A Taste of Honey – UCE (Spam) Reduction Through Deception

Nick Wallingford
Bay of Plenty Polytechnic
Tauranga, New Zealand
nick.wallingford@boppoly.ac.nz

1 ABSTRACT

Keywords: Spam, email, UCE, honeypots, spambot

"Honeypots" are being used as one effort to reduce Unsolicited Commercial Email (UCE, also known as spam). Email users, as well as organised groups of users, set up 'trap' email accounts to attract UCE to identify sources and nature of the mail received.

All email users are affected by UCE, through

<bi> cluttered inboxes,

<bi> network congestion, and

<bi> cost of transport involved in delivery.

Using honeypots reveals some of the methods that spammers obtain email addresses, as well as identifying some of the steps that can be taken to avoid having one's address become the target of UCE. An understanding of "How did they get my email address?" is of interest and value to all users, as well as being part of the process of UCE reduction.

This paper describes efforts being made by individuals and organised projects to utilise honeypots to better identify the means by which spammers obtain email addresses.

2 INTRODUCTION

It is a rare Internet email user who receives no unwanted commercial email (UCE), or spam, at all. For most users it has become both a significant problem and a topic that will generate strong debate, including such comments as:

"The uncontrolled proliferation of spam is taking one of the most important new forms of communication and killing its effectiveness" (Oliva, 2004), and

"Spamming is the scourge of electronic-mail and newsgroups on the Internet. It can seriously interfere with the operation of public services, to say nothing of the effect it may have on an individual's e-mail system." (Cerf, V., cited by Cournane and Hunt, 2004).

Cournane and Hunt (2004) categorised the problems created by UCE as:

<bi> Cost shifting – the recipients of email are forced to pay the costs of delivery that the advertiser has avoided,

<bi> Fraud – misleading subject lines (to encourage a user to open what would otherwise be deemed to be unwanted email) and misrepresentation of the origin and routing of messages,

<bi> Resource wastage – network congestion created by the routing and delivery of UCE,

<bi> Displacement of legitimate mail – overfull inboxes that exceed size limits set may mean 'real' mail is rejected and lost, and

<bi> Black lists – the banning of servers and domains may impact on users who were not necessarily responsible for the abuse of email systems.

For many email users, the issue is emotive and highly-charged – they simply wish it would all go away, along with the people who produce and distribute the unwanted emails.

This paper describes one component of on-going efforts to deal with UCE by technical means: the use of honeypot addresses to attract UCE with the longer term aim of reducing the overall amount of unwanted emails being distributed.

3 PROJECT HONEYPOT

Biever (2005) describes Project Honeypot well:

"Webmasters who want to help fight spam can download Project Honey Pot's software, which is designed to turn their website into a magnet for harvesters. If the site detects that a crawler is visiting it the software generates a fake email address for the crawler to grab, and records the address of the crawler and the time and date."

"The fake address then vanishes from the site, but remains valid as a mailbox. Because it

is a fake, no one will send it legitimate mail. If any mail arrives it can only have come from the spammer who grabbed it off the Honeypot site, and this fingers the computer that crawled the site as belonging to the spammer."

The project is then able to provide both individual site and collective statistics on the numbers of email address harvesters that have visited the site and the quantity of UCE that resulted from the harvesting.

The results are shared with anti-spam developers and researchers with the intent that it will assist in the development of tools to ultimately reduce the quantity of UCE. (Project Honey Pot, 2005)

4 A PERSONAL HONEYPOT BY THE AUTHOR OF PEGASUS EMAIL

David Harris, author of the Pegasus Mail and Mercury Mail Transport programs, utilises a honeypot to attract, identify and ultimately 'blacklist' spammers.

In the footer of his webpages, he includes a simple '|' character that has an email link to the address "shibboleth@pmail.gen.nz". Alternative text that displays if a mouse moves over the character warns "Never, ever use this link – it is a honeypot address".

Harris is a self-proclaimed lover of words – his use of 'shibboleth' for the email account carries an intentional irony. A shibboleth is a word or phrase that by its pronunciation or use indicates that a person is a member of a particular group (Kemmer, 2004). In this case, the use of the address reveals anyone who writes to it to be a spammer!

Harris reports that within 1½ hours of the time he first placed the honeypot address on his website, he began to receive UCE at the address. He currently receives approximately 30,000 email deliveries per month.

Harris chooses to reject the deliveries before they occur, and adds the sender to his 'blacklist', refusing to accept any further mail from that server/address., or he accepts the UCE [Harris, personal communication].

5 RESEARCH UNDERTAKEN

5.1 Overview

This research was specifically directed at the potential for addresses on a website to be harvested as targets for UCE. It investigates several of the methods that can be used to restrict the harvesting, but does not attempt to categorise content or identify source of the UCE that was received.

Three particular aspects were examined:

<bi> How email addresses appearing on websites are harvested,

<bi> How dictionary attacks are used to generate spam, and

<bi> How Project Honey Pot operates to identify spam.

5.2 Email addresses on websites

One of the primary means of obtaining email addresses to use for UCE is by taking them from where they appear on websites.

The software used for this harvesting, sometimes called 'spambots', crawl from one webpage to another, collecting and collating anything that appears to be a validly-formatted email address.

Addresses that appear on websites are there to make it easy for website users to address emails, allowing the user to simply click on the link rather than having to type in an email address into their email client program. It is this convenience, however, that leads to the majority of address harvestings.

5.2.1 Munging and Obfuscation

Munging is a term referring to either making an email address technically invalid, but still potentially useable by the website visitor, or by a process of obfuscating the address in some way such at it appears visually and performs technically as expected, but is not likely to be picked up by the automated harvesting software.

In this research, addresses were 'munged' by replacing the @ symbol in the address with the word AT (with the expectation that a human would reinstate the symbol before using the address to send email).

Addresses were also obfuscated through the use of HTML entity substitution and hexcode entity substitution. In each of these methods, the address when rendered by an Internet browser will still appear 'normal'. The coding that generates the address, however, consists of a string of characters that may (hopefully) not be recognised by address harvesting software.

5.3 Dictionary Attacks on Common Names

Another means of obtaining email addresses targets email accounts with common names, or uses a brute force method to find variations on those names.

Delio (2003) refers to users stating that "… within a day of creating a new Hotmail account the spam starts flowing in", blaming dictionary attacks for having harvested the addresses.

Dictionary attacks involve the submission of a large number of random email addresses to a mail server, recording which are "live" based on the server's response. Common names (john@domain.com) and variations on those (john01@domain.com, john02@domain.com) are typically targeted by the software, acknowledging the increased likelihood of success with those email address formats.

Cook (2004) described such an attack from the viewpoint of recipient of the 'catch all' mail account – the account that any misaddressed mail to the domain is delivered. He suggests that in some cases, it may be that a spammer's list of addresses has been inflated by simply making up account names before on-selling the list of email addresses.

For this research email accounts were created for the 20 most common first names, 10 male and 10 female. (Lusby, 2005). In fact, the names are those that appear most often within the latest US census; no attempt was made to incorporate names from other cultures or countries.

5.4 Project Honey Pot

Registering a site with Project Honey Pot involves creating a webpage that will contain code provided by the project and registering that page's Universal Resource Indicator (URI, sometimes referred to as a URL) with the project.

As part of the research, a page on the server used for the research was set up and monitored for harvesters' visits and UCE received at the addresses it promulgated to the Internet.

6 MATERIALS AND METHODS

6.1 Addresses on Website

The BCS server (http://www.bcs.net.nz) used to support the degree programme at BOP Polytechnic, was used to create a series of email accounts.

Twenty email accounts were created using common first names. Those email addresses were not placed on any webpages or advertised in any way. Any mail that came to them would have been generated simply with the expectation that there might be an account with that name on the server involved.

A further 18 accounts were created with randomly generated names of the form 'aaa###' where 'a' is a random letter A-Z and '#' is a random digit 0-9.

Three of the 18 addresses were not advertised on any website in order that they might remain as controls.

The remaining addresses, in replicates of three, were placed in the footer of each page of the BCS server.

'As is' - Three of the email addresses were simply written into the text of the footer, such as: xsw572@bcs.net.nz.

'mailto:' - Three were included as properly formatted mailto: links, such as Email.

'Munged with AT' - Three were 'munged' by replacing the @ character with AT in a mailto: link, such as Email.

'HTML encoded' - Three were obfuscated with HTML entities, obtained using an online generator (Neumüller, 2005). While long and unreadable to the eye, when the page was viewed with a browser, they appeared and acted like normal mailto: links. An example was: <a href="mailto:gyo553@bcs.net.nz">Email</A>.

'Hex encoded' - The final 3 addresses were similarly obfuscated using the same online tool, but were instead encoded with hexcode entities. An example was: Email.

6.2 Project Honey Pot

A page on the BCS server was set up using the instructions provided by Project Honey Pot. The page was named with a non-obvious name (studentlist.php) and links to the page were made from the footer of each of the pages on the BCS website.

Monitoring of the results was done through the Project Honey Pot website by logging into the account that had been created.

7 RESULTS

Unwanted emails began to arrive at the test account addresses four days after they were first advertised on the website. For most of the accounts, UCE has continued every two to three days, even though the advertising of the addresses stopped after 25 days, and a further 29 days have passed since then.

7.1 Munging and Obfuscation

Of the methods used in the research, 'munged with AT' and 'Hex encoded' were the only two that did not receive any UCE during the term of the research.

Most of the other methods resulted in a similar level of UCE received. All of the 'As is' and 'mailto:' addresses had similar results, with the number of messages from 36-40, and the total size for each account ranging from 209kb to 284kb.

Each of the addresses 'HTML encoded' received one message – with the same sender, subject and message body. While HTML encoding of email addresses to appear on websites is described as effective, these results would indicate otherwise.

7.2 Dictionary Attacks on Common Names

No mail was received by any of the 20 'common name' accounts, indicating that no specific dictionary attack was apparent to target the mail server being utilised.

During the same period, however, the server was under attack with attempts to login to SSH (Secure Shell server) using a series of common names. In one 24 hour period, 384 attempts were made, using a series of common first names, hoping there might be accounts with those names that had insecure passwords. These dictionary attacks on the SSH server targeted 8 of the 20 names that were being used in the research.

7.3 Summary of Addresses on Website

Treatment	Number	Size(kb)
Control	0	0
Common names	0	0
As is	36-40	209-284
Mailto:	35-37	207-239
Munged with AT	0	0
HTML encoded	1	7
Hex encoded	0	0
Total	**233**	**1.526MB**

7.4 Project Honey Pot

During the time of the research, only one suspected spambot email address harvester visited the page set up for Project Honey Pot on the BCS server. Of the 27 email addresses that had been presented on that page through the course of the research, only 1 had been the recipient of UCE, according to the Project Honey Pot statistics.

With the first UCE arriving 10 days after first setting up the Project Honey Pot page, the BCS server would appear to have been 'found' earlier than the Project's average of 29 days.

8 CONCLUSIONS

Honeypots can be used effectively to attract UCE with the purposes of identifying sources, analysing the nature of the emails or simply blacklisting the mail servers that deliver it.

This research examined aspects of email address harvesting from webpages and dictionary attacks on common names. Other means such as the misuse of email addresses collected as part of web-based registration services may also contribute to the collection process required by the originators of spam.

Email users should be particularly wary of allowing their email address to appear on webpages in any form that might leave them open to harvesting by spambots.

9 REFERENCES

Biever, C. (2005). Project Honeypot to trap spammers. Scientific American 185(2485): 26.

Cook, C. (2004). On-going dictionary attack. [Available online: http://geek.focalcurve.com/archive/2004/06/on-going-dictionary-attack. Retrieved 13 April 2005.]

Cournane, A. and Hunt, R. (2004). An analysis of the tools used for the generation and prevention of spam. *Computers and Security, Elsevier, UK, 23(2): 154-166. March 2004, pp154-166.*

Delio, M. (2003). Hotmail: A spammer's paradise? [Available online: http://www.wired.com/news/infostructure/0,1377,57132,00.html]. Retrieved 19 April 2005.]

Harris, D. (2005). Pegasus mail. [Available online: http://www.pmail.com]. Retrieved 19 April 2005.]

Kemmer, C. (2004). The story of the Shibboleth. [Available online: http://www.ruf.rice.edu/~kemmer/Words/shibboleth.html. Retrieved 19 April 2005.]

Lusby, A. (2005). The top 100 most common first names for both genders. [Available online: http://www.namestatistics.com/list.php?type=firstall. Retrieved 10 May 2005.]

Neumüller, R. (2005). Anti mail spam. [Available online: http://www.katpatuka.org/pub/doc/anti-spam.html. Retrieved 14 March 2005.]

Oliva, R.A. (2004). Spam! – Separating the good from the bad and ugly. *Marketing Management, 13(1):* 50-52.

Project Honey Pot. (2005). How to avoid spambots. [Available online: http://www.projecthoneypot.org/how_to_avoid_spambots.php. Retrieved 13 April 2005.]