Proceedings of the 2005 IEEE
Workshop on Information Assurance and Security
United States Military Academy, West Point, NY

An e-mail honeypot addressing spammers' behavior in collecting and applying addresses

Guido Schryen

Spam has become one of the most annoying and costly phenomenon in the Internet. Valid e-mail addresses belong to the most valuable resources of spammers, but little is known about spammers' behavior when collecting and harvesting addresses and spammers' capabilities and interest in carefully directed, consumer-oriented marketing have not been explored yet. Gaining insight into spammers' ways to obtain and (mis)use e-mail addresses is useful in many ways, e.g. for the assessment of the effectiveness of address obscuring techniques and the usability and necessity of hiding e-mail addresses on the Internet. This paper presents a spam honeypot project in progress addressing these issues by systematically placing e-mail addresses in the Internet and analyzing received e-mails. The honeypot's conceptual framework, its implementation, and first empirical results are presented. Finally, an outlook on further work and activities is provided.

Spam, ham, e-mail, honeypot, address obscuring technique, address taxonomy

I. Introduction

Spam, that is unsolicited e-mail to a large number of recipients, is generally recognized as an increasingly disturbing and costly issue for electronic business and Internet traffic. Companies, non-profit organizations, and individuals receive this kind of e-mail to such an extent that the issue has certainly gone beyond what is merely "annoying". More than two-thirds of the world-wide e-mails is categorized as spam: Symantec reports that the percentage of spam e-mails reached 67% in December 2004 – 106 billion e-mails were scanned – and continues to grow at a fairly steady rate. However, by January 2005 – a month later – the total volume of e-mail increased by nearly 19% over the preceding month [1]. MessageLabs announced that the average global ratio of spam was even more than 81% in December 2004, although the sample of e-mails inspected was much smaller, comprising some million per day [2]. The content of spam e-mails covers a broad range of topics: spammers' e-mails mainly offer or advertise general goods and services (23% of all e-mails categorized as spam), contain references or offers related to money, the stock market or other financial "opportunities" (15%), contain or refer to products or services intended for persons above the age of 18 (14%), or offer or advertise health-related products and services (12%). The increased payload of networks and e-mail servers and also the consumption of employees' attention and time is not the only harm spam e-mails cause. Fraudulent messages, e.g. e-mails that appear to be from a well-known company but are not, also known as "brand spoofing" or "phishing" e-mails, are often used to trick users into revealing personal information such as e-mail addresses, financial information and passwords (16%). Furthermore, viruses, worms, and Trojan horses (opening backdoors for botnets using the infected computer as spam client) are distributed over the Internet. The economic damage caused in total by spam e-mails is estimated at several billion US$ [3].

This central economic aspect motivated anti-spam activities embracing many facets: national laws and international regulations – http://notebook.ifas.ufl.edu/spam/Legislation.htm) provides a good overview – , organizational provisions including abuse systems (e.g. http://spam.abuse.net/) and lists of suspicious domains and IP numbers, and technical solutions mainly applying blocking, filtering, or authenticating mechanisms [4]. However, statistics and e-mail users' daily experience show that the spam problem is far from being solved. Nevertheless, the application of technical anti-spam is necessary and has prevented our Internet e-mail system from collapsing.

Alongside these mainstream activities efforts to analyze spammers behaviour or to even attack them have come up implementing honeypots and honeynets [5,6]. The honeypot presented in this paper contributes to this field by setting up a technical environment that allows to analyze where spammers get their e-mail addresses from and how they exploit them (or if they simply use any harvested e-mail address).

II. Motivation and goals

Valid e-mail addresses belong to the most valuable resources of spammers, and identifying address sources and spammers' exploiting procedures is crucial to preventing from getting addresses and from misusing them. It is widely known that beside generating addresses with brute force mechanisms, spammers get valid e-mail addresses from harvesting the Internet or, illegally, from organizations. Some address obscuring techniques (AOT) restricting the availability and usability of e-mail addresses have been proposed : already in 1997 Hall described e-mail channels [7], and single-purpose addresses encapsulating policy in the address were

0-7803-9290-6/05/$20.00 ©2005 IEEE.

presented by Ioannidis in 2003 [8]. Many users also use temporary addresses and abolish them when they feel that the spam portion has become too high.

Gaining insight into spammers' ways to obtain and (mis)use e-mail addresses is useful in many ways:

- Assessment of the effectiveness of AOT and input for their improvement
- Identification of spammers leading to their prosecution
- Assessment of the usability and necessity of hiding e-mail addresses on the Internet
- Discovery of specific marketing and addressing activities

The last item aims at the quality of e-mail addresses. Spammers are known to collect as many valid e-mail addresses as possible but little is known about spammers' capabilities and interest in carefully directed, consumer-oriented marketing. A taxonomy of quality for e-mail addresses is shown in figure 1.

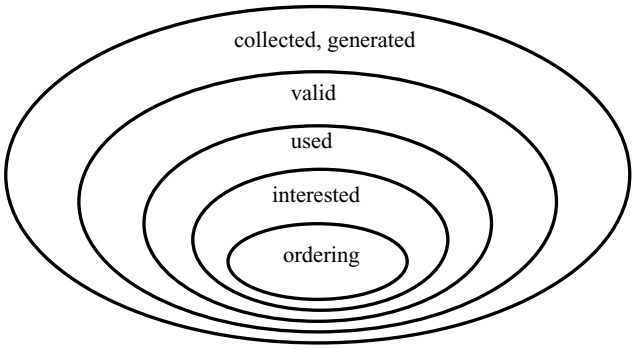

Figure 1: Taxonomy of e-mail addresses

The inner ellipses are more valuable for spammers than the outer ones due to their losses caused by non-selective advertising . Only a portion of collected or generated e-mail addresses are valid ones, i.e. e-mails addressed to non-valid ones are refused by the addressee's host as these mailboxes do not exist. The valid ones can be divided in used addresses and those that are no longer accessed and thus useless for spammers. A way to distinguish between the two is provided by an opt-out option included in some spam e-mails, which when used incautiously by the spam recipient indicates that the address is in use. Spammers can even go a step further when adopting physical marketing strategies using knowledge about consumer-specific interests and behaviour: for example, an Internet user participating actively in a German discussion group that focuses on medical products is presumably interested in medical product offers in the German language. The innermost ellipse contains the e-mail addresses belonging to users who buy products and and, thus, from whom the spammer profits.

The goal of the honeypot presented in this paper is (1) to penetrate spammers' behavior in harvesting e-mail addresses from Internet services such as newsgroups and the web and (2) in discovering the extent to which spammers have already shifted from simply employing used e-mail addresses towards acquiring addresses of users likely to be interested (specific marketing).

III. CONCEPTUAL FRAMEWORK

In order to cover a broad range of locations which are attractive for spammers to harvest e-mail addresses it is necessary to inspect many Internet services. Integrated in this honeypot are newsletters/mailing lists, web pages, web chats, ICQ chat, and the Usenet in which e-mail addresses are placed. In order to detect language- and/or region-specific particularities each of these mediums is split up into German (language) oriented ones and US-based ones which, as a second dimension, renders the study readily extensible to other languages and regions. To allow inspecting spammers' behavior regarding specific marketing activities a third dimension of the survey focuses on the topic of the Internet service. For example, web pages and newsletters/mailing lists are divided into those ruled by an individual, a discussion board, greeting card service etc; for a complete list of topics see annex A in which the topics are grouped by types administration, content, connection, context and commerce. It should be noted that topics are service-specific. Figure 2 shows the classification of Internet locations as used in the empirical study. Each type of location is represented by a cube, each cube contains three locations (a location is a specific web side or a specific newsletter), each location gets four addresses (de-, com-, net-, and org-address), i.e. for each cube 12 e-mail addresses have to be reserved. This procedure allows to detect if the top level domain of an e-mail address is relevant. So far, German and US newsletters/mailing lists and web pages have been addressed for almost all topics listed in annex A, i.e. the number of e-mail addresses placed for getting harvested is almost 2*2*36*12 which is 1728. Of course, no e-mail address must be seeded more than once.

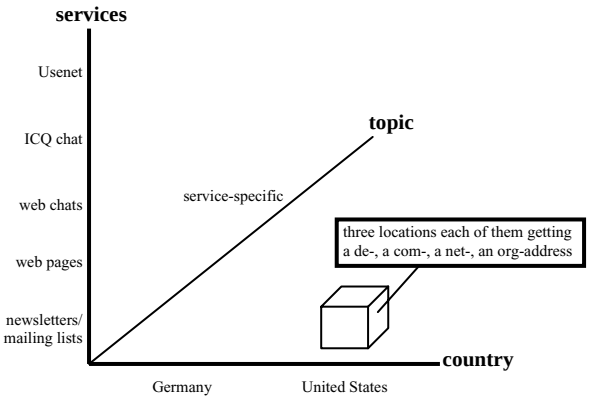

Figure 2: Classification of Internet locations

IV. Implementation

A mail server has been set up, namely charlie.winfor.rwth-aachen.de, and three domains have been reserved, wforasp.com, wforasp.net, and wforasp.org for covering e-mail addresses of four top level domains. All e-mails addressed to these domains are directed to this mail server. As thousands of e-mail addresses had to be created they were generated automatically by using a random generator for the user part of the addresses. In order to prevent e-mail addresses from being guessed or generated with brute force attacks it is necessary to define them randomly as well as to give them an appropriate number of characters. An example of an e-mail addresss created this way is wasp10208@wforasp.com.

The Internet locations serving as lures were chosen manually just as the placement of the e-mail addresses had to be done manually. As soon as an e-mail address is spread, its location and activation date is stored.

All incoming e-mails are classified into regular e-mails (ham e-mails), e.g. regular newsletters or such containing comments from users of discussion forums, and spam e-mails. This procedure is currently mainly executed by humans but supported by a mail parser (written in PHP) which uses an increasing white list containing pairs of (recipient-address, IP) entries: each time a host was manually assessed qualified to send an e-mail to recipient-address, its IP number is linked to this e-mail address and stored in the white list. A second task of the mail parser is to decompose each incoming e-mail: all entries of the header are analyzed as well as the content and the (MIME) structure of the body. A detailed description of the (relational) data model the procedure is based on is beyond the scope of this paper. Next, the e-mails' elements are stored in the (MySql) database broken down into spam and ham e-mails. The database is intended to be used by data mining tools and (simpler) statistical analyzers. Figure 3 provides a survey of the implementation infrastructure.

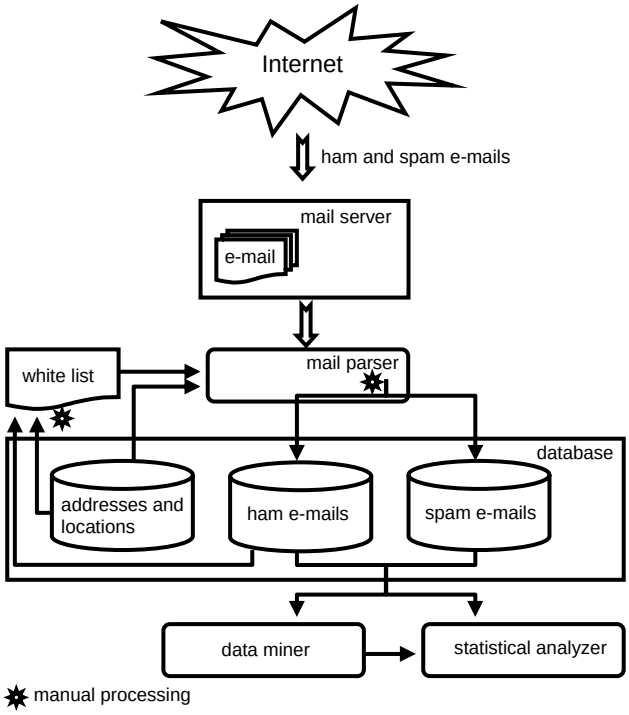

Figure 3: Infrastructure of the e-mail honeypot environment

V. First empirical results

Due to the early stage of the project the results presented here are preliminary and include only numbers of spam and ham e-mails received in different categories. In total, 7968 ham e-mails and 2482 spam e-mails have been recorded by our mail server. Table 1 shows the results for (e-mail addresses placed in) newletters/mailing lists and on web pages. The first column represents the service (nl/ml=newletters/mailing lists; wp=web pages), the column "tld location" refers to the dimension "country", "topics" classifies according to the topics given in annex A. "addressees" refers to the top level domain of the e-mail address placed in the locations. The entry "--" indicates that no spam e-mails arrived in the respective category. For example, the highlighted entry refers to the number of spam e-mails that were directed to addresses used for subscribing US administration newsletters/mailings lists. 19 spam e-mails were received in total, 17 of them sent to com-addresses, one to an org- and one to a net-address.

service	tld location	topics	all	[2;7]	[8;11]	[12;18]	[19;20]	[21;36]
		addressees	all de net org com	all de com org net	all de com org net	all de com org net	all de com org net	all de com org net
nl/ml	de		--	--	--	--	--	--
	com		19 -- 17 1 1	**19 -- 17** **1 1**	--	--	--	--
wp	de		515 76 108 211 120	240 20 40 107 73	5 2 2 1 0	7 2 3 1 1	81 45 18 12 6	182 7 45 90 40
	com		1803 311 656 537 299	860 123 299 298 140	409 106 124 111 68	--	--	534 82 233 128 91

Table 1: Number of spam e-mails received on addresses placed on web pages or used for newsletter/mailing list subscription

Although the project is in a very early stage, some facts seem worth being mentioned:

- No spam has been sent to addresses that were used for subscribing German newsletters/mailing lists.
- Only a few spam e-mails have been received due to US newsletter/mailing list subscription. The few ones are all due to administration topics.
- Not surprisingly, many more spam e-mails arise from placements on web pages. Interestingly, German web pages caused only a third of the number of spam e-mails that are due to US web pages. Independent of the country, net-addresses seem to be more interesting for spammers than de- and org-addresses, on US web sites com-addresses have been even more used by spammers.

VI. SUMMARY AND OUTLOOK

Spammers are known to collect as many valid e-mail addresses as possible but little is known about spammers' capabilities and interest in carefully directed, consumer-oriented marketing. Gaining insight into spammers' ways to obtain and (mis)use e-mail addresses is useful for the assessment of the effectiveness of AOT and input for their improvement, for identification of spammers leading to their prosecution, for assessment of the usability and necessity of hiding e-mail addresses on the Internet, and the discovery of specific marketing and addressing activities. The paper presents a honeypot to penetrate spammers' behavior in harvesting e-mail addresses from Internet services such as newsgroups and the web and in discovering the extent to which spammers have already shifted from simply employing used e-mail addresses towards acquiring addresses of users likely to be interested (specific marketing). The honeypot's conceptual framework classifies Internet locations as used in the empirical study using three dimensions: Internet services (e.g. the Usenet, web pages, newsletters), service-specific topics such as education, infotainment, and auctions, and countries. Each location gets four addresses (de-, com-, net-, and org-address) allowing to detect if the top level domain of an e-mail address is relevant for spammers. When e-mails arrive at the honeypot's mail server they are classified into spam and ham e-mails (regular e-mails), decomposed by a parser, and stored in a database that is intended to be used by data mining tools and (simpler) statistical analyzers. Preliminary results of the honeypot study are presented showing that no spam has been sent to addresses that were used for subscribing German newsletters/mailing lists, only a few spam e-mails have been received due to US newsletter/mailing list subscription, many more spam e-mails arise from placements on web pages, and net- as well as com-addresses seem to be especially interesting for spammers.

The project is at an early stage. More services and countries have to be integrated, and to allow more reliable results and to apply time-series analysis more data have to be collected. Another branch to be extended is the functional part, i.e. the application of data mining procedures and statistic procedures aiming at detecting differences between spam and ham e-mails. They can be used for the improvement of spam filters.

ANNEX A

class	No	Topic
	1	personal web page (not included yet)
administration	2	departments, authorities, offices
	3	federations, unions
	4	social welfare organizations
	5	churches, sects
	6	associations, clubs
	7	institutions of education
content	8	information
	9	entertainment
	10	education
	11	infotainment
connection	12	discussion board
	13	peer-to-peer (not the service itself)
	14	chats (not the service itself)
	15	greeting cards
	16	Internet providers
	17	community providers
	18	(content) mobile providers
context	19	search engines, web catalogues
	20	meta search engines
commerce	21	auctions
	22	payment
	23	logistics and transport
	24	web portals
	25	shops, malls
	26	finance, insurances
	27	tourism
	28	jobs
	29	motor vehicles
	30	property
	31	social contacts
	32	health
	33	gambling, lottery
	34	adult material
	35	(computer) hardware
	36	software

Topics specific to the services „newsletter/mailing lists“ and “web pages“

VII. REFERENCES

[1] Symantec: Spam statistics. http://www.symantec.com/region/de/PressCenter/spam.html [Accessed 04/01/05].

[2] MessageLabs: Email Threats. http://www.messagelabs.com/emailthreats/default.asp [Accessed 04/01/05].

[3] OECD: Background Paper For The OECD Workshop On Spam, 2003.

[4] Schryen, G: Effektivität von Loesungsansaetzen zur Bekaempfung von Spam. Wirtschaftsinformatik 46 (2004) 4, pp. 281-288. (English version is not published but available from the author)

[5] The Honeynet Project. http://honeynet.org.

[6] Project Honey Pot. http://www.projecthoneypot.org.

[7] Hall, R.: Channels: Avoiding Unwanted Electronic Mail. Proceedings DIMACS Symposium on Network Threats DIMACS, 1996.

[8] Ionnadis, J: Fighting Spam by Encapsulating Policy in Email Addresses. Network and Distributed System Security Symposium (NDSS’03), 2003.

VIII. ACKNOWLEDGEMENTS

The set up of the honeypot was strongly supported by Reimar Hoven, manual work regarding the classification of incoming e-mails had to be done, Stephan Hoppe sacrified much time in performing this task. Many thanks also go to Jan Herstell and to Katrin Ungeheuer for proofreading.