IGI PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

ITJ 5038

This paper appears in the publication, International Journal of Digital Crime and Forensics, Volume 1, Issue 3
edited by **Chang-Tsun Li © 2009, IGI Global**

Spam Image Clustering for Identifying Common Sources of Unsolicited Emails

Chengcui Zhang, University of Alabama at Birmingham, USA

Xin Chen, University of Alabama at Birmingham, USA

Wei-Bang Chen, University of Alabama at Birmingham, USA

Lin Yang, University of Alabama at Birmingham, USA

Gary Warner, University of Alabama at Birmingham, USA

ABSTRACT

In this article, we propose a spam image clustering approach that uses data mining techniques to study the image attachments of spam emails with the goal to help the investigation of spam clusters or phishing groups. Spam images are first modeled based on their visual features. In particular, the foreground text layout, foreground picture illustrations and background textures are analyzed. After the visual features are extracted from spam images, we use an unsupervised clustering algorithm to group visually similar spam images into clusters. The clustering results are evaluated by visual validation since there is no prior knowledge as to the actual sources of spam images. Our initial results show that the proposed approach is effective in identifying the visual similarity between spam images and thus can provide important indications of the common source of spam images. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: *Botnet; Clustering; Computer Forensics; Cybercrime; Data Mining; Spam Image*

INTRODUCTION

Spamming is a problem that affects people all over the world. Spam is an unsolicited email which has been sent to many people. There can be legal spam, where the sender gave proper contact information and also has an option to no longer receive the messages. However, in almost all situations, spam is illegal. It is an unsolicited mail that the recipient did not ask to receive and did not give the sender permission to send. Spam falsifies the sender information to prevent anyone from finding out where it has been sent from. Botnets are machines that keep on sending spam. Today, botnets are the main choice for cyber criminals who seek to conceal their identities by using third-party computers as vehicles for their crimes (www.cnn.com/2007/TECH/11/29/fbi.botnets). The FBI has identified at least 2.5 million unsuspecting computer users who have been victims of botnet activities (www.cnn.com/2007/TECH/11/29/fbi.botnets). Spam sometimes attempts to sell a product, convey some messages, or they might also try to trick the recipient to become infected, or attempt to lure them into visiting a website that can infect them.

Spam can cause a lot of problems to internet users. More than 90% of the emails sent on the internet are spam. Billions of dollars of counterfeit software, electronics, as well as shoes, watches, etc., are being sold because of spam advertisements. In this way, huge financial loss occurs to the companies. Spam emails, claiming to be from banks, might also lure users to give their usernames and passwords. Besides software piracy and viruses, spam is also the primary means of phishing and identity theft. Therefore, spam email analysis is one of the most important topics in cyber security. The most effective way of controlling spam emails at the moment is filtering (Carreras & Mrquez, 2001; Clark, Koprinska, & Poon, 2003; Drucker, Wu, & Vapnik, 1999; Sanpakdee, Walairacht, & Walairacht, 2006). However, filters can only differentiate spam emails from non-spam emails but cannot tell the origins of spam. In order to hide their origins, escape detection and spam filter analysis, and to conceal the fact that there are relatively few organizations creating the vast majority of these unsolicited emails, criminals use a variety of intentional obscuring techniques. For example, one of the techniques is to present text primarily as an image, to avoid traditional computer-based filtering of the text. Spam images are sent for two reasons: 1) for advertisement purposes; 2) to hide the textual contents of an email from spam filters. Having no words in the message will not allow spam filters to understand the nature of the message.

Spam images are harder to detect than text spam. Spam images are created when text is embedded into images and content obscuring technologies are used to defeat spam blocking techniques. Spammers use certain methods to defeat traditional anti-spam technologies such as fingerprinting (e.g., md5 (Rivest, 1992)), OCR, and URL blocklist.

1. A text can be embedded in an image which appears as normal text to the recipient but the spam blocking technologies will never be able to “see” the text as it is actually an image.
2. Spammers vary the space between words and lines and also add random speckles to make messages look different to different recipients, though all of them have the same text. By this way, they evade fingerprinting technology such as md5 (Rivest, 1992) by making the images appear unique to standard spam analysis.
3. Use of different colors and varying font size makes it impossible for OCR techniques to find out spam. Also, splitting up one word into two halves with a gap in between deceives OCR techniques.