Email Traffic: A Quantitative Snapshot

Richard Clayton
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, United Kingdom
richard.clayton@cl.cam.ac.uk

ABSTRACT

It is common to think of email as a one-to-one communication medium, but at the ISP level, many email flows are mailing-lists (one-to-many) or forwarded traffic (many-to-one). Some anti-spam systems have foundered on misapprehensions as to the nature and importance of these flows. However, although understanding has grown, there are no quantitative studies in the literature as to the relative importance of these different types of email flow. This brief study is a snapshot of the types of email that can be distinguished amongst the 331 million items that arrived at a medium-sized ISP in March 2007, and is intended to provoke the publication of further data, to better illuminate the relative importance of different types of email.

1. INTRODUCTION

Email is usually considered to be a one-to-one communication medium, but at the Internet Service Provider (ISP) level, many email flows are mailing-lists (one-to-many) or forwarded traffic (many-to-one). However, the literature contains few, if any, quantitative measures of what is meant by "many". This paper analyses four weeks of data from a particular ISP to provide some hard figures.

The email traffic dataset is described in Section 2. In Section 3, an estimate is made of the incoming email that has been automatically forwarded from other sites, and in Section 4 heuristics are presented that provide an estimate of incoming mailing-list email. In Section 5 there is a discussion of the implications of this data for anti-spam proposals, and the paper concludes by calling for further work to provide accurate quantitative views of email volumes and patterns, in both space and time.

2. THE DATASET

The dataset analysed in this paper is the incoming email to Demon Internet, a United Kingdom ISP with c 150 000 customers: a mix of individuals, and small and medium-sized businesses. Demon sets the MX records for generic customer sub-domains (e.g.: `example.demon.co.uk`) as well as many specific customer domains (e.g.: `example.co.uk`) to point at its main email servers, and hence they handle the vast majority of email arriving at the ISP. The exceptions are larger companies (where MX records point at customer machines) and intra-ISP email – Demon Internet customers sending email to each other.

Traffic data (the date, time, source, destination and size) of incoming email was collected for the four week period 1–28 March 2007; normal working weeks with no UK public holidays. An attempt was made to deliver a total of 331 858 366 emails (just under 12 million per day) addressed to 355 559 372 different addresses (i.e. 1.15 addresses/email).

In addition, 41 565 269 (about 1.5 million a day) deliveries failed to complete all the steps of the SMTP protocol, and little more can be said about these. Also, 22 400 218 of the delivered emails had a "null" (`<>`) sender (an average of 800 000 per day). The nature of these will vary, but the majority will be "backscatter", where emails cannot be delivered elsewhere and the "bounces" are being delivered to the (forged) source address at Demon. This leaves 309 458 148 "real" emails actually delivered.

On their incoming email servers, Demon Internet operates a spam detection system provided by Cloudmark. This marked an average of 73% of the "real email" as spam, so that it was not accepted (a `5xx` permanent failure return code is used to indicate this to the sender). Hence 83 720 106 emails (an average of 2.99 million per day) addressed to an average of 1.18 destinations per email were non-spam items to be delivered (along with the bounces already mentioned) to Demon's customers. The ratio of destinations per email for the items detected to be spam was 1.10.

The low ratio of addresses to emails shows that spammers, who used to regularly use multiple (often hundreds) of destinations per delivery, now – far more often than not – use individual addressing. Blocking emails for high destination counts is a common heuristic, for which email servers provide simple configuration options, and so it is unsurprising to find that high numbers of recipients is now a rarity. Previous data is hard to find, but in 2004 Gomes et al. examined a somewhat smaller dataset and found the destination-per-email ratios at that time to be 1.4 overall, but 1.7 for spam email [3].

Figure 1 shows how email "spam" took little account of which day of the week it was, generally varying between 6 and 10 million items per day (although with a substantial spike to over 13 million items on the 20th March – having been just 5 million spam items on the 18th). However, the pattern of non-spam email varies consistently by the day of the week, with somewhat less email arriving at weekends. In passing, it should be noted that the spam detection system is far from perfect, and therefore some of the variation in non-spam email must be assumed to be new types of spam, that

CEAS 2007 - Fourth Conference on Email and Anti-Spam, August 2-3, 2007, Mountain View, California USA

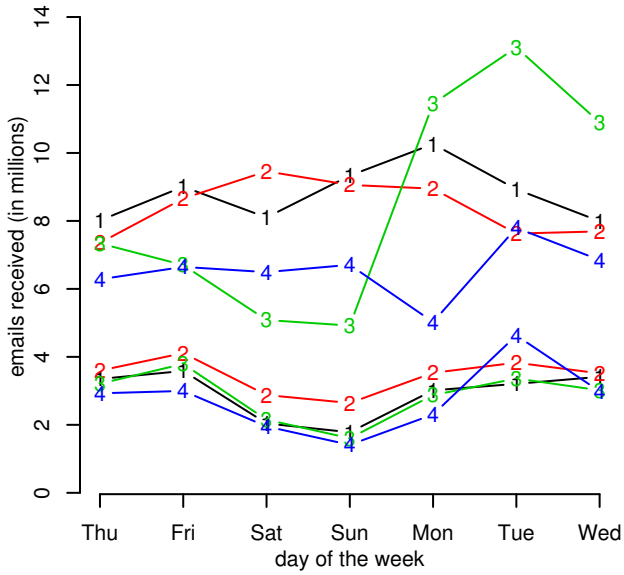

Figure 1: Spam (the upper values, in the 5 to 13 million range) and non-spam email (the lower values, in the 2 to 4 million range) that arrived at Demon Internet during the first four weeks (1st to 28th) of March 2007. The data excludes the approximate 800 000 "bounces", viz: email with a null SMTP senders, that also arrived each day – mainly "back-scatter" from spam delivery attempts elsewhere.

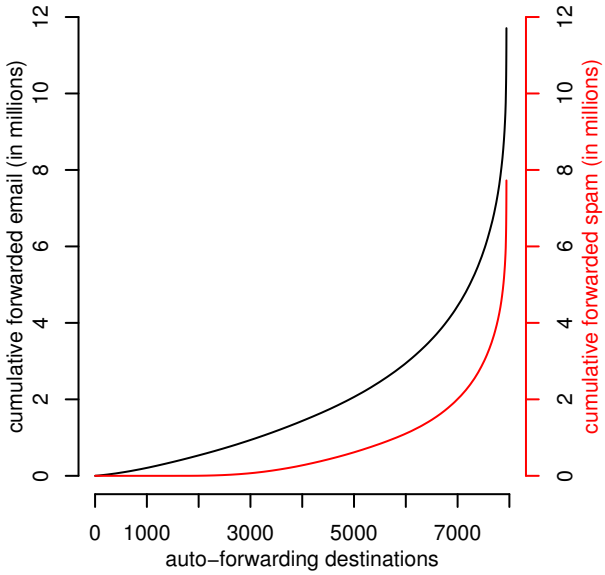

Figure 2: Auto-forwarded email received by Demon Internet customers during the period 1–28 March 2007. The upper curve is the cumulative count of forwarded email for recipient addresses in ascending order of amount received. The lower curve is the cumulative count of received spam, with the addresses re-ordered into ascending order of the amount of spam received.

were not yet being detected effectively. This is especially noticeable on the Tuesday of week 4 (the 27th) where a new type of German language "pump and dump" spam evaded the filters for several hours.

3. FORWARDED EMAIL

Email is forwarded from one site to another for many reasons. Some individuals forward interesting material to friends, relatives, or colleagues, or just pass on email that would be better dealt with by someone else. This *ad hoc* forwarding is supplemented by automated forwarding that sends all email for a particular domain or user to another site. Some companies auto-forward email from their main site to branch offices or out-workers. Domains are often hosted by third parties, along with websites and other services, and email to that domain is forwarded to addresses at the owner's ISP.

To identify auto-forwarded email, the dataset was scanned for remote machines that sent more than 10 items in a single day to a single email address. When this occurred on more than half the days in the sample (i.e. 14+ days), then this was deemed to be auto-forwarding, and a complete count made of all items (even when less than 10) transferred (incomplete sessions were ignored since they were more likely to be transient faults, than misbehaving sources).

This method of counting forwarded email will slightly underestimate the total – the main omission being where entire domains are forwarded without rewriting the local parts, and some of those local parts are only used occasionally. The other likely error is that the count may include some spam sources, but fortunately, current spam senders appear to change source and destination very quickly, so the 10 and 14 values are sufficient to exclude these. Conversely, where the source of email was a cluster of machines, there may be undercounting whenever some machines in that cluster are not identified as a source.

This approach counted 11 709 190 emails during the month that were forwarded to 7 943 different email addresses (used by 5 427 different customers). Hence, approximately 3.5% of all incoming email can be detected to be auto-forwarded. Of this email, 66.0% was determined to be spam and was not accepted.

However, these totals hide considerable disparity, as is indicated in Figure 2 which shows the cumulative count of received emails and spam for the recipients, ranked in each case by the number of items received. The top 5 destination addresses account for 9.4% of all forwarded email, and the top 100 for 31.4%.

The percentage of forwarded email that is spam also varied immensely. 1 522 addresses (19.2%) received no spam at all (but 13% of the email), and 2 368 (30%) had less than 10% of their forwarded email (which was 20% of the total) rejected as spam. These low spam figures occur because either these customers have disabled spam filtering on their accounts, or the remote site is being effective at removing spam before forwarding the email. Conversely, 3 781 (48%) of the email addresses received more than the average amount of spam during the month (73%, see above). But, because these tended to be the recipients of higher volumes generally, (and 178 received 99%+ spam), this meant that the 64% of the forwarded email being sent to them averaged 90% spam.

4. MAILING-LIST EMAIL

The dataset was re-examined, but excluding the auto-forwarded email, to identify mailing-lists. Unfortunately, it proved impossible to develop effective criteria based upon regularity, number of sources, or number of destinations, so as to pick out mailing-lists and not include spam runs. Mailing-lists may send out dozens of messages a day, or just one a month. They may be sending copies of email to dozens of customers, or to just one. Although they usually sent email from a single source (or machines hosted on a /24 subnet) this was not always the case.

To avoid these difficulties, an alternative, heuristic, approach was taken. Email was defined to be mailing-list email if it appeared to be being sent by standard mailing-list software. The local part of the sender string was checked for strings such as `-bounce-`, `-owner-`, `-admin-`, `-notify-` or `listmanager@` and the hostname was checked to determine if it began `list.`, `lists.`, `bounce.`, `return.`, or if it was one of a handful of special cases such as `message.myspace.com`. The resulting list of addresses (from 25 792 domains) that were sending email was then checked by eye to ensure it looked plausible.

This method underestimates mailing-list email, because it will fail to identify *ad hoc* lists sent from addresses such as `alice@example.com`. It also excludes mass commercial mailings (such as from `info@example.com`) for companies that do not use mailing-list software to automate their newsletters. However, it does appear to exclude most outright spam (from `service@paypal.com` etc.).

The total amount of incoming mailing-list email identified in this way was 5 753 383 emails of which 550 136 (9.6%) were identified as being spam. Figure 3 shows the cumulative frequency of email and spam (in the same way as Figure 2 above, with the x-axis sorted by sending domain volume. Once again, the most striking feature of the results is that many (12 824 or 50%) of the sources are completely spam free. Also, the data is again dominated by a small number of sources (though these were running a large number of individual mailing-lists, so this is somewhat misleading).

The figures for mailing-list email are unexpectedly low. A similar heuristics-based count at Demon in 2004, reported in [5], found that approximately 40% of non-spam email could be identified as coming from mailing-lists, whereas the proportion – on an entirely comparable basis –is now about 7%. There are two possible explanations for the substantial decrease. The first is a change in the type of customer that Demon Internet attracts, with individuals being replaced by small and medium size businesses. The second, and more likely, explanation is that a lot of traffic that used to be on email mailing-lists has migrated to web-based forums. Although the main reason for using the web will be ease-of-use and richness of features, a contributory factor will be the increasing difficulty, well-documented by Cohn and Newitz [2], of delivering bulk email, albeit solicited material, to over-enthusiastic and unhelpful spam filtering systems.

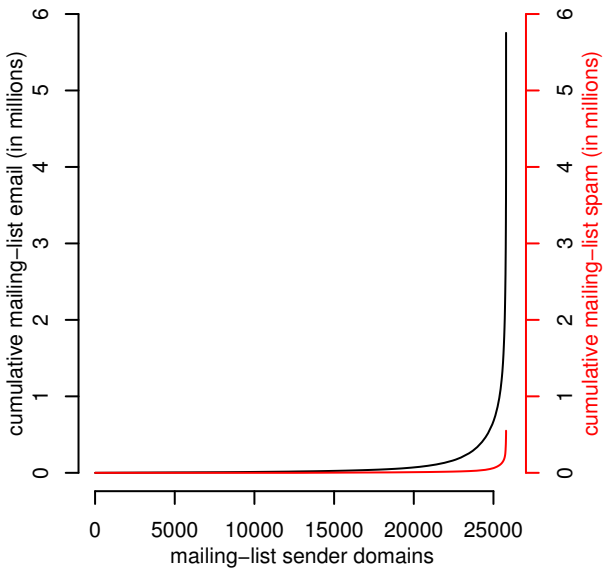

Figure 3: Mailing-list email received by Demon Internet customers during the period 1–28 March 2007. The upper curve is the cumulative count of email by sending domain in ascending order of amount received. The lower curve is the cumulative count of received spam, with the addresses re-ordered into ascending order of the amount of spam received.

5. DISCUSSION

The sending of bulk unsolicited email ("spam") has been a major problem for over a decade, and considerable effort has been put into attempts to prevent it from reaching user in-boxes. Most practical systems use some combination of blacklists (blocking known senders of spam, or hosts sited within subnets known to be used by mass-market ISP customers), pedantic approaches to the SMTP protocol (greylisting [4] will catch out spammers who do not re-queue temporary failures) and inspection of the email content itself (Bayesian filters etc.).

Many people wish to bolster these systems with reputation systems tied to the actual sender of the email. One strand of this work is Sender Policy Framework (SPF) [10] and the related SenderID proposal [6]. The idea is for email servers to reject email when a forged sender is used in the protocol – with the domain name owner specifying which servers are authorised to send outgoing email. SPF schemes break when email is forwarded without any rewriting – and it is generally believed that forwarding is commonplace.

The data from Demon Internet shows that forwarding does occur, but it is quite a small proportion of all email, and hence might not be an insurmountable obstacle for SPF-style schemes. What is also noteworthy is that a considerable amount of spam is being forwarded (about two-thirds of the total) and then rejected by Demon. If the sending machines follow the SMTP protocol, this will cause the generation of a Delivery Status Notification (DSN) – which, most likely, the machine will attempt to deliver to some innocent third party whose identity was forged as the sender. This volume of forwarded spam (approximately 276 000 items a day) is of a similar order of magnitude as the 800 000 incoming "bounces" per day – suggesting that this could be a major contributor to "back-scatter" bounces, and hence it could be worthwhile considering email system enhancements that mitigate this mechanism.

A different approach to reputation is that of Domain Keys

(now being standardised as DKIM [1]) where messages are cryptographically signed to make it impracticable for spammers to forge a sender identity, and hence email from senders with a good reputation can be accepted without any fuss. DKIM is designed to be unaffected by forwarding schemes or distribution by mailing-list software, provided that the signed header fields and body text remains intact. It would be valuable to determine how often signatures failed – something not currently logged. This would prove a useful adjunct to efforts such as "The Forwarding Project" [8] which checks whether particular email server software causes any signature failures.

A third type of anti-spam proposal attempts to create an economic disincentive to spamming by using cryptographic primitives as a "proof-of-work" to show that the sender has generated a limited number of emails per day. Proof-of-work systems have to fudge around the issue of mailing-lists (and similar amplifiers) because it is assumed that these systems are common and cannot afford to generate the necessary proof-of-work for each individual email that they send. If future work confirms that mailing-lists no longer form 40% of all email, but a mere 7%, then this impediment reduces in importance. Nevertheless, the arguments set out in [5] remain relevant – that spam for high-profit-margin items will remain economic, and that spammers can use their "botnets" for the proof-of-work calculations and continue to send spam at unacceptable rates.

It is quite striking how few figures about email traffic appear in the published literature. Most measures of email are either overall estimates of global volumes, or reports upon the fraction that is being detected by a particular service provider to be spam. Since individual ISPs and spam filtering providers have recently been reporting immense day-to-day swings of spam volume (as indeed is visible in Figure 1), this lack of data is even more surprising.

It is widely believed that there are quite a small number of senders of the majority of email spam, with organisations such as SpamHaus headlining "200 Known Spam Operations responsible for 80% of your spam" [9]. However, the huge day-to-day variability in volume has also been reported by others – MessageLabs call the phenomenon "spam-spikes" [7]. The Demon figures indicate a background spam rate of some 5 million items per day, which may indeed be sent by 200 gangs whose changes of target tend to cancel each other out. However, the indications are of a mere handful of gangs sending 8 million items a day between them – with their spikes not being obscured by other activity of similar size. If there are indeed just a handful of gangs contributing so significantly to the spam problem then this has major public policy implications – the number of investigations required to locate those responsible is several orders of magnitude less than even SpamHaus's estimate.

6. CONCLUSION

This short paper is intended to be just a small contribution towards measuring email. It is, necessarily, reporting on just one ISP, in one country, in one month, but the methodology is explained clearly enough for others to duplicate and refine; and to repeat in future years to detect changing patterns. Of particular value would be figures from large, million-subscriber, consumer ISPs, which might be expected to show considerably different patterns of forwarded email and mailing-lists – but the nature of that difference can currently only be guessed at.

Acknowledgements

We thank Demon Internet for providing access to their email traffic data. Richard Clayton is presently employed by the spamHINTS project, supported by Intel Research.

7. REFERENCES

[1] E. Allman, J. Callas, M. Delany, M. Libbey, J. Fenton, and M. Thomas. Domainkeys identified mail (DKIM) signatures. IETF RFC 4871, May 2007.

[2] C. Cohn and A. Newitz. Noncommercial email lists: Collateral damage in the fight against spam, 2004. `http://www.eff.org/wp/SpamCollateralDamage.html`.

[3] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and W. Meira Jr. Characterizing a spam traffic. In *IMC '04: Proceedings of the 2004 Internet Measurement Conference*, Oct 2004.

[4] E. Harris. The next step in the spam control war: Greylisting, 2003. `http://projects.puremagic.com/greylisting/whitepaper.html`.

[5] B. Laurie and R. Clayton. Proof-of-work proves not to work. In *Third Annual Workshop on Economics and Information Security, WEIS04*, May 2004.

[6] J. Lyon and M. Wong. Sender ID: Authenticating e-mail. IETF, RFC4406, April 2006.

[7] MessageLabs Ltd. MessageLabs intelligence: May 2007 "Spam spikes – the battering ram of spam", 2007.

[8] The Forwarding Project, 2007. `http://forward.sp.am`.

[9] The SpamHaus Project. The ROKSO list, 2007. `http://www.spamhaus.org/rokso/`.

[10] M. Wong and W. Schlitt. Sender policy framework (SPF) for authorizing use of domains in e-mail, Version 1. IETF, RFC4408, April 2006.