Do Zebras get more Spam than Aardvarks?

Richard Clayton
Computer Laboratory
University of Cambridge
CAMBRIDGE, CB3 0FD, UK

Abstract

Analysis of traffic logs of email received by a large UK ISP shows considerable disparity between the proportions of spam received by addresses with different first characters. This disparity is quite marked when only email addresses that appear to be 'real' are considered. The root cause is likely to be spammers using 'dictionary' or 'Rumpelstiltskin' attacks to guess valid email addresses. There is limited evidence for these attacks taking place in real-time, suggesting that 'fake' email addresses were constructed sometime in the past and are now immortalised within spammer databases.

1 Introduction

The recipients of unsolicited bulk email (spam) report very differing experiences of how much spam they each receive. Some of these differences are undoubtedly due to how visible individual email addresses are, or how they are used [Hann 2006], and studies regularly find variations between different business sectors [MessageLabs 2008]. In this paper, we consider whether some of the different perceptions may arise from as simple an issue as which letter of the alphabet an individual's email address begins with; that `zebra@example.com` receives a lower proportion of spam than `aardvark@example.com` might expect to.

Having shown that the first letter of the local part of an email address does indeed make a difference, we then discuss why this occurs, and try to quantify some spammer behaviour that might account for it.

2 Data Collection

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 eight week period 1 February–27 March 2008. This period included the Easter weekend (with two bank holidays). Data from the same ISP for a four week period in 2007 was examined in [Clayton 2007].

On their incoming email servers, Demon Internet operates a number of spam mitigation strategies – which makes direct comparison with the 2007 data rather problematic. For example, connections are rejected from sites which are listed in the SpamHaus Policy Block List (PBL) [SpamHaus 2008a], viz: where the responsible ISP has declared that machines in particular IP address blocks will only send email via their own 'smarthost' machines, so that email which arrives directly can reasonably be assumed to be spam. Additionally, greylisting [Harris 2003] is applied to machines that appear in the SpamHaus ZEN list [SpamHaus 2008b]. Any email that comes from machines that are not on blocklists, or which is retried after initial greylisting, will be passed through a spam detection system provided by Cloudmark.

In this paper, the only email considered is that which reaches the Cloudmark system. This will determine whether the email appears to be spam and if so delivery will be refused. If the email is not categorised as spam then it will be placed into customer mailboxes. A small proportion of customers completely opt out of the Cloudmark system; all of their email will be considered 'non-spam', whatever its actual nature.

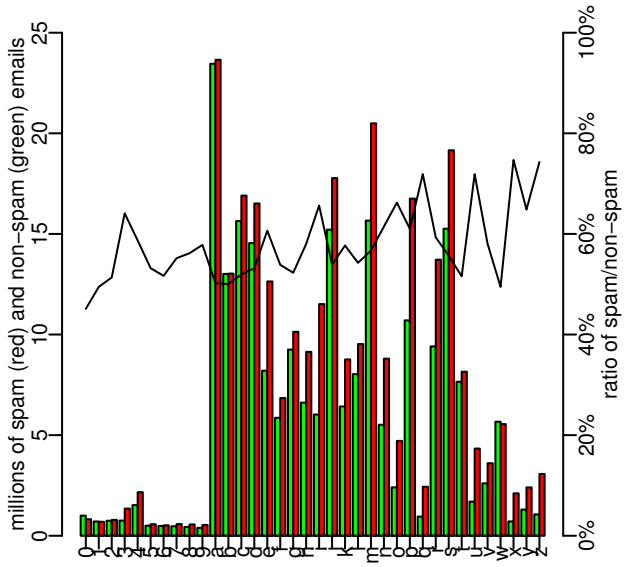

Figure 1: Spam (red) and non-spam (green) email for 8 week period, where local parts begin with particular letters. Line shows percentage of email that is spam.

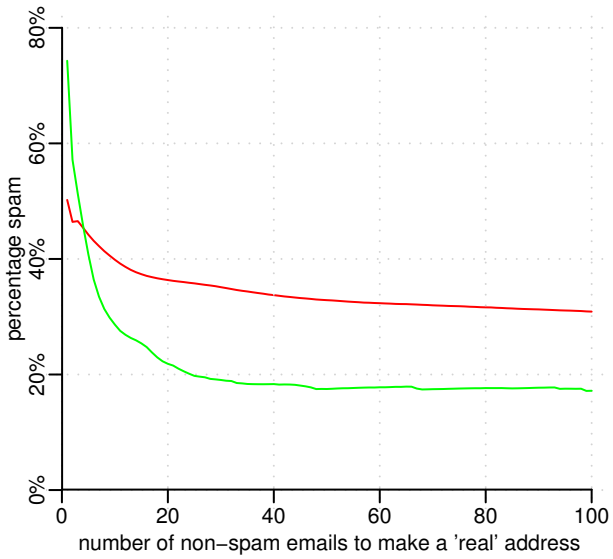

Figure 2: Proportion of email that is spam for 'real' email addresses beginning with 'a' (red line) or 'z' (green line). The x-axis is the number of non-spam emails that must be received for the address to be considered 'real'.

Some email arrived with multiple destination addresses at Demon Internet customers. Except where otherwise indicated, an email which is to be delivered to n different Demon customers is counted as if it were n different emails. Additionally, since we expect such emails to be mainly 'backscatter', for purposes of the present paper we ignore email that appears to be a 'bounce'. This was done by the inexact expedient of failing to count any email to a single destination that has a null sender (<>).

Overall, for the eight week period considered and using the above definitions, 550 596 270 emails arrived (8.94 million/day), of which 56.0% were deemed to be spam and delivery was refused.

3 The first letter of local parts

We examined the first character of the local part of the destination email addresses, ignoring the 321 730 (0.06%) emails where this did not begin with a letter or digit. For each starting character (combining upper and lower case), the number of spam and non-spam emails was counted, and the results plotted in Figure 1.

As can be seen, 'zebras' (people with email addresses beginning with a 'z') collectively do not receive very much email, but their perception is that 74.3% of all email is spam. In comparison, 'aardvarks' (people with email addresses beginning with the letter 'a') have the perception that 45.1% of all email is spam.

However, the situation changes when we consider 'real' email addresses, which are likely to be reach an individual's mailbox. We take all of the email addresses that begin with 'a' and 'z' and count, for each address individually, how many emails they receive that are spam or non-spam. We deem an email address to be 'real' if it receives at least n non-spam emails during the eight week period we are considering.

Because the spam detection system is not perfect, it is unlikely that detecting a single non-spam email will be a good indicator of whether an email address is 'real'. Varying the value of n, the cut-off point for 'realness', gives the results shown in Figure 2.

Clearly we would always expect some diminution in spam percentage as the amount of non-spam email is required to be higher – but what is most striking about the results is that an individual 'real zebra', on average, will find that less than 20% of their email is spam, whereas a particular 'real aardvark', on average, will detect that over 30% of their email is spam.

It can be seen that a reasonably steady-state is reached at around 28 non-spam emails. In other words, it is plausible to define a 'real' address as one which, on average, receives one non-spam email every second day. Using this criterion, the customers whose 62 784 'real' email addresses start with 'a' perceive 35.2% of all email to be spam. The equivalent figures for the number of 'real' addresses and the amount of email these

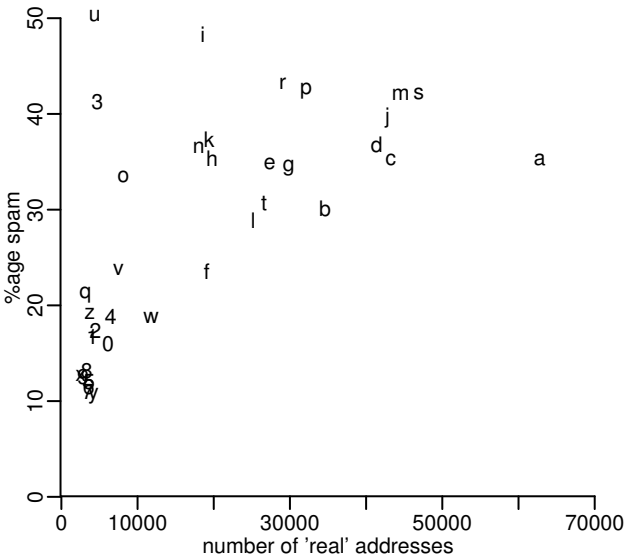

Figure 3: Relationship between number of email addresses receiving 28 or more non-spam emails over 8 weeks, and the proportion of email these addresses receive that is spam.

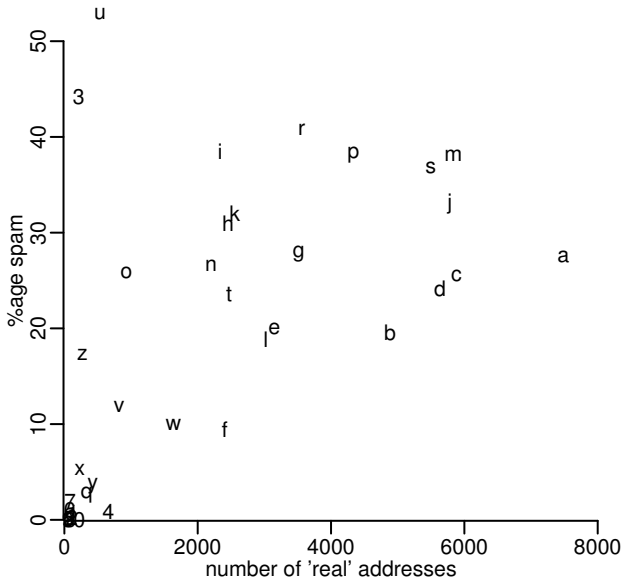

Figure 4: Relationship between number of email addresses receiving 500 or more non-spam emails over 8 weeks, and the proportion of email these addresses receive that is spam.

customers perceive to be spam are shown in Figure 3. The scatter plot shows that as the number of 'real' addresses per letter increases, the proportion of spam increases; though there are a lot of outlying points. Some outliers can be easily explained (addresses starting with '3d' have been incompetently harvested from HTML pages); others deserve further investigation.

Taking a much more cautious approach to what is a 'real' email address and requiring that over 8 weeks the address receives 500 or more non-spam emails gives the very similar relationship shown in Figure 4.

4 Spammers and spam lists

One reason for the behaviour that we have just measured is the way that spammers create and use lists of email addresses. Initially they collected valid addresses by consulting mailing list archives, scanning Usenet feeds, 'scraping' websites and so on. Systems that would once have permitted these addresses to be validated (delivery failures, the SMTP VRFY command etc.) are disabled nowadays – because other spammers were guessing addresses and using these 'oracles' to validate their guesses.

At some point, it occurred to the spammers that if `john@example.com` was a valid email address then perhaps `john@another.com` was valid as well, so they started to combine local parts (to the left of the `@`) with other domain names. This method of creating email addresses to attempt delivery to is called a dictionary attack (or sometimes a Rumpelstiltskin attack).

It ought to be possible to estimate the extent to which spammers are using lists and the extent to which they are doing the Rumpelstiltskin attack in real time, by examining runs of deliveries to the email servers.

If the spam sender is using a sorted list of email addresses – of whatever quality (possibly including Rumpelstiltskin names, possibly not) then we would expect to see incoming email addresses appearing in order. Of course the spammer may not be sorting their list, but many spammers will wish to weed out duplicates, and sorting is normally a prerequisite for this.

For each sending IP address we examine the destination of each email. Difficulties with the logging of multiple destinations email forces us to ignore this type of email in this part of the paper. This means that we are only considering 276 million emails rather than the full 300 million pieces of spam.

For each of the incoming email servers in turn, we count runs of ascending (and also descending) addresses using normal case-sensitive alphabetical ordering rules. Since we are only interested in spam senders (many senders of genuine email will also have ordered lists) we ignore any run where less than half of the emails were detected to be spam, and we also ignored any run whose length was less than 5.

Besides considering runs where the full address is in

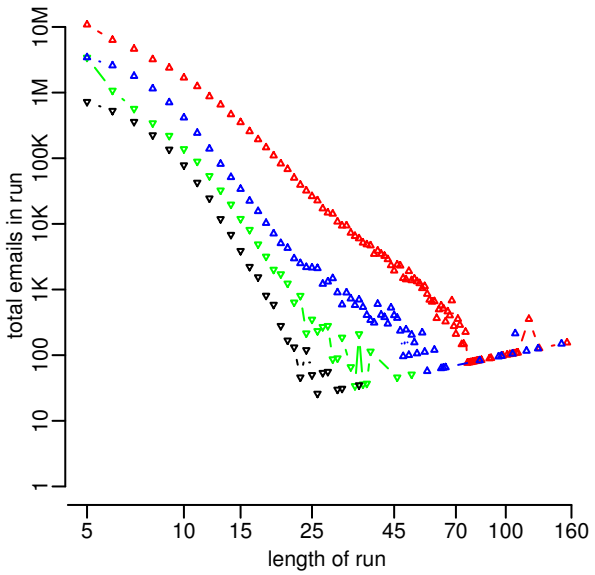

Figure 5: Email that is runs of spam in ascending (△) or descending (▽) alphabetical order of email addresses. The red/green pair match only local parts, the blue/black pair is where the run is within a single domain. Note that both axes are logarithmic.

order, we also consider where the local parts are ordered, but the domain varies.

The rather disappointing results are shown in Figure 5 (note that both axes are logarithmic). The unexpected result is that all of these runs put together only account for about 2.9% of all spam, so drawing conclusions from them is problematic. It is noteworthy that ascending runs considerably outnumber descending runs, and that ignoring the domain makes runs more common, but little else can be gleaned. Since ordered lists appear very commonly when outgoing email spam is detected using the methods outlined in [Clayton 2004], the most likely reason for failing to detect them in this data is the generic anti-spam defences using the PBL and ZEN lists. Additionally, many sources sent very low numbers of emails – the likelihood being Demon Internet received a very small fraction of all the email they sent, so that picking out patterns was never going to be likely.

5 Conclusions

Measuring incoming email has shown that the first letter of email addresses makes a difference to the proportion of incoming spam. As a group, 'zebras' receive a higher proportion of spam than 'aardvarks'. However, when considering 'zebras' that actually exist (in that they receive non-spam email), they receive a lower proportion of spam than actual 'aardvarks'.

There is some evidence that this effect is caused by Rumpelstiltskin attacks, but little evidence that incoming email has been alphabetically sorted by recipient – suggesting that these attacks are not being done in real time.

Turning the results around – there are some hints here about viable anti-spam policies. Although classifying email addresses by the amount of non-spam email received is not a very sensible way of deciding whether future messages will be spam or not, it does seem clear that there is a significant gain in spam filtering efficiency to be gained from making the ISP email reception systems aware of all valid email addresses.[1] This data indicates that around half of all the email which is being given to the spam detection system is destined for non-existent mailboxes.

Acknowledgements

We thank Demon Internet for providing access to their email traffic data.

References

[Clayton 2004] R. Clayton: Stopping Spam by Extrusion Detection. *First Conference on Email and Anti-Spam (CEAS 2004)*, Mountain View CA, USA, 30–31 July 2004.

[Clayton 2007] R. Clayton: Email Traffic: A Quantitative Snapshot. *Fourth Conference on Email and Anti-Spam (CEAS 2007)*, Mountain View CA, USA, 21–22 Aug 2007.

[Hann 2006] I. Hann, K. Hui, Y. Lai, S.Y.T. Lee, I.P.L. Png: Who gets spammed? *Comm ACM*, 49(10), 2006, pp. 83–87.

[Harris 2003] E. Harris: The Next Step in the Spam Control War: Greylisting. `http://projects.puremagic.com/greylisting/whitepaper.html`

[MessageLabs 2008] MessageLabs Inc: MessageLabs Intelligence: Q1/March 2008.

[SpamHaus 2008a] SpamHaus: PBL – The Policy Block List. `http://www.spamhaus.org/pbl/`

[SpamHaus 2008b] SpamHaus: ZEN. `http://www.spamhaus.org/zen/`

[1] Although some ISPs give customers a small number of email addresses (`example1@aol.com`, `example2@aol.com`, etc.) Demon Internet, in common with many other business-oriented ISPs, provides email services for entire domains (`anything@example.co.uk`) or subdomains (`anything@example.demon.co.uk`). This means that the email systems are not *de facto* aware of which email addresses might be valid.