SPAM CONSTRUCTION TRENDS

Barry Irwin[1], Blake Friedman[2]

Rhodes University
Department of Computer Science
South Africa

[1]b.irwin [at] ru.ac.za, [2]blakef [at] rucus.ru.ac.za

ABSTRACT

This paper replicates and extends *Observed Trends in Spam Construction Techniques: A Case Study of Spam Evolution.* A corpus of 169,274 spam email was collected over a period of five years. Each spam email was tested for construction techniques using SpamAssassin's spamicity tests. The results of these tests were collected in a database. Formal definitions of *Pu and Webb's* co-existence, extinction and complex trends were developed and applied to the results within the database. A comparison of the *Spam Evolution Study* and this paper's results took place to determine the relevance of the trends. A geolocation analysis was conducted on the corpus, as an extension, to determine the major geographic sources of the corpus.

KEY WORDS

Spam, Geolocation

SPAM CONSTRUCTION TRENDS

1 INTRODUCTION

Unsolicited commercial email, more commonly known as spam, has placed an increasing burden on global human, computational and bandwidth resources. There is little argument over the proliferation of spam, which has seen significant increases in the quantity and frequency of its distribution into users' inboxes [2, 4]. Current estimates of the scale of the spam problem have identify that up to 80% [10] of all attempts to send email are spam related. The advent of filters which adapt to statistically identifiable components of spam has been met with spammers using increasingly complex construction techniques [1]. Spam has been shown to have a detrimental effect on the end user's perception of the integrity of email and their overall Internet experience [2]. Due to the quantity of spam and the effect this is having, there is a need to improve upon existing anti-spam techniques.

Significant research has been conducted into methods of spam detection, however little attention has been given to the analysis of spam construction trends, particularly the continuity of these techniques [5]. An understanding of whether spam emails' structures significantly vary is a critical factor in dealing with spam. Changes in the structure of spam emails, over a period, can be used to ratify specific anti-spam efforts' effect. This paper extends the framework developed by Pu and Webb [11], hereon referred to as the *Spam Evolution Study*, to further the analysis of spam construction trends.

A large corpus of spam emails was collected and processed through SpamAssassin [9] using a distributed processing architecture. SpamAssassin identifies the components which make up each spam email using a number of rule based *spamicity* tests. A complete history of the relative frequency of each component over the period of a corpus of emails is developed. Each component is then classed using Pu and Webb's original trends: co-existence, extinction and complex. Formalised descriptions of these trends are developed. The results of our trend analysis is then compared to Pu and Webb's results. Further extensions are made by associating each spam email with its geopolitical origin, based on the IP addresses of the sending *mail transfer agent* (MTA).

2 THE CORPUS

Two significant corpora were collected, combined and analysed. The first corpus consisted of a personal MTA's spam collection of 101,170 cataloged spam emails. This corpus was collected between July 2003 and July 2007, using a combination of hand sorting, Bayesian filters, *DNS blacklists* (DNSBL) and SMTP protocol conformity tests to updated the corpus. The second corpus consisted of 68,104 spam emails, collected from January 2006 until August 2007. This corpus represents a user base of approximately 3,000 schools users. This corpus is particularly of interest, as it contains spam which has evaded a far-side MTA performing DNSBL and SMTP protocol conformity tests. A large portion of this corpus consisted of spam containing MIME-encoded viruses, amounting to 2.4Gb of decompressed data.

Emails which originated from local hosts, as well as erroneous files, were removed from the combined corpora of 201,288 emails. The final size of the combined corpora is 169,274 spam emails. As with the *Spam Evolution Study* fluctuations in the quantity of spam are normalised. The normalisation is performed by dividing the spamicity count by the total number of messages per month, determining the relative state of the various spamicity tests each month.

3 DISTRIBUTED PROCESSING ARCHITECTURE

SpamAssassin 3.2.3 [9] formed the most computationally intensive portion of the study. SpamAssassin is the open-source project used in the *Spam Evolution Study*, and was the basis for characterising the various components of a spam email. SpamAssassin uses a number of methods to evaluate the likelihood of an email being spam, these include header and text analysis, DNSBL, statistical and collaborative filtering. These methods are collectively referred to as *spamicity* tests. Initial testing indicated that SpamAssassin was prohibitively computationally intensive when applied to the complete corpus. A distributed processing architecture was developed to decrease the analysis period. The architecture distributes the corpus amongst a number of processing nodes, each running a SpamAssassin instance. Spam emails are then processed in parallel on each node. The results are then submitted by each node to a database for analysis. The details of the implementation and performance of the architecture are further described in [3].

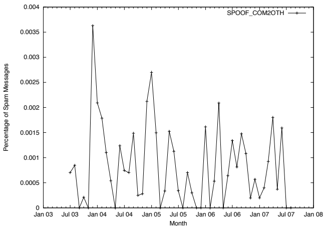

Figure 1: Complex

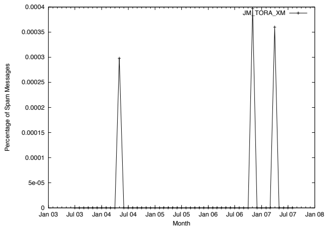

Figure 2: Extinction

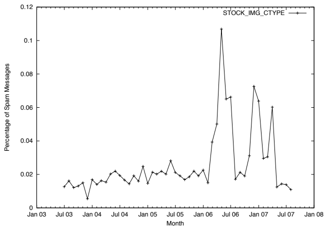

Figure 3: Co-existence

4 SPAM MODEL

Formal definitions of the *Spam Evolution Study's* trends are developed in this section. An iterative process, built on a testing framework, was used to develop the algorithms and extract the trend groups from the corpus. The framework generated a graph of each spamicity test. Each graph depicted the spamicity test's frequency, as a percentage of the total number of email for each month, over the duration of the study. Examples of these graphs can be seen in figures 1, 2 and 3. Further details of this framework are found in [3]. Each graph was categorised based on the trend algorithms, and a comparison to the data would follow to determine the accuracy of the trend.

4.1 Environment Model

To allow for more formal definitions of these algorithms, further definitions of the environment are required. The *months* during which the testing took place occurred between the start month 1 until the final month M, and are defined as $months := \{m \in \mathbb{N} | 1 \leq m \leq M\}$. The total period, P_{tot}, describes the entire testing period, which is defined as $P_{tot} := \bigcup_{i=1}^{M} \#P_i$. The sub-period, which is to say the days within a month, is defined as $P_t := \{n \in \mathbb{N} | n_t, \ldots, n_{t+1}\}$where $t \in months$. During this period, we measure each spamicity test out of a possible set of spamicity tests. We shall refer to a particular spamicity test s where $s \in spamicity$ and *spamicity* is defined as the set of all spamicity test, $spamicity := \{BAD\_CREDIT, HELO\_OEM, \ldots\}$. Emails are, for the purposes of this analysis, *only seen as subsets of spamicity tests.* A particular email is referred to within the period of the testing, denoted by e_t, where $t \in P_{tot}$, such that $e_t \subset spamicity$. It is also useful to view a particular spamicity test's frequency on a particular month as a percentage of the total number of emails during this month. The values represented in figures 1, 2 and 3 use the frequency function $f(s,t)$,which is defined as $f(s,t) = \frac{\sum_{i \in P_t} \#(e_i \cap \{s\})}{\#P_t}$.

4.2 Complex Trend

The complex trend "combine different trends or contain high variability" [11]. The complex trend's algorithm would have to identify fluctuations between monthly results and mixed candidate spamicity tests.

The complex trend is predominantly identified by fluctuations between each months proportional appearance. The difference between the percentage of email which contains test s in month n and $n+1$ would have to be measured for the duration of the testing. The cumulative value is represented by the function $c(s)$ defined as:

$$c(s) = \sum_{n}^{M-1} |f(s,n) - f(s,n+1)| \tag{1}$$

The set of all complex spamicity tests C for a given spamicity s is then defined as:

$$C(s) = \{s \in \mathit{spamicity} | c(s) \geq \mathit{min\,bound}\} \cup \{\ \mathit{spamicity} \backslash E \backslash X\}$$

Where E is the set of co-existent spamicity tests and X the set of extinct spamicity test, the definitions of which will follow. The value of $\mathit{min\,bound} = 8.4$, which was determined from the ordered-by-magnitude results of $c(s)$ for all elements of *spamicity*. Values above the *min bound* were found to clearly indicate a significantly increased quantity of fluctuation. This process is elaborated in [3].

4.3 Co-Existence Trend

The second trend, "co-existence, [was] indicated by a sustained population of a strain of spam, particularly through the end of the study period" [11]. The "co-existence group consists of curves that remain flat" [11], indicating that there must be little fluctuation in the month-to-month values. The co-existence trend algorithm was required to identify a consistently sustained population, and react to variations from the sustained population, particularly towards the end of the study period.

In considering co-existence, it was found that grouping certain ranges and assigning a collective value was reasonable. Spamicity tests which were found in $(0\% \ldots 80\%]$ of the emails in a given month were considered viable co-existent candidates. A particular spamicity test's appearance in 80% and above emails for a month was considered a fluctuation, and carried a lesser weighting. Spamicity tests which were not found in a month were negatively weighted, particularly if this occurred in the final month of testing. A failure to appear in the final month resulted in the exclusion of a spamicity test from the co-existent group. The grouping is represented in the bucket function $b(s,t)$ with s being a spamicity test, where $s \in \mathit{spamicity}$, and t is a month in the testing period, where $t \in P_{tot}$. The bucket function is defined as:

$$b(s,t) = \begin{cases} 1 & \text{if } f(s,t) > 0.8, \\ 10 & \text{if } 0.1 < f(s,t) \leq 0.8, \\ 5 & \text{if } 0 < f(s,t) \leq 0.1, \\ -10 & \text{if } f(s,t) = 0, \\ -1000 & \text{if } f(s,t) = 0 \text{ and } t = M \end{cases}$$

The bucket function is then applied to the entire range of the corpus, and each months value is adjusted to give greater weighting to the latter range of the corpus. The co-existence function $e(s)$ for a particular spamicity test is defined as:

$$e(s) = \sum_{n}^{M} \frac{b(s,n)}{(M-n+1)^2} \tag{2}$$

The set of all co-existent spamicity tests, E, is defined as:

$$E \quad = \quad \{s \in \mathit{spamicity} | e(s) > \mathit{accept\,bound}, c(s) \leq \mathit{min\,bound}\}$$

This set excludes all spamicity tests which display a high degree of fluctuation, and are considered complex. The *accept bound* responds to the bucket function, where $\mathit{accept\,bound} = 0$.

4.4 Extinction Trend

The final trend is “extinction, indicated by the population of a strain of spam declining to zero or near zero during the study period” [11]. Extinction presented significant problems in attempts to define a reasonable algorithm, and because of this it is based off the two existing algorithms. The definition requires that extinct spamicity tests identify a consistently sustained population and have no monthly population or decline to a near-zero population.

	Main Corpus		Spam Evolution Study		Relative Difference
Trend	#	%	#	%	%
Co-existent	197	31	64	13	18
Extinction	316	51	236	48	3
Complex	111	18	195	39	21

Table 1: Comparison of the distribution of the spamicity tests amongst the trends.

As has already been shown in section 4.2, a value greater that *min bound* for the $c(s)$ function indicated a high degree of fluctuation in the monthly spamicity test results. Values less than or equal to *accept bound* for the $e(s)$ function indicate a spamicity test which has significantly declined for periods, or is consistently absent. The set of all extinct spamicity test is defined as:

$$X = \{s \in spamicity | e(s) \leq accept\,bound \,\text{and}\, c(s) \leq min\,bound\}$$

5 SPAM EVOLUTION ANALYSIS

A comparison between the *Spam Evolution Study*'s distribution and the distribution of the corpus is shown in table 1. The corpus has approximately 82% of the tests falling under the co-existent and extinct trends. The *Spam Evolution Study* has approximately 61% of the spamicity tests falling under similar trends. The two corpora do not reflect a similar distribution of the spamicity tests outside of the complex trend. The differences between the two corpora's co-existent and extinct trends shows that over a longer period extinction is more prominent than co-existence.

The maximum range for each spamicity test and the average range indicates a correlation between the extinction and complex trends of both corpora. The majority of these two trends are found in the $[0.0 \ldots 0.1)$ range. This is to say that the majority of these tests, which identify the corpus' emails as spam, are dispersed over less than 10% of the corpus emails on average or at a maximum each month. The corpus' co-existence trends, in particular, show a significantly higher proportion located in this low range. This is not in keeping with the *Spam Evolution Study's* co-existence trend, which is dispersed amongst the higher ranges of both the maximum and average spamicity test results.

Assuming that SpamAssassin is able to consistently identify the components of a spam email using its spamicity tests, the locality of the majority of spamicity tests in range could be caused by two conditions. Firstly the types of spam captured are from a large number of spammers, or secondly spammers employ a diverse number of techniques, or both. In either instance the average and maximum distribution suggests a large and varied number of spamicity tests per an email in the corpus. This is reciprocated by further analysis which shows that an average of 8.96 spamicity tests are found for every email in the corpus.

One hypothesis for the dominance of the extinction spamicity tests is a natural extension of the evolutionary metaphor used by *Pu and Webb*. All spamicity tests inevitably tend towards extinction, while some may co-exist for longer periods: their existence relies on their evolving beyond the means of their respective spamicity tests. This evolution implies that the older spamicity test must adjust to these variations, resulting in their older form's extinction. We see a reflection of this behavior in the difference between spamicity test from one version of SpamAssassin and another. The above findings indicate that the trends specified in the *Spam Evolution Study* are relevant to the corpus. There are issues which mitigate these findings in a direct comparison to the *Spam Evolution Study*, which will be discussed. It does, however, hold that the process used by the *Spam Evolution Study* still has relevance in analysing the corpus.

6 DIFFERENCES TO THE SPAM EVOLUTION STUDY

The structure of the corpus was significantly varied from the *Spam Evolution Study's* corpus in two respects: quantity and period. The corpus has an average of 3,385 spam email for each month, while the *Spam Evolution Study* has 38,889 spam emails for each month. 634 Spamicity tests were applied to the corpus, while 495 spamicity tests were applied to the *Spam Evolution Study*'s corpus. If we assume the average of 8.96 spamicity tests per an email applies to both corpora, this would result in the *Spam Evolution Study* being significantly more viable and representative of spam in the wild.

The limited number of sources which make up the corpus, could have unfairly weighted certain tests, favoring specific trends. The *Spam Evolution Study's* use of the SpamArchive project allowed for a significantly more diverse series of sources. The diversity of sources increases the probability of *Pu and Webb's* results reflecting the state of spam in the wild.

The version of SpamAssassin utilised, further reduces the comparative value of this study. The *Spam Evolution Study* does not specify the exact spamicity tests it utilised, however a brief comparison between the spamicity tests of

SpamAssassin 3.1.x and 3.2.x shows significant differences. SpamAssassin 3.1.x contains 795 test and 3.2.x contains 746 tests. Only 383 of the original tests are found in the newer version, which was utilised in this paper.

The specific algorithms utilised by *Pu and Webb* to differentiate between the spamicity trends were not published. Accordingly this paper developed its own algorithms; this is the most significant variation from the *Spam Evolution Study.*

The comparison of this study and the *Spam Evolution Study* is severely limited by the above variations. More specifically the structure of the corpus, the version of SpamAssassin and the trend algorithms introduce a number of limitations to any direct comparison of this paper's results.

7 GEOLOCATION

Geographic location, or geolocation, is the mapping of an IP addresses to a series of geographic co-ordinates. The mapping of an IP address to a country was considered an adequate degree of granularity.

The IP addresses stored in a spam email must be considered unreliable. An RFC 2822 [8] email header should contain a number of *received* fields, in which the IP address of a connecting MTA is stored. Spammers abuse the standard, and often include a number of forged received fields to exploit anti-spam filters. For this reason only the IP addresses associated with connections to reliable MTAs can be trusted. A reliable MTA is defined as the border MTA, which updates an email's header with the first verifiable received field. The border MTA for both corpora was easily determined, although the structure of the anti-spam solutions was such that the connecting IP addresses of non-routable, local and far-side MTAs had to be removed. For example the far-side MTA, which is located in the United States was *one* of the border MTAs for the schools corpus. The border MTA was followed by a number of internal MTAs which append additional received fields. These fields had to be removed from consideration, with only the IP address recorded by border MTA being used for geolocation.

Once an authentic IP address had been obtained, its geolocation had to be determined. The open-source HostIP [6] database was used as a reference, and a customised local implementation was configured to map IP addresses directly to countries. The appropriately selected IP address is then mapped to a country. The email's geographic location is then updated in the database for representation and analysis.

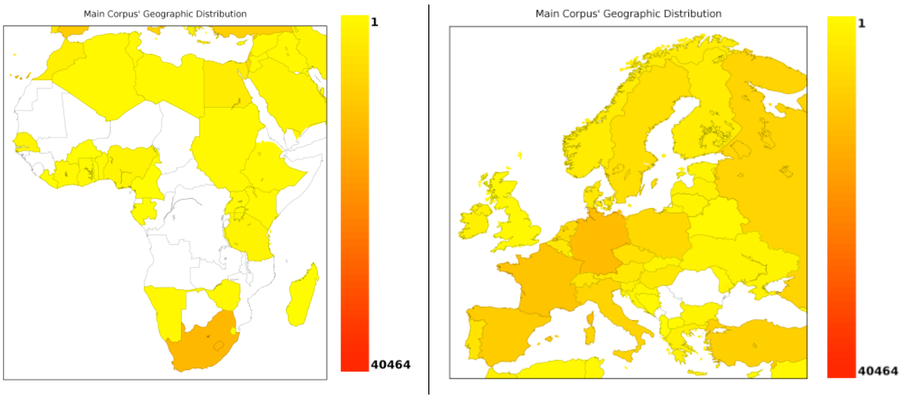
Main Corpus' Geographic Distribution
1
40464
Main Corpus' Geographic Distribution
1
40464

Figure 4: The distribution of the corpus over the African, Europe and the world.

Geolocation data was represented using map projections. A Miller cylindrical map projection was used to graphically display quantitative data. An example of this is the distribution of the corpus' sources of spam from Africa, seen in figure 4.

7.1 Results

The locality of the corpus, with the visualisation of the quantity of spam detected in specific regions, is an excellent tool for the analysis of a spam corpus. The geographic origin of an email was a factor which the original *Spam Evolution Study* was unable to explore due to corpus' structure. This paper uses African and European projections as examples. The top five contributive countries, accounting for the majority of spam in the corpus, are listed in table 2, and should be compared to the projection in figure 4.

Country	# Spam Emails	% of Corpus	Cumulative %
United States	40464	23.904%	23.904%
Taiwan	22359	13.209%	37.113%
United Kindom	17066	10.082%	47.195%
Korea, Republic of	15557	9.190%	56.385%
China	12429	7.343%	63.728%

Table 2: The top five spamming countries in the corpus.

A projection of Africa is shown in figure 4. It is clear that both South African and Egypt are the primary sources of spam in the continent. Most surprising is the lack of spam from central and western Africa, which is the largest continuously populated region in the corpus to be spam-free. Continental Europe is widely dispersed, and was a significant contributor to the corpus. With the exception of Montenegro, Serbia and Romania every country in Europe contributed.

8 CONCLUSIONS

This paper replicated the *Spam Evolution Study*, and presented map projections of the collected spam corpus. A corpus of 169,274 spam emails was collected. The corpus was analysed using SpamAssassin and a distributed processing network. The results were further evaluated by dividing each tests into the three trends of co-existence, extinction and complex. These trends were formalised as an extension to the original study. The trends were found to be applicable to the corpus, however a number of variations from the *Spam Evolution Study* reduced the comparative value of these findings. Geographic projections were created, using data collected from the corpus and findings detailed.

9 FUTURE WORK

The corpus is limited to emails which have been distributed to South African MTAs. This underutilised the distributed architecture which was specifically designed to handle a significantly larger corpus. One of the early limitations of this study is the relatively small scale of the corpus when compared to other studies [7, 11, 12]. Future research into the effects of geolocation on the evolution of spam construction would be benefited by applying this study on a substantially larger and wider ranging corpus. A closer analysis of particular provinces and states within countries could be performed.

The linking of the developed state of a country to the quantity of spam it produces would be a particularly challenging and interesting extension. An extension of this study could be conducted on further research into selecting the grouping of the various geographic locations of identified spam. One interesting possibility would be the use of spam construction techniques to probabilistically determine the identity and locations of botnets.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of our colleagues in the Department of Computer Science. We would also like to acknowledge the support of Telkom SA, Business Connexion, Comverse

SA, Stortech, Tellabs, Amatole, Mars Technologies, openVOICE and THRIP through the Telkom Centre of Excellence in the Department of Computer Science at Rhodes University.

REFERENCES

[1] James Carpinter and Ray Hunt. Tightening the net: A review of current and next generation spam filtering tools. *Computers & Security*, 25(8):566–578, November 2006.

[2] Deborah Fallows. Spam: How it is hurting email and degrading life on the internet. Technical report, PEW Internet & American Life Project, October 2003.

[3] Blake Friedman. A Formalised Replication and Extension of Observed Trends in Spam Construction Techniques: A Case Study of Spam Evolution. Honours Dissertation, Rhodes University, November 2007.

[4] Tom Gillis. Internet security trends for 2007: A report on spam, viruses and spyware. Technical report, IronPort, 2007.

[5] Joshua Goodman, Gordon V. Cormack, and David Heckerman. Spam and the ongoing batle for the inbox. *Communications of the ACM*, 50(2):25–33, February 2007.

[6] Simon Gornal. Hostip.info. http://www.hostip.info/, 2007.

[7] Geoff Hulten, Anthony Penta, Gopalakrishnan Seshadrinathan, and Manav Mishra. Trends in spam products and methods. In *CEAS 2004 - First Conference on Email and Anti-Spam*, 2004.

[8] J. Klensin. Simple Mail Transfer Protocol (SMTP). RFC 2821, April 2001.

[9] Justin Mason. SpamAssassin. http://spamassassin.apache.org/, 2007.

[10] Messaging Anti-Abuse Working Group. Email metrics program: The network operators' perspective. 3rd and 4th Quarters 4, MAAWG, March 2006.

[11] C. Pu and S. Webb. Observed trends in spam construction techniques: A case study of spam evolution. In *CEAS 2006 - Third Conference on Email and Anti-Spam*, July 2006.

[12] B. Taylor. Sender reputation in a large webmail service. In *CEAS 2006 - Third Conference on Email and Anti-Spam*, July 2006.