Spammer detection using honeypots and digital forensics

Ickin Vural, HS Venter

Abstract— At present it is very difficult to trace the identity of spammers who use identity concealment techniques. It is difficult to determine the identity of the spammer by just analysing the electronic trail.

This paper proposes the design and implementation of a spammer detection system that uses honeypot techniques to detect abnormal behaviour on a network so as to identify potential spammers.

I. INTRODUCTION

Unsolicited bulk communication also known as spam is the practise of sending unwanted email messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients [1].

The sending of unsolicited bulk communications with the intention to advertise products and generate sales is economically viable because senders have no operating costs beyond the management of their mailing lists. Because the cost of setting up a spamming operation is low spammers are numerous. Thus the volume of unsolicited bulk communications has increased dramatically over the past few years [2].

The costs of spam which involve lost productivity and fraud, these costs are borne by the general public, institutions that store and retrieve email for their employees and by Internet service providers (ISP's). Institutions and ISP's have been forced to add extra capacity to cope with the high volumes of unsolicited bulk communications [3]. Anti spamming legislation has been introduced in many jurisdictions. The problem faced by law enforcement is that spammers move their operations to jurisdictions that have no or weak anti spamming laws. At present it is very difficult to trace the identity of spammers who use identity concealment techniques. It is difficult to determine the identity of the spammer by just analysing the electronic trail using standard email tracing techniques. This paper line proposes the design and implementation of a spammer detection system that uses artificial intelligence techniques to detect abnormal behaviour on a network.

The remainder of the paper is structured as follows. The background section defines spam in more detail and also defines its cost and causes.

The next two sections are devoted to the state of the art of spamming techniques and how to trace spammers. More specifically, these two sections contrast each other in the sense that the third section looks at how spammers use bot-networks to conceal their identities, whereas the fourth section looks at techniques for tracing the identity of spammers using bot-networks. The paper then proposes the design of a spammer detection system to detect bot-networks in section V followed by the conclusion.

.

II. BACKGROUND

Unsolicited bulk email otherwise known as spam is an email sent to a large number of email addresses, where the owners of those addresses have not asked for or consented to receive the mail [4]. Spam is used to advertise a service or a product. An example of spam is an unsolicited email message from an unknown or forged address advertising Viagra.

Spam is one of the most significant threats to the Internet, accounting for around 60% of all email traffic [4]. Spam costs consumers and ISPs vast amount of money in bandwidth charges.

Spammers generally do not pay much for the sending of spam. They accomplish this by exploiting open mail servers to do their task for them. The spammer need only send one email message to an incorrectly configured mail server to reach a vast number of email addresses. Recipients in turn need to pay access costs or telephone costs in order to receive content they did not ask for.

ISPs have to bear the bulk of the cost for bandwidth overuse by spammers. This cost is often passed onto the consumer through increased Internet access fees or a degraded service level.

The following section describes how spammers use bot-nets to send spam and how they conceal their identities from persons who would attempt to identify the source of spam mail.

III. HOW SPAMMERS CONCEAL THEIR IDENTITIES USING BOT-NETWORKS

A Bot-Network consists of a set of machines that have been taken over by a spammer using Bot software sent over the internet. This Bot software hides itself on its host machine and periodically checks for instructions from its human Bot-Network administrator. Bot-nets today are often controlled using Internet Relay [5]. The owner of the computer usually has no idea that his machine has been compromised until its internet connection is shut down by an ISP. As most ISP's block bulk mail if they suspect it is spam the spammers who control these Bot-Networks typically send low volumes of mail at any one time so as not to arouse suspicions. Thus the spam mail can be traced to an innocent individuals network address and not the spammers network address.

Manuscript received June 7, 2009. Ickin Vural is with the Information and Security Architectures Research Group, Department of Computer science, University of Pretoria, Pretoria, South Africa (e-mail: ickin@tuks.co.za).

Prof Hein Venter is the leader of the Information and Security Architectures Research Group, Department of Computer science, University of Pretoria, Pretoria, South Africa (e-mail: hventer@cs.up.ac.za).

While the number of Bot-nets appears to be increasing, the number of bots in each Bot-net is actually dropping. In the past Bot-nets with over 80 000 machines were common [5]. Currently Bot-nets with a few hundred to a few thousands infected machines are common. One reason for this is that smaller Bot-nets are more difficult to detect.

IV. IDENTIFYING THE IDENTITY OF SPAMMERS BY USING HONEYPOTS ON BOT-NETWORKS

A honeypot is a closely monitored computing resource that is intended to be compromised [6]. A honeypot computer can be applied to Bot-networks, open proxies and open proxies. Thus by setting up a computer to imitate a Bot-network, investigators can attempt to trap the spammers into revealing their network addresses.

One way of identifying spammers is to set up a computer to pretend that it is part of a Bot-network [7]. By allowing the honeypot computer to become part of the Bot-network we can obtain the Bot-network software used by the spammer. Once this has been done the honeypot waits for the spammer to send new instructions and then identifies the network address of the sender. The problem with this approach is that spammers send the instructions over open relays and open proxies thus it may be impossible to discover the identity of the spammer's network address in this way.

An open proxy is a machine that allows computers to connect through it to other computers on the internet. Open proxies exist because they enable unhindered internet usage in countries that restrict access to certain sites for political or social reasons. An internet user in a country that restricts internet access can access blocked sites by using an open proxy in a country that does not restrict internet access.

Spammers use open proxies to hide their network addresses. The recipient of a spammers email will not see the spammers' network address on the email but the open proxy's network address. It is estimated that sixty percent of all spam is sent using an open proxy [7]. Thus the spammer will use an open proxy to send instructions to the machines on their bot-network to avoid detection.

V. DISCUSSION

A. *Is Spammer identification possible?*

This paper outlines the challenges facing digital forensic investigators when attempting to identify spammers using bot-networks in conjunction with open proxies. The use of bot-networks means that even if the source of the machine sending the spam is identified the person owning the machine may not be the one responsible for sending spam. The use of untraceable internet connections and open proxies to communicate instructions to bot-networks makes the use of Honeypots unlikely to succeed.

Thus any success in tracing spammers will be matched by spammers using increasingly sophisticated techniques to evade detection. Greater responsibility will have to shift to ISP's in monitoring connections to open proxies as well as attempting to shut down open relays. Nevertheless an arms race between spammers and digital forensic investigators will continue for the foreseeable future.

B. *Bot-net identification*

This paper proposes the design and implementation of a system to detect spammers by analysing network traffic for abnormal behaviour. The implementation would have to take into account spam email sending patterns to effectively identify spammers. The implementation could make use of artificial intelligence to learn behaviour and thus detect abnormal behaviour.

The proposal would be to model a network as a graph and then train an artificial intelligence agent to learn expected and unexpected behaviour so as to detect a machine that could possibly have been taken over by a bot-network.

VI. CONCLUSION

This paper outlines the challenges facing digital forensic investigators when attempting to identify spammers. The paper promotes the idea of Spammer identification as opposed to Spam identification system to halt spam.

ACKNOWLEDGMENT

The Authors would like to thank the ICSA research group members for their input.

REFERENCES

[1] Spamhaus. 2009 The Definition of spam. Available: http://www.spamhaus.org/definition.html [April 2009].

[2] Email Metrics Program. 2007. 'The Network operators perspective' , messaging Anti-Abuse working group. Available: http://www.maawg.org/about/MAAWG20072Q_Metrics_Report.pdf. [April 2009]

[3] Europa. 2009 Data protection: "Junk" e-mail costs internet users 10 billion a year worldwide.Available: h'ttp://europa.eu/rapid/pressReleasesAction.do?reference=IP/01/154&format=HTML&aged=0&language=EN&guiLanguage=en' [April 2009].

[4] Internet Service Providers' Association, 2008 Available: http://www.ispa.org.za/spam/whatisspam.shtml. [April 2009].

[5] Evan Cooke, Farnam Jahanian, Danny McPherson. 2005 . The advanced computing systems association. [Online] The Zombie Roundup Understanding, Detecting, and Disrupting Botnets. Available: http://www.usenix.org/events/sruti05/tech/full_papers/cooke/cooke_html/, [April 2009].

[6] Niels Provos. 2004. The advanced computing systems association. [Online]. A Virtual Honeypot Framework. Available: http://www.usenix.org/event/sec04/tech/full_papers/provos/provos_html/. [April 2009]

[7] Boneh, Dan. 2004. The Difficulties of Tracing Spam Email. Department of Computer Science Stanford University.