Understanding How Spammers Steal Your E-Mail Address: An Analysis of the First Six Months of Data from Project Honey Pot

Matthew B. Prince[†]
CEO, Unspam, LLC
Adjunct Professor of Law
John Marshall Law School
850 W. Adams, Suite 4e
Chicago, IL 60607-3096

Lee Holloway,
Eric Langheinrich,
& Benjamin M. Dahl
Unspam, LLC
850 W. Adams, Suite 4e
Chicago, IL 60607-3096

Arthur M. Keller
Information Systems and
Technology Management
Baskin School of Engineering
Univ. of California, Santa Cruz
& Advisor, Unspam, LLC

Abstract

This paper summarizes and analyses data compiled on the activities of email harvesters gathered through a 5,000+ member honey pot system that issues unique addresses based on a visitor's IP address and specific spidering time. The project, known as Project Honey Pot, has provided data about the geographical source of harvesting and mail processing, the sending patterns of different types of spammers as well as list management behavior. In addition to providing guidance for website administrators trying to forestall harvesting, the Project data also suggest that anti-harvesting regulations offer a new, potentially successful prosecutorial avenues against spam as well as inform potential amendments to current anti-spam laws that may help those efforts.

1 Introduction

It is axiomatic to say that the best way to stop spam is to keep spammers from getting your e-mail address. While e-postage, challenge-response systems, Bayesian filters, realtime block lists, and reputation services may be necessary once an address is widely distributed, all of these anti-spam measures can be made more effective if the process of obtaining e-mail addresses in the first place is made difficult and auditable. To that end, Project Honey Pot was created to understand the primary way by which spammers obtain new e-mail addresses.

Project Honey Pot (www.projecthoneypot.org) is a distributed honey pot network to track e-mail harvesters, and, subsequently, the spammers who send to harvested addresses. The Project was announced at CEAS 2004 and opened to public volunteers October 14, 2004. Since its launch, the Project's software has been installed by more than 5,000 users on websites worldwide. The Project's honey pots are running in at least 80 countries and on every inhabited continent.

As of June 20, 2005, the Project is monitoring more than 250,000 active spamtrap e-mail honey pots. Thousands of spamtrap addresses are distributed through honey pots each day and the Project is on pace to have more than 1 million active spamtraps monitored by the end of 2005.

This paper is the first thorough analysis of the data gathered by Project Honey Pot. Understanding the behavior of harvesters is critical to controlling the spam problem. Harvesters sit at the beginning of the spam cycle. Studies by the Pew Internet Project, the Center for Democracy and Technology, as well as the Federal Trade Commission have found that harvesting is the primary way spammers obtain new e-mail addresses.[1] Understanding harvesting and the resulting address distribution can provide not only a mechanism to keep e-mail addresses out of the hands of spammers, but may also help identify spam gangs and give law enforcement officials a new cause of action for prosecutions.

2 Technical Background

Project Honey Pot consists of two primary components: 1) the honey pot software installed on machines worldwide, and 2) the centralized server which collects data from and distributes spamtrap addresses to the honey pots. The Project currently supports honey pot software for platforms running the following scripting

[†] Corresponding author. Email: ceas05@matthew.unspam.com. Telephone: +01.312.543.3045 (direct).

[1] See "Spam: How it is hurting email and degrading life on the Internet," Deborah Fallows, Pew Internett & Amer. Life Project, Oct. 22, 2003 <http://www.pewinternet.org/report_display.asp?r=102>; "Why Am I Getting All This Spam?" Center for Democracy and Technology, March 2003 <http://www.cdt.org/speech/spam/030319spamreport.shtml>; "Email Address Harvesting: How Spammers Reap What You Sow," FTC Report, Nov. 2002 <http://www.ftc.gov/bcp/conline/pubs/alerts/spamalrt.htm>.

language: PHP, ASP, ASP.NET, Perl, mod_perl, ColdFusion, SAP Netweaver BSP, and Python. We also provide a wrapper for users of the MovableType blogging software to allow for easy installation.

Website administrators download and install the honey pot software. The static content of the honey pots, which primarily consists of a legal disclaimer forbidding the harvesting of the addresses displayed on the page, is randomized for each download in order to make the Project's honey pots difficult to recognize and avoid. On a few high-traffic websites, we have further customized the boilerplate legal disclaimer as well as the look of the honey pot for particular members' needs.

After the honey pot script is installed and activated, we provide instructions to the website administrator on linking from his current web pages to the honey pot page. These links are generally formatted to be invisible to human visitors to the website, but to be followed by web spiders and robots. We test these formats to ensure they are followed by the latest crop of spam harvesters.

When one of these links is followed and a honey pot is accessed by a visitor, the honey pot script installed on the webserver instantly contacts the centralized Project Honey Pot servers. The honey pot script passes to the centralized servers an array that includes the IP address of the visitor, the useragent of the visitor, and the referer string of the visitor. The servers record this visitor information as well as a timestamp and return a unique spamtrap e-mail address to the honey pot script.

The spamtrap address is handed out only once and is tied to both a moment in time and visitor information. The honey pot script combines the spamtrap address with the static content and displays a web page. The process from access to page display typically takes less than a second and creates little additional load for the web server where the honey pot script is installed.

While every spamtrap address is unique, they are designed to look like real addresses. There are two parts to every e-mail address: 1) the username, which appears before the @ sign, and 2) the domain, which appears after the @ sign. We construct usernames with a list of more than 6,000 common first names, 12,000 common last names, a 60,000 word dictionary, and random other letters and numbers. These components are combined to form typical usernames used by legitimate mailing systems. For example:

- `john.smith`
- `john_smith`
- `johnasmith`
- `jsmith`
- `orange42`
- `orangegrasslands`

For the domain portion of the spamtrap address, we use a number of domains controlled by Project Honey Pot as well as thousands of additional domains donated by our members. These donations take place by members pointing their donated domains' MX record to our servers.

By combining our donated domains with our possible usernames, we can currently create approximately 10 trillion unique e-mail addresses that will resolve to our mail servers. This allows us to distribute a unique spamtrap to every visitor to a honey pot for the foreseeable future. Moreover, it means it is difficult for spammers to determine what e-mail addresses on their lists are, in fact, spamtraps. To further disguise our spamtraps we rotate the IP addresses of our mail servers and are continuously looking for ways to further hide what addresses belong to the Project.

3 Data Analysis

Harvesters make up a significant percentage of the robot traffic currently trolling the Internet. Approximately 6.5 percent of the traffic visiting our honey pots subsequently turns out to be spam harvesters. While some human traffic inevitably finds our honey pots, the vast majority of visitors to these pages are automated spiders. We estimate, therefore, that harvesters make up at least 5 percent of all automated traffic online.

The average time from a spamtrap address being harvested to when it receives its first message is currently 11 days, 7 hours, 43 minutes, and 10 seconds. The fastest turnaround is less than 1 second, and the slowest is 223 days, 19 hours, 37 minutes, and 8 seconds. The slowest time is just under the total online age of the Project. As a result, we believe that the average turnaround time will continue to rise slightly as the Project ages.[2]

We've been surprised, so far, by how slow the turnaround for some spammers has been. This lends support to the hypothesis that there is a class of individuals involved in the spam trade who methodically gather addresses. These individuals could be spammers who also send to those addresses, or they could sit at the top of the spam food chain, selling the lists they obtain to the spammers who then send messages to those lists. There is additional evidence from recent legal cases that these "listmen" do exist as part of the spam economy. Identification of these listmen, with an understanding and control of their behavior following closely thereafter, we believe offers a critical choke point in the spam cycle both legally and technologically.

We have also been surprised by how clearly many harvesters identify themselves. A substantial percentage of harvesters can be identified by the "useragent" they

[2] For the latest stats, see the Project Honey Pot statistics available online at: http://www.projecthoneypot.org/statistics.php.

broadcast when visiting a website. While some harvester disguise their identity pretending to be a typical website visitor (*e.g.,* Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)), more than 50 percent of the time harvesters broadcast a useragent which is unique and identifiable. For example, the following useragents appear to be exclusively broadcast by harvesters:

- `Missigua Locator 1.9`
- `MSIE5.5`
- `Port Huron Labs`
- `Program Shareware 1.0`
- `Wells Search II`
- `Franklin Box Company`

Additionally, approximately 58 percent of the so-called "phishing" messages we have seen begin with a harvesting event by a spider broadcasting the useragent: Java/1.#.#_## (where the #s are replaced with numerals). There is a surprising lack of harvesters broadcasting the useragents of Google, Yahoo, MSN, or other "legitimate" robots. We had anticipated more masquerading of these legitimate robots by harvesters. We still believe that as more website administrators begin blocking harvesters based on their useragents harvesters will increasingly use misleading useragents.

3.1 The Two Classes of Spammers: Hucksters and Fraudsters

From the analysis of the Project Honey Pot data, we have characterized two distinct classes of harvesters. These classes break down by their turnaround time from harvest to first message, their repeated use of the spamtrap addresses, and the type of messages they send.

The first class — the hucksters — are characterized by a slow turnaround from harvest to first message (typically at least 1 month), a large number of messages being sent to each harvested spamtrap address, and typical product-based spam (*i.e.,* spam selling an actual product to be shipped or downloaded, even if the product itself is fraudulent).

The second class — the fraudsters — are characterized by an almost immediate turnaround from harvest to first message (typically less than 12 hours), only a small number of messages sent to each harvested spamtrap address, and fraud-based spam (*e.g.,* phishing, "advanced fee" fraud, etc.).

These two different classes represent two unique problems when fighting spam. Typical huckster spammers are responsible for most of the volume of spam on the Internet. The Project's data indicates that they tend to be slower to send the first message, but once they have an address they more efficiently exploit it. The good news appears to be that these huckster spammers are more reliant on a relatively small universe of harvesters. In fact, only 25 harvesters are responsible for more than 50 percent of the volume of spam that has been sent to the Project.

Fraudster spammers often harvest an address and send only a single message to it. Seventy percent of the harvesting incidents targeting Project spamtraps have resulted in only a single e-mail message. In a vast majority of these cases, the single message sent is some form of phishing scheme, advanced fee fraud, or other banking scam.

The Project sits in a unique position to capture phishing and other fraudulent messages. Since it is impossible for one of our spamtraps to, for example, sign up for a PayPal account, any messages mentioning a financial institution can be flagged as likely to be fraudulent. This allows us to quickly and efficiently identify new phishing scams as well as track the scams back to the computer used to harvest the addresses. Nearly 30 percent of the messages received by the Project so far appear to be related to some sort of phish scheme, advanced fee fraud, or other banking scam.

3.2 Establishing Spammers' Identities through Harvesting Activity

While there appears to be harvesting software on the market that runs through proxy connections, currently spammers are, by and large, not taking advantage of it. In fact, spammers, regardless of the category they fall into, are clearly not going through the same effort to obscure their identity when harvesting as they are when sending.

Harvesters appear to remain surprisingly stable. A majority of the harvester IPs that have visited honey pots at least five times have made those visits over the course of at least three months. Additionally, only 3.2% of the IP addresses used for harvesting appear in an open proxy or open relay database.[3] Compare that with the 14.6% of IP addresses used for sending to spamtrap addresses which appear in an open proxy or open relay database.[4]

The Project Honey Pot data further suggests that harvesters are generally more likely to be associated with a traceable individual responsible for the spam than many of the machines used for the actual sending. While 39.5% of spam servers are hosted on some sort of an account with a dynamic IP address, only 22.8% of harvesters are hosted on an account with a dynamic IP address.[5] This indicates that harvesting is generally occurring from more stable, established hosting space.

---

[3] The SORBS Open-Relay/Open-Proxy services were queried for this information. Queried March 21, 2005.

[4] *Ibid.*

[5] The SORBS-DUHL was queried for this information. The service lists the IP addresses that are allocated for dial-up, DSL, cable modem, and other dynamic IP space. Queried March 21, 2005.

More generally, harvesters have, to this point, slipped under the radar screen in the spam cycle. Scanning several of the major realtime block list services, 46.2% of the IP addresses of spam servers sending to Project Honey Pot spamtrap addresses were listed.[6] On the other hand, only 25.2% of harvesters appear on any of the major blocklists.[7]

Similar percentages hold true when searching mail abuse newsgroups for the spam server and harvester IP addresses. Most of the harvesters that appear on the blocklists or in mail abuse newsgroups are engaged in both harvesting and spamming. What is of note is that nearly three out of every four IP addresses engaged in harvesting appear to be untracked by the conventional anti-spam resources.

3.3 Geographic Location of Harvesting and Spamming

The geographic location of harvesters versus spam senders further evidences the difference between these two activities and demonstrates how harvester activity can be used to track the true identity of the individuals actually responsible for spam. We use MaxMind's open source GeoIP IP-to-country data to establish the geographic location of the IPs used for harvesting and spam sending.[8] While the United States tops both lists, there are significant differences below that.

Rank	Country	Percentage
#1	United States	32.1%
#2	Romania	17.1%
#3	China	12.4%
#4	United Kingdom	8.6%
#5	Japan	7.2%
#6	France	6.9%
#7	Spain	4.3%
#8	Egypt	4.0%
#9	Nigeria	3.7%
#10	Canada	3.7%

Table 1: *Top-10 Countries for Harvesting*[9]

Rank	Country	Percentage
#1	United States	38.4%
#2	China	14.9%
#3	Korea	13.4%
#4	France	7.6%
#5	Brazil	6.3%
#6	Japan	5.3%
#7	Taiwan	4.0%
#8	Spain	3.6%
#9	United Kingdom	3.6%
#10	Canada	2.7%

Table 2*: Top-10 Countries for Spam Sending*[10]

For example, Romania is the second most common country for harvesting, but does not appear in the top-10 for spam sending. Romanian harvesting almost exclusively falls into the "fraudster" category, with virtually all harvesting incidents leading to a phishing attack approximately 24 hours later.

When we first began the Project, most of these phishing messages were actually being sent out of Romania. Today, however, the messages are almost all being sent through computers in other countries. In fact, a strong association appears to exist between the Romanian harvesting and a large number of the messages being sent from France, which ranks third on the spam sending list. Because each spamtrap address is unique and allows us to track messages sent to it back to the moment and time and IP address that harvested the spamtrap, we are able to establish a definitive connection between the Romanian-based IP addresses harvesting and the French-based IP addresses sending messages.

Nigeria is another example where harvesting can help show the real identity of the individuals behind the spam sending. Famous for so-called "advanced fee" scams, Because of this, Nigerian-based IP addresses are largely blacklisted and spam servers are often set to automatically reject messages originating from the African continent.[11]

Not surprisingly, it appears Nigerian fraudsters have responded. The country ranks 46th for spam sending, near the bottom of the list of countries from which the Project has received spam messages. On the other hand, Nigeria is 9th for harvesting. Looking deeper into the data, virtually all of the messages that result from harvesting by Nigerian IP addresses appear to be advanced fee frauds. In other words, the Project shows that the blacklisting of Nigeria has not eliminated fraudsters operating from that country, but instead has displaced their outgoing mail to other countries.

3.4 The Makings of an Effective Spamtrap

In terms of what addresses spammers are looking for, there appears to be little differentiation by harvesters

[6] The following widely-used realtime blocklists were queried: SORBS, Spamhaus, SpamCop, and SPEWS. Queried March 21, 2005.
[7] *Ibid.*
[8] MaxMind GeoIP Database (http://sourceforge.net/projects/geoip/).
[9] For the latest stats, see the Project Honey Pot statistics available online at: http://www.projecthoneypot.org/statistics.php.
[10] *Ibid.*
[11] For example, one check in the popular anti-spam program SpamAssassin is whether a message originates from an IP address within Nigeria. Additional services such as nigeria.blackholes.us exists to check the Nigerian IP space. Other regularly employed blocklists exist for China, Korea, Thailand, Malaysia, Argentina, and Brazil — all countries that appear high on the list of spam senders. *See*, for example, http://www.epaxsys.net/dnsbl/.

between various constructions of e-mail addresses. We originally hypothesized that harvesters would prefer those addresses constructed with "common" top level domains (TLDs) — .com, .net, and .org — over country-specific TLDs (*e.g.,* .au, .ca, .ro, etc.), or other less common TLDs (*e.g.,* .biz, .info, etc.). We also hypothesized that harvesters would prefer those addresses constructed with 2-level domains (*e.g.,* example.com) versus domains with 3 or more levels (*e.g.,* thirdlevel.example.com).

Neither hypothesis has thus far proved true. Statistically, harvesters appear initially to be as likely to send to an address with a country-specific/uncommon TLD or a 3- or 4-level domain as they are to send to a common TLD and a 2-level domain. Moreover, the same statistical pattern continues over time: country-specific/uncommon TLDs and 3- or 4-level domain based addresses appear to be statistically as likely to remain on a spammers list and be traded with other spammers as those more commonly formed addresses. Unfortunately, this means that simply registering a domain on an obscure South Pacific island nation and using it for an e-mail address is not enough to keep your inbox spam-free.[12]

While more difficult to prove statistically, our best estimate is that an obscure username also does not dissuade spammers from including a harvested address on their mailing lists. The top-5 spamtraps by volume were constructed with approximately the following usernames: calchera9415, wellreadhoene, lionel.uan, wriestscammabell, and paulshake.[13] While not unheard of, these usernames are generally on the more obscure end of those distributed by the Project. We have been unable to tell any correlation between the obscurity of the username in an address and its propensity to receive spam.

3.5 Protecting Legitimate E-Mail Addresses

While the form of an e-mail address displayed on a page is not enough to discourage harvesters, we have found there are some steps that sites can take in displaying an address to keep it relatively safe from harvesting. Interestingly, the techniques that are successful at discouraging harvesting again break down into the two classes of harvesters: fraudsters and hucksters.

Generally, fraudsters appear to be using less sophisticated harvesting software. The software appears to take a page at its face value, do no HTML-rendering or character processing, and pick up anything meeting the basic format of an e-mail address (*i.e.,* a string of characters followed by an @ followed by another string of characters and at least one period). Because these rudimentary harvesting programs do little processing of the page, they can typically be fooled by basic address "munging." For example, we found that 52 percent of address harvesters would not recognize an e-mail address on a page if the @ sign were simply replaced with the ASCII-equivalent HTML character code (@). Simply replacing @ signs in e-mail addresses with @ is surprisingly effective at keeping less sophisticated harvesters away.[14]

More sophisticated harvesting programs that automatically decode ASCII-equivalent HTML characters are available and being used by many of the huckster-class of spammers. For these more sophisticated harvesters, basic address munging offers no protection. In some cases, however, the sophistication of these harvesters presents a new Achilles heel. Several of these harvesters have an option to "avoid spamtraps." It is possible to include specific elements in pages that will cause any e-mail addresses present on them to be skipped over by these harvesters running in "avoid spamtraps" mode.

Specifically, we have found evidence that a box that appeared at the bottom of our honey pots is sufficient, when present on a non-honey pot page, to cause some sophisticated harvesters to pass it over. The placement of the box on the page, and even rendering it invisible with CSS or Javascript, continues to offer the protective benefit. We initially left the box on the honey pot pages in order to see if harvester authors would design their software to avoid pages containing it. It appears at least some have taken the bait and, as a result, we now no longer display the box on honey pots.

Table 3: *Project Honey Pot Box, Sufficient to Keep Away Some "Sophisticated" Harvesters*[15]

Even without the inclusion of the box, some "sophisticated" harvesters can be fooled by surprisingly easy tricks. Putting the words or phrases "spamtrap," "spam harvester," or "honey pot" anywhere on a page was enough to cause at least one of the more sophisticated harvesters running in "avoid spamtraps"

[12] We were not able to test whether certain other TLDs, specifically .gov and .mil, were avoided by harvesters. Some harvesting software we have analyzed advertises an option to avoid these TLDs as "dangerous" and defaults to not harvesting from them. Because we do not currently have any domains in the Project with these TLDs, we have not been able to test whether spammers are using these options or otherwise avoiding these addresses.

[13] These usernames have been slightly altered in order to protect the integrity of the still-valid spamtrap addresses. Their "spirit," however, remains intact.

[14] For more information on "munging" addresses, see "How to Avoid Being Harvested by Spambots," Project Honey Pot <http://www.projecthoneypot.org/how_to_avoid_spambots.php>.

[15] For instructions on including the Box on a website, see "How to Avoid Being Harvested by Spambots," Project Honey Pot <http://www.projecthoneypot.org/how_to_avoid_spambots_5.php>.

to not harvest any addresses on a legitimate page. This was true even if the words or phrases appeared in the <HEAD> element of the page, or in other non-visible areas of the page's content. These tricks do not offer complete protection, but they do show how as spammers adapt to avoid honey pots, we can exploit their adaptations in order to protect legitimate addresses. While this is yet another arms race, this time the anti-spam forces are in the position of strength.

Two widely used e-mail address protections appear to still offer substantial protection. First, the inclusion of an email address in an image will protect it from harvesting from most harvesters. Second, using Javascript in order to obscure an address in the code of the page and then render it to human users still assures virtually complete protection. While it's possible that harvesters will begin compiling and executing Javascript, this would likely severely slow down their processing speed and open them up to a number of potential attacks (*e.g.,* infinite loops written into honey pot pages humans are unlikely to access).[16]

4 Legal Implications

Accurately tracking e-mail harvesters affords several new routes to attach legal liability to anyone involved in the spam industry. To begin, the harvester IP addresses provide a new data point which potentially reveals the identity of the individual behind the spam. Effectively, gathering these IP addresses is like finding more fingerprints at a crime scene.

Moreover, because harvesters generally have not used proxy networks, these "fingerprints" appear likely to be more valuable in establishing the actual identity of spammers than many of the IPs of the servers being used to send spam. While this advantage will surely diminish over time as harvesters move to proxy networks, the data from the Project appears to make it clear that this is not yet the case. While these fingerprints are still of use, prosecutors can use evidence from these honey pot addresses to more easily identify those behind the harvesting.

In addition to the fingerprints provided by the harvester IP addresses, Project Honey Pot can also help illuminate spam networks. Because each spamtrap e-mail address is unique, spammers leave a trail when they send to them. If the identity of a spammer sending to a particular spamtrap address at a particular time can be established, the Project's data can then associate that spammer with any other messages sent to the same address. Any spamtraps on spammers mailing lists then effectively become homing beacons which track the behavior of anyone who sends to such addresses. Law enforcement agencies can use this data to build cases identifying what spam messages a particular individual is responsible for as well as larger networks of conspiracies that can be prosecuted. This may bring those harvesters that generally avoid sending spam to justice in a way that was previously impossible.

Evidence that an individual has sent to a spamtrap e-mail address is also likely to help prosecutors bring cases against an accused spammer. First, since Project Honey Pot addresses cannot "opt-in" to any mailing lists, the question of whether a message is in fact "unsolicited" becomes moot. Since this question has consumed a significant amount of time in several recent spam cases,[17] the Project's data may be useful in avoiding much of the expense and uncertainty that has been inherent to spam prosecutions. Driving down enforcement costs is one way to make anti-spam laws much more successful than they have been to date.

Second, harvesting itself may lead to a cause of action in several jurisdictions. Under the U.S. CAN-SPAM Act, for instance, sending to a harvested address at least augments the applicable penalties, and may be a cause of action in and of itself.[18] To date, however, no enforcement actions have been brought under the anti-harvesting provisions of CAN-SPAM.

In addition to U.S. law, the law of other countries makes harvesting illegal. The Canadian Privacy Commission, for instance, recently ruled that the harvesting of e-mail addresses to send spam violates the Canadian Privacy Act.[19] Australia's anti-spam law also contains specific provisions which attach liability to harvesting or using harvester software.[20] Specifically, the law requires that "address-harvesting software must not be supplied, acquired or used."[21] Moreover, "an electronic address list produced using address-harvesting software must not be supplied, acquired or used."[22]

The critical element missing in order for Australia's anti-harvesting provisions to be enforced is the data proving that addresses were acquired through harvesting. Project Honey Pot and other anti-harvesting

[16] For more information on using images and Javascript to avoid being harvested, see "How to Avoid Being Harvested by Spambots," Project Honey Pot <http://www.projecthoneypot.org/how_to_avoid_spambots_3.php>.

[17] Proving messages are unsolicited has been a hurdle for virtually every anti-spam prosecution to date. Even if the prosecutions are ultimately successful, substantial time and effort is spent to establish that the messages were unsolicited. Data from honey pots may help lessen this expense, and increase the likelihood of success at trial, going forward.

[18] *See* U.S. CAN-SPAM Act of 2003, Sec. 4(b)(2)(A)(i), Sec. 5(b)(1)(A)(i) and Sec. 7(g)(3)(A) – (C).

[19] *See* Canadian Personal Information Protection and Electronics Document Act (PIPEDA). *See also* "Privacy chief takes aim at spammers with e-mail ruling," Ottawa Business J., Feb. 21, 2005 <http://www.ottawabusinessjournal.com/282670852141876.php>.

[20] *See* Australian SPAM ACT 2003, No. 129, 2003 – Sect. 19.

[21] *See* Australian SPAM ACT 2003, No. 129, 2003 – Sect. 19, simplified outline.

[22] *Ibid.*

initiatives may be essential for providing the data needed to effectively enforce these provisions.

Finally, harvesting may provide a clear, bright line that helps define what messages constitute spam and what messages do not that has, heretofore, been missing from anti-spam legislation. Defining whether a message is legitimate based on how the e-mail address was first obtained seems likely to provide a more-workable model than trying to determine after-the-fact consent and whether a message was "solicited." The Project already appears to have "split the room," with few legitimate marketers sending to harvested addresses. Future anti-spam legislation could follow Australia's lead and attach liability to the practice of harvesting and sending to harvested addresses. This would effectively penalize spammers with less risk of liability being attached to legitimate, responsible marketers.

5 Technical Implications

In addition to providing new tools and data for law enforcement, Project Honey Pot opens the possibility for a number of other technical measures to fend off harvesters and, subsequently, the spammers who rely on harvested addresses. As part of the Project, we are creating the http:BL data feed. This feed, which will be provided at no cost to any active member of the Project, allows website administrators to control the access of known harvesters. We anticipate that, over time, software authors will build the data feed into their applications in creative ways.

For example, if a known harvester IP attempts to access a website, the web server could route it instead to a gateway page containing a CAPTCHA or Javascript-based redirect. Only if the CAPTCHA is passed, or the Javascript is correctly interpreted, will access be granted. Alternatively, the web server could automatically strip email addresses out of the site's content when it is accessed by a known harvester. These measures can provide a level of technical protection for individuals that want to continue to display their e-mail address online.

Additionally, Project Honey Pot provides a way for ISPs to automatically monitor their IP space for e-mail harvesters and spammers. If an ISP provides their AS-Macro to the Project we will watch the IP space controlled by the ISP for harvesting or spamming behavior. Our systems automatically generate an e-mail, web page, or XML feed which can be checked to alert the ISP's abuse department if harvesting or spamming is occurring within their network. This monitoring service is provided at no cost to ISPs that are active participants in the Project.

Finally, the Project regularly publishes its corpus of spam for anti-spam researchers and filter authors. Because this corpus is devoid of any personal e-mails and is composed virtually entirely of spam, its publication and use does not create the same problems in terms of privacy posed by other spam corpuses that have been published. Our hope is that the data we are gathering will help the anti-spam community working on the problem further downstream to continue to develop increasingly effective tools.

6 Conclusions and Future Directions

Project Honey Pot's membership continues to grow and the volume of data it collects on a daily basis is rapidly expanding. The Project is on pace to have more than one million spamtrap e-mail addresses in circulation by the end of 2005. With each newly installed honey pot the entire membership in the Project benefits.

Going forward, we hope to use the Project's vast network of honey pots to track other spam behavior not necessarily directly tied to e-mail. For example, the Project has begun to study the behavior of referer/log spammers, who use robots to fill web server logs with bogus information. We are also designing the next generation of our honey pots to track comment spammers and other misbehaved robots. We hope to provide data to help understand these other online pest and to be able to demonstrate whether the same computers and individuals behind e-mail address harvesting and spamming are also behind these additional spamming behaviors.

Acknowledgements

Project Honey Pot owes special thanks to the volunteer coders who have ported the honey pot code to additional platforms: Eddy De Clercq (SAP Netweaver BSP), Jeffrey Greenhouse (ColdFusion), Marek Isalski (Python), Tom McIntyre (mod_perl), and Jeff Turner (ASP). Additional thanks to the thousands of Project Honey Pot members who have taken the time to install a honey pot or donate an MX to help get the results described above. We couldn't have done it without you.