of traffic analysis that picks out the distinctive patterns that spamming exhibits.

Traffic analysis is not a technique that end users can easily apply, but it is ideal for ISPs who can extract patterns from bulk traffic. It complements filtering, rather than replacing it, and – with its potential for detecting the spam whilst it is still flowing within the ISP's networks – it is a good fit to their abilities and to their problems.

Spam is being increasingly relayed via compromised end-user machines, sometimes exploiting a configuration error, but usually because the user has been tricked into running some malware delivered via an email virus or worm. The hosting ISP is held to be responsible for its users' behaviour, but the spam is often sent directly to its destination, bypassing the ISP's email systems. Although some ISPs are prepared to

block direct sending, particularly for consumers, it is unattractive to many ISPs, especially those selling services to businesses – who commonly wish to run their own email systems.

At present ISPs rely on reports from the recipients of spam to learn of problems, but these reports are increasingly rare, perhaps only one recipient in ten thousand will report spam. The reports are also expensive to process and ISPs typically end up employing two to four “abuse team” members per 100,000 customers – because it is even more expensive **not** to process the reports: the ISP's reputation suffers and they can end up in the position that none of their customers' email, even the legitimate traffic, is accepted by remote sites.

What ISPs need is timely, accurate, easy to process reports that set out which of their customers are, usually unknowingly, sending spam. This project proposes to make such reports available.

### **Picking out spam in email server logs**

Traffic analysis has already been shown to be very effective in detecting spam.

Many customers send their email via the ISP's “smarthost” email server. Processing the activity logs from this server permits ISPs to pick out customer machines that have been hijacked by spammers [refn 7]. As a bonus, the techniques also highlight email viruses and can assist in detecting resource-sapping mail-forwarding loops. In the longer term, the spammers and virus writers will succeed in hiding the most tell-tale patterns, but since what they are doing is fundamentally not the same as legitimate activity, traffic analysis must be expected to remain highly effective.

However, the email that is sent “direct” to its destination does not show up in the smarthost logs (for the obvious reason that it doesn't use those machines). The email will, of course, show up in the logs at the destination; where spam-like patterns can be detected. Some good results have already been obtained from traffic analysis of recipient logs for the special case where an ISP customer is sending email to other customers of the same ISP [refn 2].

There is currently very little inter-ISP co-operation in detecting and reporting spam. This project will, as an adjunct to its main activities, seek to improve this co-operation by, for example, establishing suitable legal frameworks for data sharing and documenting appropriate points of contact.

Unfortunately, even with increased co-operation, there is an important limitation to processing the recipient server logs. Spammers could avoid sending sufficient traffic to each destination for the distinctive patterns to be detected with any certainty. What is needed is a holistic approach that examines **all** outgoing email traffic...

### **Picking out spam in sFlow data**

The LINX, the premier Internet Exchange Point (IXP) in the UK is where almost all UK ISPs interconnect so as to pass Internet traffic to each other's networks. The LINX is currently enabling its IXP peering infrastructure to produce sFlow data, which is

packet header information about a statistically sampled subset of the packets flowing through the IXP switches.

The sFlow data does not contain any packet contents, but, because email uses a fixed port (tcp/25) it can be used to determine which machines are sending email and to give some idea as to how often. The sampling (only 1 in 8192 packets is recorded) creates significant challenges for transforming this data into useful metrics that can form the basis of reports to ISPs about problems; which is one of the reasons this is a “research” proposal and not mere “engineering”.

The title of the proposal “HINTS” comes from “Happily It’s Not The Same”. It is very seldom possible to make one type of traffic a perfect replica of another, and so the traffic type can almost always be determined without recourse to examining the content. Put more formally: the research hypothesis is that it will be possible to develop efficient heuristics for using sampled sFlow traffic data to distinguish email spam from legitimate email traffic.

For example, spam will be distinguished by having markedly different distributions:

- in time – it is not sent by humans who sleep, eat, and take the weekend off;
- in space – legitimate traffic flows from the UK to Taiwan or Korea do exist, but are uncommon; whereas spam is often relayed via machines on the other side of the world in the expectation that complaints will be less vocal than if nearby machines are used instead; and
- in volume and size – spam flows tend to consist of a great many short messages, whereas legitimate email is a mixture of sizes and is usually sent in quite small amounts.

## **Research plan**

We propose a two-year research project based within the Security Group of the University of Cambridge Computer Laboratory to identify spam sources by using traffic analysis of sFlow data.

### **Phase One: basic data collation: months 1 to 6**

In the initial phase the systems to extract detailed information about email traffic from the LINX sFlow data will be developed. The sampled nature of the traffic must be addressed; ensuring, for example, that where luck provides more than one packet for a single email conversation the two reports are collated. Looking at email traffic does have some advantages over general traffic inspection in that the spam traffic being sought is constrained to use the standard SMTP port (because the legitimate servers keep this fixed). Also, email traffic is only a small percentage of all traffic; hence the datasets involved, although large, will not be overwhelming.

The main deliverable at the end of this phase is a real-time list of email sources. Although no distinction is made between spam and non-spam, the data is still of significant value. ISPs may be in a position to know who is “authorized” to send email (or may wish to ask their customer if they know that they are sending email today for the first time). ISPs may have contacted their customer to advise them that

they have a problem, and will wish to know if it has been corrected. This data will allow them to check. Similarly, organisations creating lists of compromised machines, such as SpamHaus, and some industry anti-zombie initiatives, will be able to correlate their own lists against this data.

### **Phase Two: data metrics: months 7 to 9**

The accuracy of the system in assessing the size and frequency of email sending (that is the validity of the scaling factors used to allow for the sampled nature of the sFlow data) can be measured by comparison with ISP email server statistics, i.e. by comparing the project results with known traffic volumes from ISPs who are prepared to share this data (in particular, from project partners).

At the end of this phase of the project the real-time list of senders can be annotated with reasonably accurate estimates of volume. This can be used by ISPs to prioritise their work, and for simple heuristics such as spotting significant changes from historical traffic levels (e.g. customers sending ten times as much email as last week).

### **Phase Three: spam-detection heuristics: months 10 to 24**

This phase of the project involves the development of heuristics for classifying email flows by whether they resemble legitimate usage or whether they resemble spamming.

As indicated above, initial ideas involve assessing volume (most legitimate sources of more than 10,000 items of email a day are “well-known” ISP server machines); time of use (legitimate email is less common at 4am); smoothness of flow (legitimate email tends to clump, spam tends to be sent continuously), and so on.

This phase will involve considerable feedback from ISP partners to establish what they detect by other means (and hence indicates a pattern of interest) and what they discover when they investigate the reports generated by the project.

The deliverable from this stage is a set of effective heuristics that can be applied to sampled sFlow data to detect spam sources.

### **Dissemination: a continuous activity**

Dissemination of project results is very important. Although the full power of the system will not be apparent for some time, the work has been arranged so that, even at an early stage, data is produced that would otherwise be difficult for an individual ISP to obtain without specialist traffic data monitoring systems of their own.

A website will therefore be developed by month 6 to enable authorized access to traffic pattern data. Existing “whois” databases can be leveraged to establish ownership of IP address blocks and access to data about these blocks can be validated by the ability to respond to email sent to the email addresses recorded as owning the blocks. A similar (albeit somewhat flawed) implementation of this is already used by the MSN “Smart Network Data Services” project.

An important part of the project will be to proselytize the value of the system to ISP abuse teams, who may be initially reluctant to act upon reports that differ from those they are used to, in that no content is present. This will be achieved by speaking at industry events such as IXP and ISPA member meetings and by using industry mailing lists to reach the appropriate people. We will be exploiting our long-term involvement and reputation within the UK ISP community to persuade ISPs to take the reports from the system seriously and to provide feedback so as to tune the detection system for accuracy and sensitivity.

We also wish to build on existing work on email server log processing, in particular in the building of inter-ISP cooperation in returning details to the source ISP when other ISPs detect bad traffic on their servers. Demon Internet and NTL are expected to co-operate on a pilot project for this, and other partners are being sought.

At present there are no widely accepted formats for automated reports, so work is needed to develop suitable formats. It may also be apposite to provide appropriate tools to integrate these reports into the existing work flow of abuse teams.

## **Expected results**

By the end of the project, a set of tools will be available that will enable other IXPs to roll out similar systems. Once the scheme is proven to be effective, it is likely to be of interest to major ISPs who will be able to collect data within their own networks at higher sampling rates, giving them more accurate indications of problems.

Making the detection system work, and ensuring that its reports are acted on, will have a significant impact on the length of time that spammers will be able to exploit compromised end-user machines. That should lead to a reduction in the amount of spam in the UK. Provided that the reports are seen as trustworthy, there is also considerable potential for automating the response to the reports, which should lead to reductions in staff costs in ISP “abuse teams”.

## **Project Extensions**

The compromised end-user machines that are being detected by this project are only one part of the overall spam sending system. It may also be possible to extend the traffic analysis to detect common features of other connections to and from these machines – with a view to locating the next level of controlling machine. There are a number of policy and legal issues to address, but there is certainly scope for disrupting the spam sending operation per se, rather than just fixing insecure ISP customers after the event.

Information from the project may enable some categorization of the spam senders themselves by collating information about their modus operandi. SpamHaus claim that there are less than 200 “spam gangs” in total. Tackling these gangs requires different tactics than, for example, tackling 200,000 small-scale spam-sending entrepreneurs. Assessing the accuracy of the SpamHaus estimates, and providing data on estimated traffic levels may assist the public policy debate and will in particular help the regulators and enforcement agencies in finding the correct approach to preventing spam at a non-technical, legal or society level. In addition, estimates of

traffic volumes may assist the courts in correctly assessing damage – and hence appropriate sentences – when spammers appear before them.

When end-user machines are found to be sources of spam, ISPs face a dilemma. Disconnecting the machine stops the flow of spam, but it is far easier to “clean up” a machine that is still connected to the Internet. A recent trend has been to quarantining machines within a “walled garden” so a handful of websites, such as anti-virus vendors, can be accessed, but the rest of the Internet remains inaccessible. There is a need to develop better and more effective cleanup software to assist end users in removing malware from their machines so that they are no longer compromised. If the traffic analysis has indicated which spam gang was responsible and hence which technique was likely to have been employed, this can be leveraged to provide very specific assistance in cleaning things up.

There will also be opportunities to better understand the nature of spam-sending trojans and propose systems that would interfere with their activity. It may also be possible, for example, to infer the presence of malware from traffic patterns in the DNS protocol and then disrupt that traffic. Similar techniques are at the forefront of the current attacks on “botnets”.

One of the difficulties encountered with anti-spam activity is in distinguishing poorly operated legitimate activity from spamming. False positives arise with existing systems, especially with poorly maintained mailing lists. A particular need is to assist the operators of these systems with automated methods of handling delivery failures.

## **Intellectual Property**

All project results will be placed in the public domain. All code, scripts and systems will be made available under appropriate open source licenses. Project partners will have early access to results and prototypes.

This emphasis on openness and the public domain is not only the “right way” to ensure that this technology is rolled out in a rapid manner so as to make an impact on the spam problem, but it should also be noted that the LINX would refuse to fund or, even more importantly, contribute data to a proprietary research agenda.

## **Project Mechanics**

The research is to be carried out within the Security Group of the Computer Laboratory of the University of Cambridge. This group has been in existence for many years and ground-breaking research has been done on cryptography, formal methods, security policies, electronic commerce, steganography and information systems. At present it comprises, 6 academic staff, 2 post-doc researchers and 21 PhD students. There is a strong tradition of funding from industry sources and much of the group's most interesting work has been inspired by tackling real problems. Details of current group members, interests and publications can be found at:

<http://www.cl.cam.ac.uk/Research/Security/index.html>

The work proposed in this submission will be performed by “Named Researcher” Richard Clayton. He is used to handling large amounts of information and picking out small amounts of signal from the noise. He also has a background at a senior level in

the ISP industry in the UK and has considerable experience of what abuse teams need in order to do their job. An important part of this work is not only to detect the sending of spam but to generate reports of this in a timely manner – which will enable the hosting ISP to take appropriate action. Clayton’s background and contacts will enable him to proselytize the accuracy and usefulness of traffic based analysis (rather than the content-based reports that are the current stock-in-trade of abuse teams).

The “Principal Investigator” (the responsible member of staff) will be Professor Ross Anderson, <http://www.cl.cam.ac.uk/~rja14> who has an international reputation as a leading researcher within the computer security field.

The LINX will be providing the sFlow data and the systems on which the initial selection of relevant traffic is made. A desktop machine for the project will need a specification suitable for high speed analysis of significant amounts of stored data. It is expected that LINX will provide web server resources for access to traffic data. There are no other specialist equipment needs.

## **Budget**

The project budget (using standard University of Cambridge costings) is:

Senior RA (total cost including all overheads)	= £82,314
Travel (2 research conferences, participating ISPs, IXP meetings)	= £ 3,000
Equipment (PC for processing sFlow data)	= £ 1,500
TOTAL for first year	= £86,814
With inflation at a standard rate, second year =	£89,694

## **Funding Partners**

The project is expected to be co-funded by four distinct types of organization:

- LINX;
- “academic” funding from Intel Research
- a few major ISPs;
- Department of Trade & Industry

The initial concept was for a LINX only project, but leaving aside the cost, LINX believed that it was very important to have other contributors.

The ISP partners will have early access to project results and will be a key part of the evaluation. It is notoriously difficult to get ISP commitment to assist in research, especially over the medium to long term, but if they are helping to fund it they will be much more motivated to stay involved.

The contribution from Intel Research, who have already agreed to put up approximately half the total cost, reflects their long term interest in traffic monitoring (the spamHINTS system will be integrated with their existing CoMo project).

The DTI involvement is conditional on ISP industry funding, but ensures that at a time when the industry is extremely cost-conscious, research into new approaches to problems remains possible.

## Risks + Rewards

The project uses existing technology that is already being deployed by LINX for other purposes, so most risks relate to the processing of the data that is captured.

Although the sampling rate (1 in 8192) means that data rates are low, there may still be some difficulty in performing complex analysis on large datasets. Conversely, the sampling may make it hard to pick out anything other than the most egregious senders of spam – making the system of limited use. Tuning the systems to be more sensitive may generate too many “false positives” for the information to be useful to ISPs; making it less sensitive may mean that only a small proportion of compromised machines are detected. These risks cannot be avoided, but existing success with email server log processing systems suggests that a workable system can be created and that spammers will have some difficulty in evading detection by small changes to their sending methods.

Other areas of risk are that the traffic analysis methods may be seen as too esoteric for ISPs to act upon the reports they receive. This is addressed by involving some ISP project partners in the project from the beginning and by ensuring that their positive experiences are promulgated to others.

It would be implausible to set a success criterion for this project as being the elimination of spam within the UK. However, what does look achievable is to reduce the “time to detect” for compromised end user machines, and to improve the efficiency of ISP abuse teams in dealing with problems.

The project can also be judged a success if it leads to the deployment of similar traffic analysis schemes at other IXPs and within major ISP networks.

## Related Work

Almost all work on email spam has concentrated on looking at content, usually with a view to filtering the unwanted material. Clayton has looked at unusual traffic data patterns in the server logs for outgoing email “smarthost” servers provided by ISPs for their customers [refn 7]. He obtained further success by examining incoming server logs [refn 2], because spammers (and email virus/worm malware) send small amounts of traffic to other customers at the same ISP, which can then be detected.

There has been considerable research on detecting DDoS attacks by examining network layer traffic data, but looking for email spam using this data is unusual. Gupta and Sekar [refn 9] looked for traffic anomalies (there is lot more email today than yesterday). Whyte et al [refn 11] assumed that sending email without generating any DNS traffic was suspicious (because it indicated a relay) and Mushashi et al [refn 10] examined the content of DNS traffic to distinguish which particular mass-mailing virus was present and Ishibashi et al [refn 12] used Bayesian inference to separate DNS activity so that malware-infected machines could be detected.

This proposal differs from earlier work because it intends to examine SMTP traffic at the network layer, extending Clayton’s existing approach of looking at traffic patterns, but with considerably less protocol level information than is available from

email server logs. The scale of the work (at an IXP) and the statistical nature of the sFlow information are further challenges.

There is limited deployment of the existing research results at present. The email server log processing is used daily at Demon Internet in the UK and is expected to be adopted by some other major ISPs in the UK in the near future. The other work appears to be mainly academic and we are not aware of it being deployed in industry, although Mushashi et al work for NTT in Japan and are clearly in a position to apply their results within that major ISP. DDoS detection systems are of course a growth industry, and there are a number of products based upon that work.

## **Commitments**

TBA

## **CV for Richard Clayton:**

Richard Clayton obtained the Top First in Computing Science from the University of Manchester in 1977. For the next twenty or so years he worked in the software industry developing software for the mass market. From 1983 to 1995 he co-owned a successful software house that developed the software for the Amstrad home computers and the Amstrad PCW word processor, a best-selling icon of the 1980s. In 1995 his company was acquired by Demon Internet who wished to exploit the Internet access software, Turnpike, which he had designed and developed.

Clayton worked for Demon until 2000 in a variety of roles, mixing regulatory issues with programming projects and analysis of logging-data to produce statistics and locate problems within systems. In October 2000 he seized the opportunity to study for a PhD in the Computer Laboratory at the University of Cambridge. His thesis on “Anonymity and Traceability in Cyberspace” was completed in August 2005 [refn 1].

Clayton was employed from October 2003 to October 2005 on a CMI Institute research project, but through consultancy has continued his links with Demon and with the ISP industry generally. He is the author of a number of academic papers (see Bibliography below), has edited many of the LINX “BCP” documents and has a record of fighting spam that goes back longer than most!

## **Bibliography**

[1] Richard Clayton. Anonymity and Traceability in Cyberspace. PhD Thesis, August 2005.

<http://www.cl.cam.ac.uk/~rnc1/thesis.pdf>

[2] Richard Clayton. Stopping Outgoing Spam by Examining Incoming Server Logs. Second Conference on Email and Anti-Spam (CEAS 2005), Stanford CA, USA, July 21-22 2005.

<http://www.cl.cam.ac.uk/~rnc1/incoming.pdf>

[3] Andrei Serjantov and Richard Clayton. Modelling Incentives for Email Blocking Strategies. Fourth Annual Workshop on Economics and Information Security, WEIS05, Boston MA, USA, June 2-3 2005.

<http://www.cl.cam.ac.uk/~rnc1/emailblocking.pdf>

- [4] Richard Clayton. Failures in a Hybrid Content Blocking System. Fifth Privacy Enhancing Technologies Workshop, PET 2005, Dubrovnik, Croatia, May 30-June 1 2005.  
<http://www.cl.cam.ac.uk/~rnc1/cleanfeed.pdf>
- [5] Richard Clayton. Insecure Real-World Authentication Protocols (or Why Phishing is so Profitable). Thirteenth International Workshop on Security Protocols, Cambridge, UK, April 20-22 2005.  
<http://www.cl.cam.ac.uk/~rnc1/phishproto.pdf>
- [6] Richard Clayton. Who'd phish from the summit of Kilimanjaro? Financial Cryptography and Data Security: 9th International Conference FC 2005, Roseau, The Commonwealth of Dominica, February 28-March 3 2005, volume 3570 of LNCS, pages 91-92, Springer Verlag.  
<http://www.cl.cam.ac.uk/~rnc1/kilimanjaro.pdf>
- [7] Richard Clayton. Stopping Spam by Extrusion Detection. First Conference on Email and Anti-Spam (CEAS 2004), Mountain View CA, USA, July 30-31 2004.  
<http://www.cl.cam.ac.uk/~rnc1/extrusion.pdf>
- [8] Ben Laurie and Richard Clayton. Proof-of-Work Proves Not to Work. Third Annual Workshop on Economics and Information Security, WEIS04, Minneapolis MN, May 13-14 2004.  
<http://www.cl.cam.ac.uk/~rnc1/proofwork2.pdf>
- [9] Ajay Gupta and R. Sekar: "An Approach for Detecting Self-Propagating Email Using Anomaly Detection", Proc. Recent Advances in Intrusion Detection (RAID 2003), Sept 2003  
<http://seclab.cs.sunysb.edu/seclab1/pubs/papers/raid03.pdf>
- [10] Yasuo Musashi, Ryuichi Matsuba, and Kenichi Sugitani: "Indirect Detection of Mass Mailing Worm-Infected PC terminals for Learners", Proc. ICETA2004, 2004.  
<http://www.cc.kumamoto-u.ac.jp/~musashi/musashiICETA04.pdf>
- [11] David Whyte, Evangelos Kranakis, and P. C. van Oorschot: "DNS-based Detection of Scanning Worms in an Enterprise Network", Proc. NDSS'05, 2005.  
<http://www.scs.carleton.ca/~kranakis/Papers/whytednswormv3.pdf>
- [12] Keisuke Ishibashi, Tsuyoshi Toyono, and Katsuyasu Toyama: "Detecting Mass-Mailing Worm Infected Hosts by Mining DNS Traffic Data", SIGCOMM 2005.  
<http://www.sigcomm.org/sigcomm2005/paper-IshToy.pdf>