Filtering Spam E-mail on a Global Scale

Geoff Hulten
Microsoft
One Microsoft Way
Redmond, WA 98052
(425)705-8692
ghulten@microsoft.com

Joshua Goodman
Microsoft
One Microsoft Way
Redmond, WA 98052
(425)705-2947
joshuago@microsoft.com

Robert Rounthwaite
Microsoft
One Microsoft Way
Redmond, WA 98052
(425)706-9791
robertro@microsoft.com

ABSTRACT

In this paper we analyze a very large junk e-mail corpus which was generated by a hundred thousand volunteer users of the Hotmail e-mail service. We describe how the corpus is being collected, and analyze: the geographic origins of the e-mail; who the e-mail is targeting; and what the e-mail is selling.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues – *abuse and crime involving computers, transborder data flow, privacy.*

General Terms

Measurement, Economics, Legal Aspects.

Keywords

Junk E-mail, spam, international e-mail.

1. INTRODUCTION

This paper describes a massive corpus containing over two million hand-classified e-mail messages that were sent to Hotmail accounts between April and June of 2003. The data came from the Hotmail Feedback Loop: a mechanism that allows over a hundred thousand randomly selected Hotmail users to give feedback about which of their messages are good and which are spam. In particular, every day we randomly select one message from the mail stream of each feedback loop user and ask the user to classify it for us. Thanks to the Feedback Loop users, we currently receive tens of thousands of classified messages every day. We have carried out a series of analyses on this data. We are only aware of one other large scale study of spam, the FTC report on false claims in spam [1]. Our study differs from this one in several ways. Perhaps most importantly, our data was collected by randomly sampling over the entire mail stream, rather than by relying on users to report e-mail that offended them. This allows us to include a large sample of good mail in our analysis, and also mitigates the problem of biased sampling (anecdotally, the probability with which a user reports a spam e-mail is proportional to how offensive the user found the content of the e-mail). Our study also differs from the FTC study by focusing on the geographic origins of e-mail, while theirs focused on false claims.

2. The Geographic Origins of E-mail

Each of the messages that arrived via the Feedback Loop was tagged with the IP address of the computer that connected to Hotmail to deliver the message.

In order to determine the geographic origins of these messages, we gathered data from the four major entities responsible for allocating IP ranges: ARIN, APNIC, LACNIC, and RIPENCC. Their data contains a record for every IP allocation that describes (among other things) which IP addresses were included in the allocation, the date of the allocation, and the country where the entity that received the allocation is based. Using this data we were able to determine the country to which over 99% of the IP addresses that sent mail into our data set were allocated. It is important to note, however, that there are at least two ways that the country of allocation of an IP address can be different from the country of origin of the e-mail. First, allocations can be transferred from entity to entity (even over country borders) without notifying ARIN, APNIC, LACNIC, or RIPENCC – so the data we got from them may be out of date. Second, the computer that delivers a message to Hotmail is not necessarily the same computer that the mail was sent from: it is often a mail server at an ISP, or some other intermediary. Despite these caveats, we believe this data gives us a very good snapshot into the geographic nature of spam.

The two million e-mail messages in our data set came from 214,000 distinct IP addresses. These IP addresses were allocated to 157 different countries (although 67 of these countries each sent less than 100 messages into our data set). Figure 1 shows the volume of e-mail by country superimposed on a map of the world. The USA was the largest sender by far, accounting for just under half of the total e-mail volume. The rest of the Americas accounted for 8% of the mail; Europe for 21%; Asia for 17%; and Africa for just 0.2% of the mail.

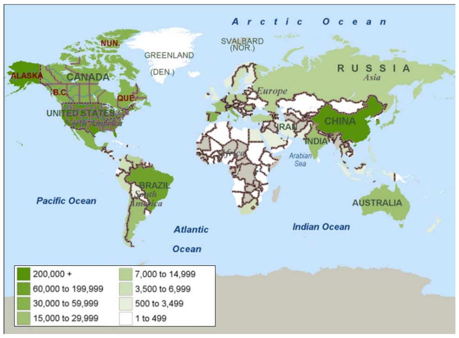

Figure 1 : Volume of E-mail by Country

Figure 2 shows the nature of the mail from the fifteen countries that sent the highest volume of e-mail into our data set. For example, about 53% of the spam and 59% of the good mail came from the USA, while about 15% of the spam and 2.5% of the good mail came from China.

WWW 2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

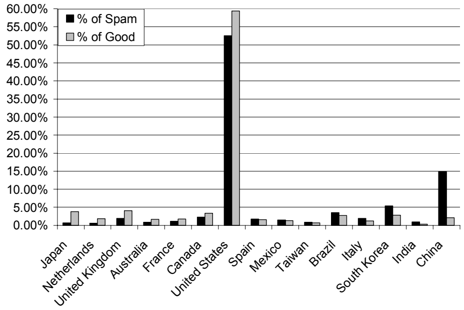

Figure 2 : Portion of "good" versus spam e-mail from the 15 countries that sent the highest volume of e-mail

There is a great deal of diversity in the ratio of good e-mail to spam e-mail received from various countries, and Figure 3 depicts this graphically. Countries that sent relatively more spam e-mails per good e-mail are colored in oranges, and countries that sent relatively more good e-mails per spam are colored in greens. In general Western Europe, Japan, and New Zealand sent more good e-mail than spam e-mail to Hotmail users; while Asia, The Middle East, and Africa sent more spam than good mail to Hotmail users.

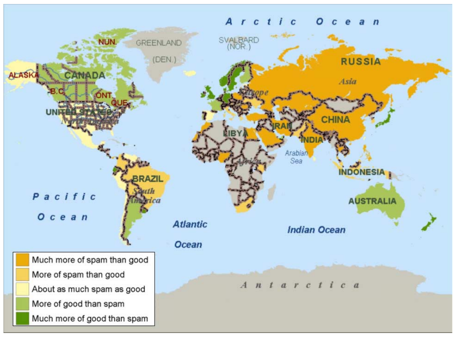

Figure 3 : Purity of E-mail by Country

3. The Languages of E-mail

About half of the e-mail in our data set came from servers outside of the United States, and we wanted to determine what language this international mail was written in. We used the character set of the messages as a surrogate for language detection. For example, mail with Japanese characters in it (and thus written in Japanese) will be encoded in the *ANSI/OEM - Japanese Shift-JIS* character set, and most English mail will be encoded in the *US-ASCII* character set. Table 1 contains some details about the five most common character sets in our data set.

Table 1 : Five Most Common Character Sets

Character Set	% of Good	% of Spam
ISO Latin (Spanish)	1.41%	0.31%
ANSI/OEM (Korean)	2.33%	0.45%
ANSI/OEM (Japanese)	4.16%	0.18%
ANSI Latin (Spanish)	29.18%	7.52%
US-ASCII (English)	60.31%	90.80%

3.1 What Spammers are Selling

In the previous two sections we found that a large potion of the spam being sent to Hotmail comes from outside of the United States, but is written in English. Intuitively, much of this spam is coming from businesses selling products and services to consumers in the United States from businesses based in foreign countries. This is somewhat troublesome, as it will be more difficult to affect the behavior of such businesses with legislative solutions than to affect the behavior of fully domestic businesses. To further evaluate the ability of spammers to move their operations internationally, we examined a sample of the spam from our data set. We identified the type of product or services promoted by each of the spam in our sample and categorized these as:

- ***Domestic***
  Products or services that require a domestic presence to sell, such as: financial services, insurance, government grant programs, and items we deemed too expensive to ship internationally. Legislation has the potential to greatly discourage or even stop the use of spam to promote such products.

- ***Semi-domestic***
  Products that require shipping but which we deemed small enough to be shipped from nearby countries such as Canada, Mexico, etc. Such products include Viagra and other medical products, college diplomas, magazines, etc. Legislation in the United States has the potential to stop domestic businesses from promoting such products with spam, but may not be able to discourage international ones.

- ***International***
  Products or services that do not require physical shipping or a domestic presence. These include porn sites, software, and scams. It will be difficult for legislation to affect this type of spam.

We categorized 30% of the spam in our sample as domestic; 32% as semi-domestic; and 38% as international—that implies that 70% of the spam in our data set could be sent from international locations (thus potentially avoiding U.S. legislation). It is also interesting to note that about 16% of the spam in our sample was advertising for pornographic web sites.

4. Summary

This paper reported on an analysis of over two million messages that were sent to Hotmail accounts between April and June of 2003. In it we found that about half of the spam and 40% of the good mail comes from sources outside the U.S.; that countries are very diverse in the ratio of spam mail to good mail that they send; that most of the spam is in English (or at least in the character set that is indicative of US English text); that over two thirds of the businesses that send spam could be run with no domestic presence in the United States.

4.1 Acknowledgements

Thanks to Carol Brown, Nathan Howell, Anthony Penta, and John Mehr for helping us identify useful tools and data sources. And thanks to the Hotmail team for setting up the Feedback Loop, especially Pablo Stern, Eliot Gillum, and Raj Pai.

5. REFERENCES

[1] FTC Division of Marketing Practice. False claims in spam. http://www.ftc.gov/opa/2003/04/spamrpt.htm. April 30, 2003