Email Address Harvesting and the Effectiveness of Anti-Spam Filters

A Report by the Federal Trade Commission's Division of Marketing Practices

November 2005

I. Overview

This report describes the results of a Federal Trade Commission ("FTC") staff study of three aspects of spam in the current Internet environment. First, the study explored the current state of email address harvesting – the automated collection of email addresses from public areas of the Internet. The study found that addresses posted on websites were at risk of being harvested by spammers, but that addresses posted in chat rooms, message boards, USENET groups and weblogs ("blogs") were far less likely to be harvested. Indeed, some chat room operators took proactive measures to prevent the harvesting of email addresses posted by FTC staff.

Second, the study explored the effectiveness of spam filtering by Internet Service Providers ("ISPs"). The study showed that the anti-spam filters utilized by two free web-based ISPs effectively blocked the vast majority of spam sent to harvested addresses. The implication of this finding is that ISP spam filtering technologies are substantially reducing the burden of spam on consumers. Nevertheless, spam sent to harvested addresses imposes costs on ISPs receiving the spam.

Third, the study measured the effectiveness of using "masked" email addresses as a possible technique in preventing harvesting. The "masking" of an email address involves altering the appearance of an email address so that it is understandable by a person who sees the address, but less likely to be discernable by automated harvesting software. For example, to mask an unmasked email address such as "johndoe@ftc.gov," the words "at" and "dot" can be written out, and segments of the email address can be separated by spaces. The masked version of the address would appear as "johndoe at ftc dot gov." The study found that the "masking" of an email address was very effective in thwarting harvesting.

II. Background

A 2002 study conducted by FTC staff and state law enforcement authorities found that the harvesting of email addresses from public areas of the Internet was widespread.[1] A year later, Congress reached the same conclusion when it made address harvesting an aggravated violation of the Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003 (the "CAN-SPAM Act").[2] This report examines the current validity of these findings and analyzes whether anti-spam filters substantially reduce the impact of email address harvesting.

III. Methodology

To measure the prevalence of harvesting and the effectiveness of two major ISPs' anti-spam filters, FTC staff created 150 new undercover email accounts.[3] FTC staff established 50 of these email addresses at an ISP that employs no anti-spam filtering technologies (the "Unfiltered Addresses") and 50 addresses at each of two free web-based ISPs that pass incoming email through anti-spam filters ("Filtered ISP 1" and "Filtered ISP 2").[4]

FTC staff then posted sets of three of these newly-created email addresses – consisting of an Unfiltered Address, an address at Filtered ISP 1, and an address at Filtered ISP 2 – on 50 Internet locations. The 50 Internet locations included websites controlled by the FTC[5] and several popular message boards, blogs, chat rooms, and USENET groups which had high hit/visit rates, according to ranking websites such as www.message-boards.com and Google popularity searches.[6] All of the 150 addresses were posted during a three day period in July 2005.

Graphic 1

Locations On Which Email Addresses Were Posted

Type	Number
FTC Website Pages	12
Message Boards	12
Blogs	12
Chat Rooms	12
USENET Groups	2

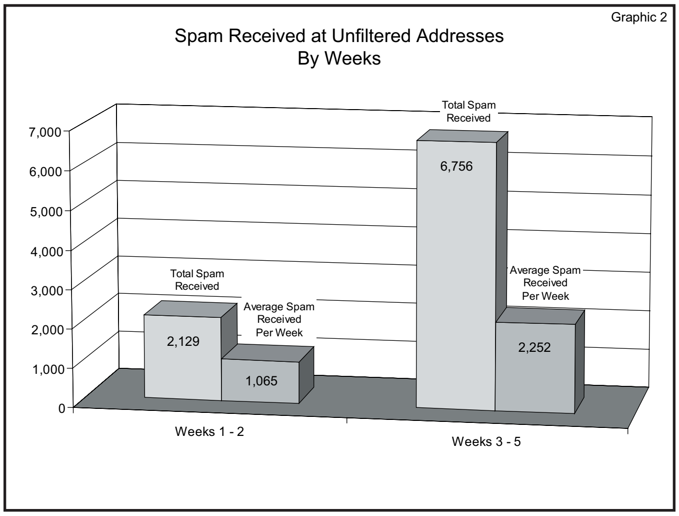
After a two week period, and again three weeks later (after a five-week period), FTC staff tallied the total number of spam messages in the inbox of each of the email accounts. The receipt of messages by the Unfiltered Addresses indicated whether harvesting had occurred. It also indicated whether the posting of these email addresses on different types of Internet locations – such as websites, chat rooms, message boards, blogs, or USENET groups – resulted in different levels of harvesting. In addition, because at each site the FTC staff had posted a triad of email addresses – one from each of the three groups we had created (Unfiltered ISP, Filtered ISP 1 and Filtered ISP 2) – FTC staff was able to calculate the percentage of spam messages blocked by the two ISPs' spam filters by comparing the number of messages received in each of the Unfiltered Addresses to the number of messages received in Filtered ISP 1 and in Filtered ISP 2.[7]

To measure the effectiveness of masking techniques in preventing harvesting, FTC staff created four additional masked email addresses at the ISP that employs no anti-spam filters (“four masked Unfiltered Addresses”). FTC staff then posted these four masked addresses, alongside four unmasked email addresses (“four unmasked Unfiltered Addresses”), on four websites. To mask an email address, FTC staff wrote out the words “at,” “dot,” and “com” and separated each segment of the email address with spaces. For example, using this masking technique, the unmasked email address “johndoe@ftc.gov” would appear in its masked version as “johndoe at ftc dot gov.” After a two week period, and again three weeks later (after a five week period), FTC staff tallied and compared the total number of spam messages received by the masked and unmasked addresses posted on the same four websites.

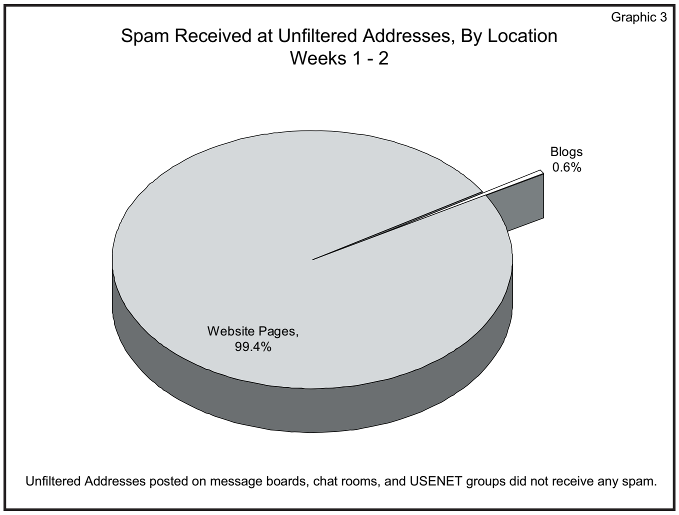
IV. Findings

A. Harvesting Is Still Observed, But Some Sites Appear More Vulnerable Than Others

Spammers continue to harvest email addresses. At the conclusion of the two week study period, the 50 Unfiltered (and “unmasked”) Addresses had received a total of 2,129 pieces of spam. At the conclusion of the five week study period, these same addresses had received 8,885 pieces of spam. The total weekly amount of spam sent to the Unfiltered Addresses more than doubled from weeks one and two to weeks three through five. In weeks one and two of the study, the 50 Unfiltered Addresses received an average of 1064.5 messages per week. In weeks three through five, these same addresses received an average of 2,252 messages per week.



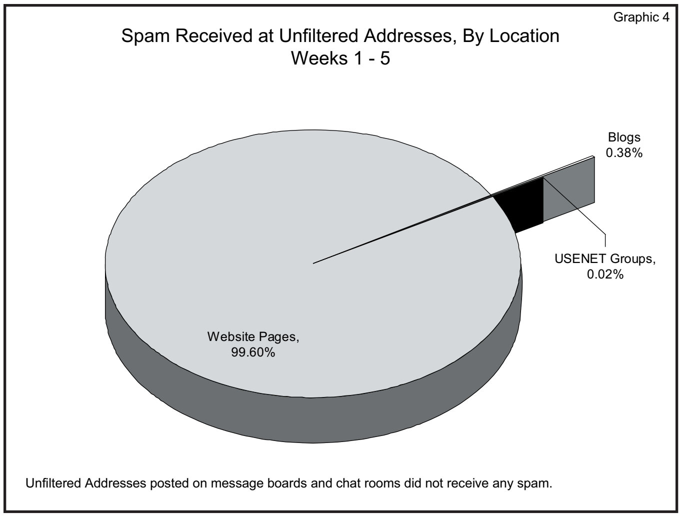
At the conclusion of both the two week and five week study periods, email addresses posted on particular types of Internet locations – such as websites – were far more likely to be harvested than email addresses posted on other types of Internet locations – such as message boards, chat rooms, blogs or USENET groups. Indeed, nearly all of the spam received was received by the Unfiltered Addresses that we had posted on website pages. At the conclusion of the two week study period, 99.4 percent of the total amount of spam received were received by Unfiltered Addresses posted on 11 of the 12 website pages, and only 0.6 percent of the spam received were received by addresses posted on two of the 12 blogs. By contrast, the Unfiltered Addresses that had been posted on 12 message boards, 12 chat rooms, and two USENET groups, received no spam at all.[8]


Graphic 3

Spam Received at Unfiltered Addresses, By Location
Weeks 1 - 2

Unfiltered Addresses posted on message boards, chat rooms, and USENET groups did not receive any spam.

At the conclusion of the five week study period, the results were similar. Of the 8,885 total spam messages received by the Unfiltered Addresses during the five weeks, 99.6 percent of the total amount of spam received were received by Unfiltered Addresses that had been posted on 11of the 12 website pages, whereas only 0.38 percent of the spam messages were received by addresses posted on two of the 12 blogs, and 0.02 percent of the spam messages were received by one address posted on one of the two USENET groups.[9] By contrast, the Unfiltered Addresses that had been posted on 12 message boards,12 chat rooms, and 10 of the 12 blogs, received no spam at all.[10]

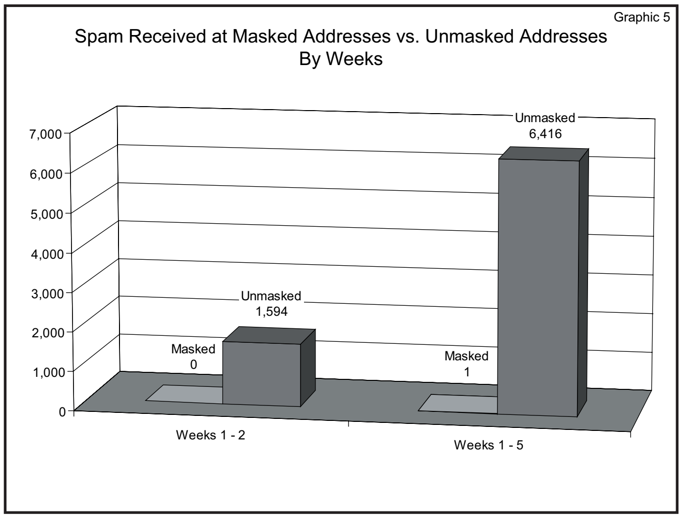

Graphic 4

Spam Received at Unfiltered Addresses, By Location
Weeks 1 - 5

Unfiltered Addresses posted on message boards and chat rooms did not receive any spam.

B.   Harvesting Can Be Prevented

When FTC staff posted email addresses in chat rooms, operators of some of these rooms quickly removed the postings because they violated the chat rooms' policies against the posting of personally-identifiable information. Morever, the operator of one USENET group in which FTC staff posted email addresses automatically masked the addresses by replacing some of the characters in the email addresses so that the true addresses would not appear in the USENET posting. Vigilance by these chat room and USENET group operators ensured that posted addresses were not harvested. It should be noted, however, that because the FTC's study selected popular USENET groups and chatrooms with high hit/visit rates, no conclusion can be drawn from these results as to whether less frequently visited groups and chatrooms would be as vigilant in preventing harvesting.

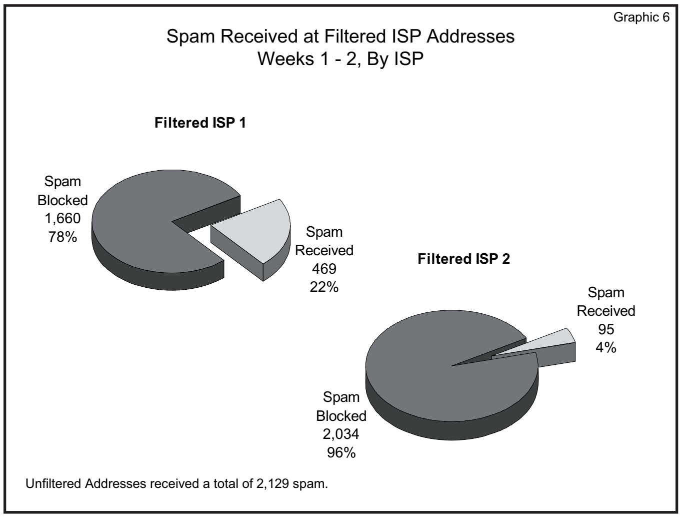

Graphic 5

Spam Received at Masked Addresses vs. Unmasked Addresses
By Weeks

At the conclusion of the two week study period, the four masked Unfiltered Addresses posted by FTC staff received no spam, while the four unmasked Unfiltered Addresses posted on the same four websites received 1,594 spam messages. Similarly, at the end of the five week study period, the four masked Unfiltered Addresses received one spam message, while the four unmasked Unfiltered Addresses posted on the same websites received a total of 6,416 spam messages. Thus, it appears that consumers who must post their email addresses on websites can use masking techniques to significantly reduce the risk of harvesting.[11]

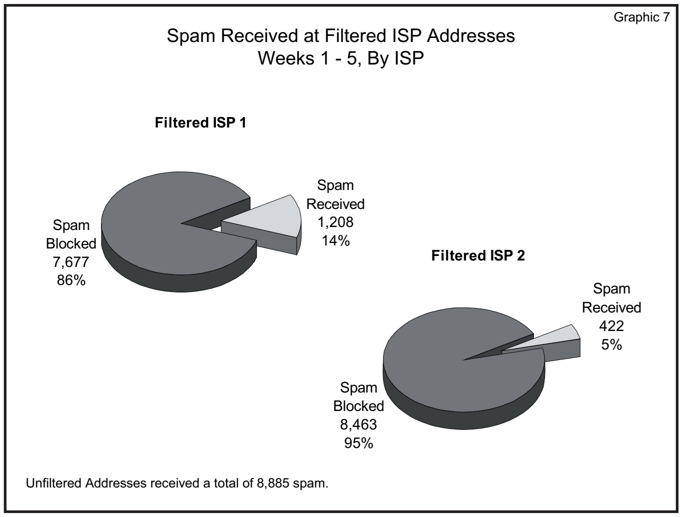
C. Spam Filters Prevent The Delivery Of The Vast Majority Of Spam

Although FTC staff posted the 50 Unfiltered Addresses, the 50 addresses at Filtered ISP 1, and the 50 addresses at Filtered ISP 2 on the same 50 locations on the Internet, the Unfiltered Addresses received dramatically more spam than the addresses located at the two Filtered ISPs. After the two week study period, the 50 Unfiltered Addresses received a total of 2,129 spam messages, while the 50 addresses established at Filtered ISP 1 received 469 messages, and the 50 addresses established at Filtered ISP 2 received 95 messages. Thus, after two weeks, Filtered ISP 1 effectively prevented 78 percent of spam emails from entering its users' inboxes, and Filtered ISP 2 blocked 96 percent of spam messages.[12]


Graphic 6

Spam Received at Filtered ISP Addresses
Weeks 1 - 2, By ISP

Unfiltered Addresses received a total of 2,129 spam.



After five weeks, the results were similar. While the 50 Unfiltered Addresses had received a total of 8,885 spam messages, the 50 addresses opened at Filtered ISP 1 had received a total of 1,208 messages, and the 50 addresses opened at Filtered ISP 2 had received 422 messages. Thus, at the conclusion of the five week study period, Filtered ISP 1 effectively prevented 86.4 percent of spam messages from entering its users' inboxes, and Filtered ISP 2 blocked 95.2 percent of spam messages.

V. Conclusion

This study indicates that spammers continue to harvest email addresses posted on websites, and, to a much lesser extent, those posted on blogs and USENET groups. But email addresses posted by FTC staff in popular message boards and chat rooms were not harvested. It also appears that consumers who must post their email addresses on websites can substantially reduce the risk of harvesting by masking their addresses.

Notably, the fact that the vast majority of spam sent to harvested addresses in this study was never delivered to consumers' inboxes demonstrates the relative effectiveness of the two ISPs' spam filters. This encouraging result suggests that anti-spam technologies may be dramatically reducing the burden of spam on consumers.

Endnotes

1. "Email Address Harvesting: How Spammers Reap What You Sow." http://www.ftc.gov/bcp/conline/pubs/alerts/spamalrt.htm.

2. Under the CAN-SPAM Act, one who initiates the transmission of an email message that violates the Act may be subject to triple damages if the violative email is sent to an email address that he or she knows was harvested. 15 U.S.C. §§ 7706(f)(3)(C) and (g)(3)(C). In passing the CAN-SPAM Act, Congress found that "[m]any senders of bulk commercial electronic mail use computer programs to gather large numbers of electronic mail addresses on an automated basis from Internet websites or online services where users must post their addresses in order to make full use of the website or service." 15 U.S.C. § 7701.

3. Each of the 150 email accounts had an email address consisting of a combination of letters and numbers to make them less resistant to a "dictionary attack" – a spamming technique in which the spammer sends out thousands or millions of emails to likely real email addresses (such as common names) in the hope that some of the messages will be addressed to actual email addresses.

4. In all of the email accounts created with both of the filtered ISPs, FTC staff selected the default level of filtering to mimic the actions of a typical consumer.

5. For the study, FTC staff posted the email addresses on different pages of a website controlled by the FTC. The addresses posted on the FTC website pages were inserted in the same color as the background color of the web pages. Thus, these addresses would not be viewable to a human user viewing the page, but would be "seen" by a harvesting program.

6. In all of the message boards, blogs, chat rooms, and USENET groups, FTC staff used an undercover email account to register onto the site. FTC staff then posted the undercover email address, from each particular email account, directly into the body of a message. Where applicable, in the space for the Subject Line, FTC staff inserted neutral-sounding phrases, such as "My Views," "My Opinion," and "My Thoughts."

7. We assumed that spammers who harvested the addresses were not biased in favor or against a particular ISP when sending spam. It is possible, however, that the number of spam messages sent to the Unfiltered Addresses differed from the number of messages sent to Filtered ISP 1 or Filtered ISP 2.

8. Out of the 50 Unfiltered Addresses, 13 addresses received spam. Of these 13 addresses, 11 addresses were posted on website pages, and two of these addresses were posted on blogs. The remaining 37 of the 50 Unfiltered Addresses - which included all of the Unfiltered Addresses posted on 12 message boards, 12 chat rooms, 2 USENET groups, and 10 of the 12 blogs, received no spam at all.

9. Out of the 50 Unfiltered Addresses, 14 addresses received spam. Of these 14 addresses, 11 addresses were posted on websites, two addresses were posted on blogs, and one address was posted on a USENET Group. The remaining 36 of the 50 Unfiltered Addresses – which included all of the addresses posted on 12 message boards, 12 chat rooms, and 10 of the12 blogs, received no spam at all.

10. In FTC staff's 2002 Spam Harvest study, 86 percent of email addresses posted in USENET groups received spam and chat rooms received spam almost instantaneously after the addresses were posted. However, it is difficult to draw direct comparisons between the 2002 study and the current analysis of spam harvesting. For the current study, addresses were posted in different chat rooms. Moreover, the Internet has changed significantly in the past three years, as demonstrated by the rise in popularity of blogs and Internet Relay Chat rooms.

11. The masking technique used by FTC staff, while very effective, was not foolproof. One harvesting program apparently captured the masked address and converted the spelling out of "at" and "dot" into the "@" and "." symbols.

12. The results of this study do not suggest that Filtered ISP 2 employs a better spam filter than Filtered ISP 1. Among other things, this study does not take into consideration the increase in false positives – the inappropriate blocking of non-spam email – that may result from a tighter spam filter.

FEDERAL TRADE COMMISSION ftc.gov

1-877-FTC-HELP FOR THE CONSUMER