# On the difficulty of counting spam sources

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

## ABSTRACT

A great deal of spam comes from botnets and there is considerable interest in arranging for the bots (the compromised machines) to be made secure. In practice, the owner of the compromised machine can only be contacted via their ISP, and their helpfulness is known to vary. This variation has led to attempts to count the bots on particular networks and thereby assess the ISP's reputation. This paper presents a model for bot incidence and explains the measurement difficulties that arise from not only from the ebb and flow of botnet membership, but also from the dynamic nature of the spam sending, and the use of dynamic IP addresses. It then considers three months of data (several million emails) sent from a very large $O(10^6)$ ISP to a medium size $O(10^5)$ ISP and attempts to calculate the daily incidence of spam-sending bots at the large ISP. The wide disparity between the estimates of the upper and lower bounds was predictable from the model, and suggests that reputation values should only be considered to be rough approximations.

## 1. INTRODUCTION

Data is regularly published that purports to show that customers of particular ISPs (or residents of a particular country) are responsible for some proportion of email 'spam'. Sometimes counts are made of the number of senders, with the intent of providing some sort of measurement of how 'bad' that particular ISP or country might be at controlling the constituent machines ('bots') of spam-sending botnets.

What this short paper demonstrates is that the measurement of the number of senders can be extremely complicated, and that the methodology chosen will have a significant impact on the statistics.

The dataset analysed in this paper is for the incoming email to a United Kingdom ISP with *c* 150 000 customers: a mix of individuals, and small and medium-sized businesses. Traffic data was examined for the three month period 1 Oct–31 Dec 2009. The ISP operates a pipeline of spam detection methods, culminating in a content filtering system provided by Cloudmark. We consider email to be spam if it is rejected at any stage – unless the sender is null (<>), since this is less likely to be spam, and more likely to be associated with 'backscatter' or 'sender callback' mechanisms.

The email considered in this paper was all of the incoming items that were sent by the several million customers of one of the largest UK ISPs. This ISP also caters to individuals, along with various sizes of business. A copy of the global routing table from the end of the time period was used to identify the blocks of IP address space announced by the ISP and hence distinguish which email came from the ISP's customers. Although routing announcements do vary over time, there was little change over the relevant period.

Over the three months, 2 555 938 attempts were made to send email from the large ISP (27 782/day). Of these, 366 684 (14% 3 986/day) emails were detected to be spam. 14 957 IP addresses never sent any spam (they sent a mean of 39 messages/IP). Of the remaining 5 421 IP addresses, 2 451 sent spam and nothing else (their mean of 4.6 messages/IP was extremely small), and 2 970 sent a mixture of spam and not-spam (a mean of 120 spam and 541 not-spam messages/IP).

The mean number of spam sending IP addresses per day was 190, but the minimum was 78 and the maximum 303, and a glance at Figure 1 shows strong evidence of a weekly rhythm. If we count spam sending IP addresses on a weekly basis, and ignore December when the Christmas holidays distort the figures, then the minimum number of IP addresses is 636, the maximum 836 and the mean is 725.

So how many bots might there be on the large ISP's network? Is it 5 421, 2 451, 303, 836 or 190? In Section 2 a simple model of an ISP's customers is presented so as to explain why measurements are theoretically difficult. In Section 3 some practical attempts are made to make sense of a substantial quantity of real world data to further illustrate the difficulties that arise. In Section 4 we discuss how other work on botnet measurement has encountered similar problems. Finally, in Section 5, some conclusions are drawn.

## 2. MODELLING BOTS AT AN ISP

The basic model is that from time to time, ISP customer machines are compromised and they end up as a part of a spam-sending botnet. In the fullness of time the customer learns of the problem, probably via the ISP's abuse@ team, and the botnet software is removed. Unless the ISP has taken special steps to block or redirect email traffic, the spam that is sent will originate at the customer IP address. This address will usually be dynamically allocated – and will change whenever the customer machine is disconnected (perhaps overnight) and may also be changed by the ISP on an arbitrary, but regular, schedule. There may also be several machines, in the same household or small business, sharing a single IP address using Network Address Translation (NAT) techniques.

*CEAS 2010 - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference* July 13-14, 2010, Redmond, Washington, US

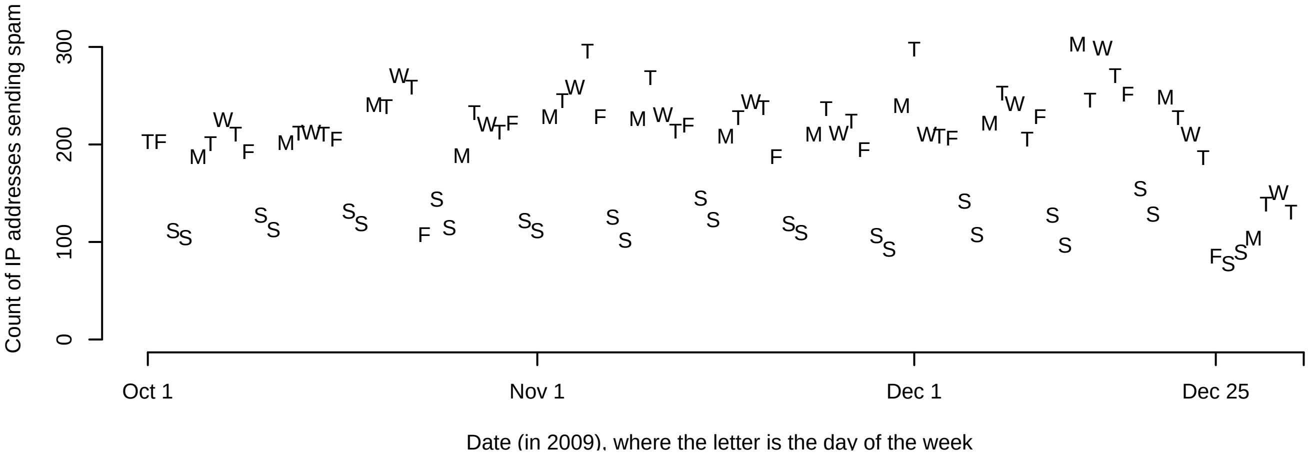

**Figure 1: Daily count of IP addresses sending spam. Quite clearly, there are fewer active sources of spam operating at the weekends.**

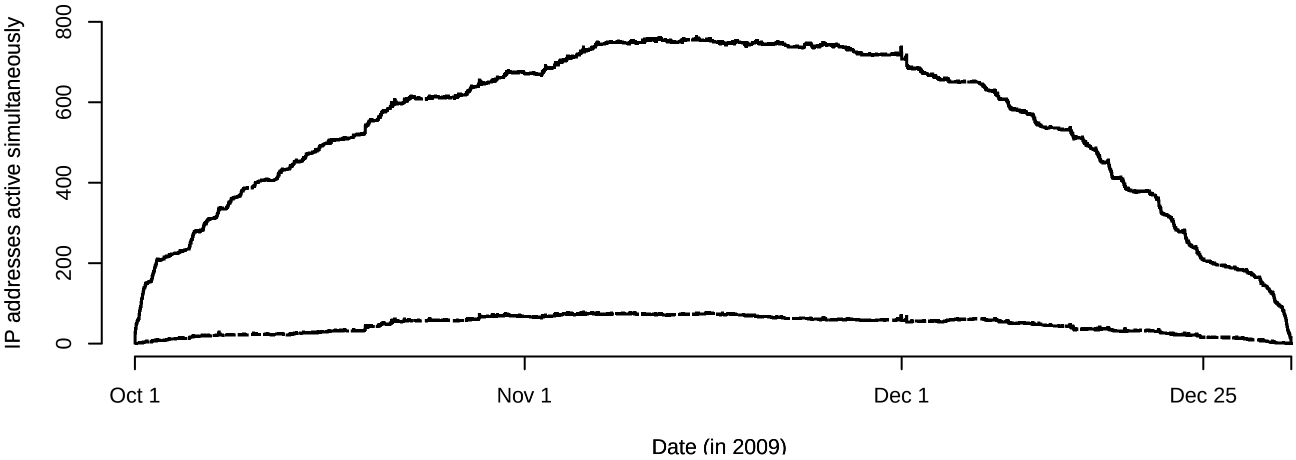

**Figure 2: Count of machines which are actively sending spam at the same time. The upper line is for all spam sending machines, the lower for machines that only send spam. Note that the ramp up in October and ramp down in December are edge effects caused by taking data from a specific time period.**

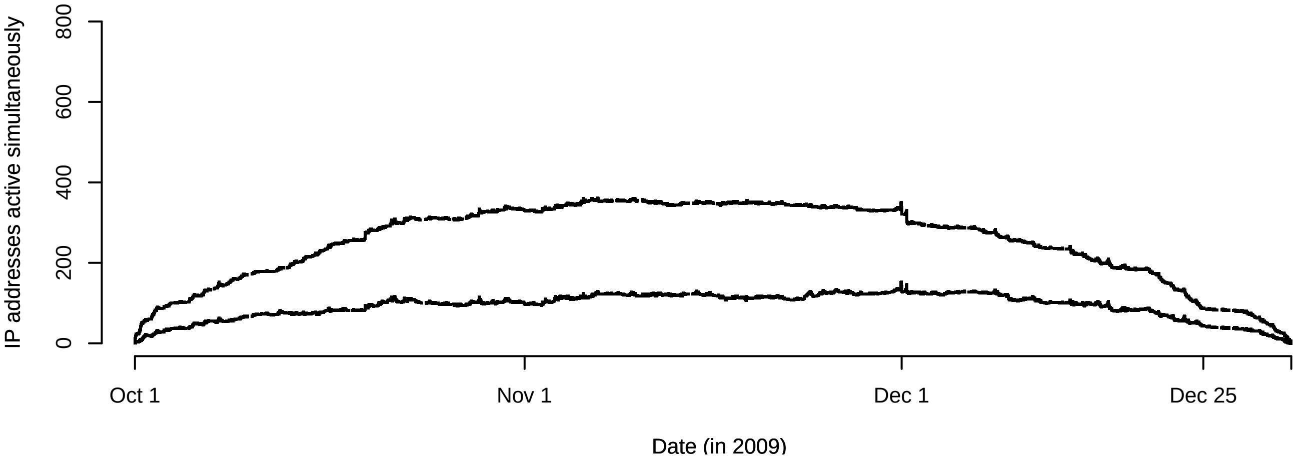

**Figure 3: Count of machines which are actively sending spam at the same time. The upper line is for all machines sending $>75\%$ spam, the lower for machines that send spam for fewer than 30 days.**

When the bots send out spam, it may reach a particular observer and they can log the sending IP address. However, the observers cannot know whether a cessation of spam means that the bot has decided to send email to other destinations, whether the machine has been cleaned up, or whether the bot has now moved to another IP address. Equally, where the spam continues, the observer cannot be sure that it is the same machine, because a completely different bot may have, entirely coincidentally, been allocated the IP address at a later time. However, at the sort of infection rates we are discussing (a few thousand IP addresses amongst millions), such coincidences would be very rare.

Further complications arise because the observer may not correctly identify the email they receive as spam. It may be novel enough to evade Bayesian filters, or it may be from a new location – detection systems often rely on the reputation of the source, and so when bots move to new IP addresses, there can be a delay before email can be correctly categorised.

Even when email is categorised as spam, the source may not be a bot. The email may be spam that is being forwarded, it may be legitimate email that unintentionally trips spam filters, or it may be a 'smarthost' providing email relay services to multiple clients – one or more of which is actually a botnet member.

An accurate count of bots is only possible if one can observe all email (getting rid of the bias caused by the location of the observer); if one can map dynamic IP addresses back to a customer account; and if appropriate adjustments are made for NAT. However, with some careful examination of the data, it may be possible to slightly improve on the rather gross estimates set out in the introduction.

## 3. MAXIMUM AND MINIMUM COUNTS

We can put a lower bound on the number of bots by seeing how many exhibit activity that overlaps in time. We first note that the incidence of bots is small, so that coincidental re-use of IP addresses by different entities will be rare. Now we consider the case where bot 'A' sends an email, then bot 'B', then bot 'A' again. When such a pattern occurs, we can deduce that 'A' and 'B' are different entities. The result of applying this analysis to all of the IP addresses that send any spam at all is a peak value of 763 on the afternoon of 14 November. However, repeating this for the machines that send spam and nothing else gives a much lower peak of 79 on November 5.

Graphs of these overlapping activity values are shown in Figure 2 from which it can be seen that there are huge 'horizon' effects – that is, the failure to consider events outside the October/December timeframe is significantly distorting the results. That is, we have a fair number of machines that are sending spam before and during November, along with machines that start to send spam during November and continue thereafter. This overlap is causing the distinctly higher values during November.

There are two obvious explanations for this effect – either that normal machines, which are regularly sending mail throughout the period have sent occasional items that are identified as spam, or that we are actually seeing a fair number of very long-lived bots.

We can try and deal with the first explanation by only considering machines where the proportion of traffic identified as spam is 75% or more. There are 3 092 IP addresses in this category (i.e. 2 451 sending 100% spam, and 641 sending between 75% and 100% spam). How many of these are active at once is the topmost plotted line in Figure 3; the maximum is 252 in early November.

Another way of excluding the normal, non-bot, machines is to take the view that any machine that sends spam for more than one month (30 days) from the same IP address is not a bot. If we process the data using this definition then we find that there were 4 719 IP addresses belonging to bots, and the maximum active at once was 152, at the end of November. This is the lower plotted line in Figure 3.

In both cases, it can be seen that the horizon effects are far less pronounced, showing that both approaches are a reasonable way of excluding long-lived senders, and this suggests that is it quite likely that a great many of the long-lived senders are normal machines and not bots.

The 30 days was chosen on a fairly arbitrary basis. As the plot in Figure 4 shows, the period during which spam is sent from a particular IP address is generally very short. One third of all sources cease after less than a single day, and another quarter within two days. However, after that, there is no clear-cut pattern with a fairly constant number of sources lasting each extra day. This tends to suggest that almost a half of all bots stay online for some time, whereas a slight majority are on machines that are switched off, or have their IP addresses compulsorily reassigned, on a regular basis.

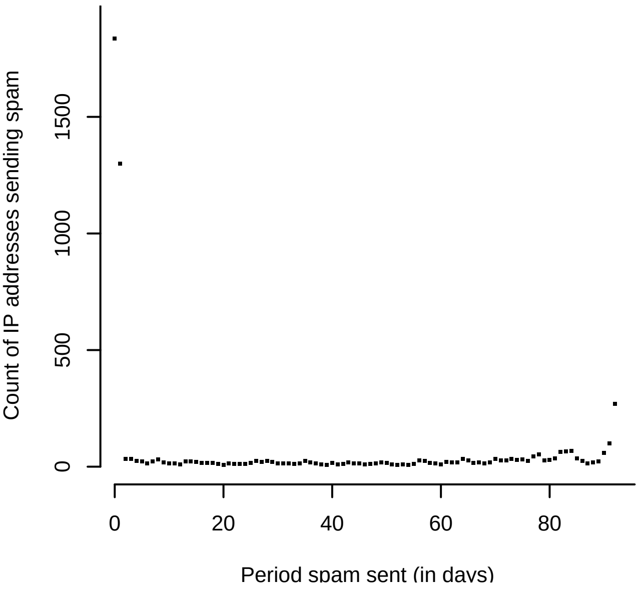

**Figure 4: Lifetime of spam sending machines.**

## 4. RELATED WORK

Counting the size of botnets is known to be a difficult problem. For example, in 2006 while Dagon et al. [1] were reporting on botnets with up to 350 000 members, Rajab et al. [2] were finding that the effective sizes of botnets rarely exceeded a few thousand bots. In 2007, Rajab et al. [3] considered these discrepancies. They looked specifically at botnets that communicated over IRC and found, in essence, exactly the same problems as are discussed in this paper – not all bots are visible in the IRC channel and IP addresses are often dynamic. They also considered how many bots were live on the IRC channel at the same time (the equivalent of the overlapping above) and after detailed discussion, concluded that no single metric was adequate for describing all aspects of a botnet's size.

Rather more recently, Stone-Gross et al. [4] controlled the Torpig botnet for a short period. Torpig gives all bots a unique identifier, enabling exact numbers of bots to be tracked. On average, they observed 4 690 new IP addresses per hour, but only 705 new bot identifiers per hour. They calculated this ratio for different countries, finding wide variations. Their UK figure, which is relevant here, was 4.48 IP addresses per bot identifier.

## 5. CONCLUSIONS

It is difficult to draw any firm conclusions from this short study. If one assumes that every IP address is a different bot, then the ISP we are considering had 5 421 bots. If we assume that at least three quarters of the email coming from a bot is detected as spam, then the total is 3 092. If we think that bots only send detectable spam, the total reduces to 2 451.

However, if we think that what we are seeing is a small number of bots that are using a wide range of IP addresses then by counting how many are active simultaneously we can find the minimum number of bots there must have been. The three cases we've just explored then reduce to counts of 763, 252 and 79.

If we assume that the ISP's abuse@ team is extremely efficient and bots are always cleaned up within a month, then there were between 179 and 4 719 bots.

Quite clearly, there is huge disparity in our estimates, and so if we were building a reputation system based on these estimates there would be considerable doubt as to its accuracy. It is not even possible to argue that the reputations are relative without examining all aspects of the model we've proposed and checking that factors such as dynamic IP address lifetimes are constant.

Finally, it is important to caution that this short paper has not even started to consider the sampling error inherent in using the logs from a single ISP to assess botnet activity. If any particular bot fails to send any email to the medium-sized ISP where we are measuring, then it will not be detected at all. Since there are likely to be enormous biases arising from the measurement location, it would be most unwise to draw any conclusions at all about the true infection rate at the large ISP, but merely to stress once again, that measurements based on counting IP addresses have enormous margins of error.

## 6. REFERENCES

[1] D. Dagon, C. Zou, and W. Lee. Modeling botnet propagation using time zones. In *Proceedings of the 13th Network and Distributed System Security Symposium NDSS*, 2006.

[2] M. Rajab, J. Zarfoss, F. Monrose, , and A. Terzis. A multifaceted approach to understanding the botnet phenomenon. In *Proceedings of the 6th ACM SIGCOMM on Internet Measurement (IMC)*, 2006.

[3] M. Rajab, J. Zarfoss, F. Monrose, , and A. Terzis. My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging. In *USENIX Workshop on Hot Topics in Understanding Botnets (HotBots'07)*, 2007.

[4] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna. Your botnet is my botnet: analysis of a botnet takeover. In *CCS '09: Proceedings of the 16th ACM conference on Computer and communications security*, pages 635–647, New York, NY, USA, 2009. ACM.