A Model for Identifying Phishing E-Mail Based on Structural Properties

A.S. Zadgaonkar, Suraj Prasad Keshari, Savita Ajay

Abstract: The widespread use of email caused the number of warnings being made about the dark side our technological revolution to increase and we are becoming uniquely vulnerable to many mysterious and malicious threats. Viruses, worms, and other forms of malicious software started targeting our email inboxes to propagate. Spam and other forms of unsolicited bulk electronic commerce started filling our email inboxes and invading our privacy. Phishing and other forms of fraud attacks have been using email as their primary communication channel to trick users into giving out their credentials. Email could have been a killer application for the Internet if none of the problems mentioned above exist.

Index Terms—Phishing, email, privacy, software, Virus, computing.

I. INTRODUCTION

There are an increasing number of emails purporting to be from a trusted entity that attempt to deceive users into providing account or identity information, commonly known as “phishing” emails. Traditional spam filters are not adequately detecting these undesirable emails, and this causes problems for both consumers and businesses wishing to do business online. From a learning perspective, this is a challenging problem. At first glance, the problem appears to be a simple text classification problem, but the classification is confounded by the fact that the class of “phishing” emails is nearly identical to the class of real emails.

II. PHISHING ATTACK

1. URL and Host Name Obfuscation Attacks*:* Phishing attacks require that the victims visit the phisher's website by making them believe that the forged website is the real one. This is achieved through URL and other hostname obfuscation techniques using DWORD, HEX, UTF-8, and other encodings appearing in the characteristic e-mail. Therefore, to circumvent these forms of obfuscation attacks, as a feature for detection we also consider URL untangling tools. However, a significant majority of the present day e-mails have the URLs displayed in dotted decimal format, thereby increasing the suspicious factor.

2. Embedded e-mail Attachment: E-mails posing to appear from legitimate domain may contain embedded HTML forms requesting the user’s credit card numbers and other financial information. As the existing browser based defense solutions fail to detect these attacks, in order to protect against these attacks, the body of the received messages is parsed and HTML forms with suspicious field names are immediately tagged as malicious.

3. Browser Vulnerabilities: Browsers in their quest to accommodate increased features and functionalities, accidentally inject security loopholes making them prone to phishing attacks. Browsers are also susceptible to homographic attacks like International Domain Name (IDN) spoofing and pop-up hijacking. Also, vulnerabilities in ActiveX controls and other browser helper objects (BHO) can install Trojans, which can modify the system's /etc/hosts file to redirect the request of a legitimate website to a phisher's IP address. With new vulnerabilities being discovered and patches released, it becomes extremely difficult for a naive user to constantly update and protect against the attacks. Disabling vulnerable features like ActiveX controls, Java runtime environments (JRE), IDN support is also viewed as a trade-off between extended functionality and security. As we have restricted our current setup to target e-mail based phishing attacks, some of these approaches can circumvent our classification framework.

4. Cross-site Scripting (XSS) and Session Hijacking Attacks: Phishers can also exploit the security loopholes in web applications and web server's software to make the users unknowingly execute malicious scripts. These scripts are usually embedded through encoded characters in the URL for the purpose of redirecting the users to a malicious server. Also, by installing packet sniffers and extracting session ID from the server side exploits, it is possible for the phishers to hijack the user's current session. Since these attacks do not propagate via e-mail messages, it is beyond the scope of our proposed work.

III. PROBLEM STATEMENT

With the widespread usage of Internet for online banking and trade, phishing attacks and allied form of identity theft based scams are becoming extremely popular among the hacker communities. The anonymity in the Internet, coupled with the potential for large financial gains, serves as a strong motivation to perpetrate such seemingly low risk, yet high return crimes. In 2004 alone, more than 50 million phishing e-mails were sent out resulting in 10 billion dollars of damage to banks and financial institutions alike. Phishing attacks pose a serious threat to end-users and commercial institutions alike.

Manuscript received on May 9, 2013

Dr. A.S. Zadgaonkar, vice chancellor, Dr. C.V.Raman University, Kargiroad, Kota, Bilaspur(C.G.),India

Suraj Prasad Keshari, Asst. Professor, Dr. C.V.Raman University, Kargiroad, Kota, Bilaspur(C.G.),India

Savita Ajay, Research scholar, Dr. C.V.Raman University, Kargiroad, Kota, Bilaspur(C.G.),India

Majority of the present day phishing attacks employ e-mail as their primary carrier, in order to allure unsuspecting victims to visit the masqueraded website.

IV. WAYS OF PHISHING ATTACKS

Phishing was a term originally used to describe email attacks that were designed to steal your online banking username and password. However, the term has evolved and now refers to almost any email-based attack. Phishing uses social engineering, a technique where cyber attackers attempt to fool you into taking an action. These attacks often begin with a cyber criminal sending you an email pretending to be from someone or something you know or trust, such as a friend, your bank or your favorite online store. These emails then entice you into taking an action, such as clicking on a link, opening an attachment or responding to a message. Cyber criminals craft these emails to look convincing, sending them out to literally millions of people around the world. The criminals do not have a specific target in mind, nor do they know exactly who will fall victim. They simply know the more emails they send out, the more people they may be able to fool. Phishing attacks work one of four ways:

1. Harvesting Information: The cyber attacker’s goal is to fool you into clicking on a link and taking you to a website that asks for your login and password, or perhaps your credit card or ATM number. These websites look legitimate, with exactly the same look, imagery and feel of your online bank or store, but they are fake websites designed by the cyber attacker to steal your information.

2. Infecting your computer with malicious links: Once again, the cyber attacker’s goal is for you to click on a link. However, instead of harvesting your information, their goal is to infect your computer**.** If you click on the link, you are directed to a website that silently launches an attack against your computer that if successful, will infect your system.

3. Infecting your computer with malicious attachments: These are phishing emails that have malicious attachments, such as infected PDF files or Microsoft Office documents. If you open these attachments they attack your computer and, if successful, give the attacker complete control.

4. Scams: These are attempts by criminals to defraud you. Classic examples include notices that you’ve won the lottery, charities requesting donations after a recent disaster or a dignitary that needs to transfer millions of dollars into your country and would like to pay you to help them with the transfer. Don’t be fooled, these are scams created by criminals who are after your money.

V. RELATED WORKS

There are only a few research efforts that focus entirely on tackling the problem of phishing attacks. Phishing e- mails are often related to spam and most of these techniques target spam control as a mechanism to prevent such identity theft scams. The primary difference is that the spam messages lack proper feature selection that appropriately demarcates spam from phishing messages. In this section we briefly review these approaches to put our work in perspective.

Application of support vector machine for classification is diverse. Vapnik et al. [4] has shown the usability of SVM for spam classifications. They also compared their algorithm with other techniques such as boost trees and Inverse document frequency (IDF) metric. For binary classification tests, in a comparatively less training time SVM achieved the highest detection and least false positive rates. SVM has been successfully used in other areas of computer security like e-mail author attribution [8], text classification [6], masquerade detection [11], document forensics etc.

To illustrate the need for a better approach, we first discuss the weaknesses present in the existing schemes. Although these schemes are sufficient in majority of scenarios, they need to be complemented with other approaches to improve the overall performance. Several commercial and open-source toolbars exist that perform spoof tests and verify SSL certificates for establishing the validity of a website. Spoofstick [3] is widely used tool which employs reverse DNS lookup on the visited website, displaying the site's actual IP address on its toolbar. Although it can detect simple URL obfuscation, it still requires human in the loop to make the decision. NetCraft [4] anti-phishing toolbar is another monitoring tool that engages client-server architecture to detect phishing attacks. Each user with the toolbar acts as clients, who actively report masqueraded websites to the server. The server is responsible for processing the incoming requests and informing its client about the authenticity of the website. As these techniques rely on user's feedback for its decision making, it may be subjected to increased false positives and denial of service attacks in cases where a group of hackers may tag a legitimate website malicious. Also, since the masqueraded websites are short-lived, it is highly unlikely that such responses are propagated to the clients before their lifetime. Tools, which depend on black lists for detection, also suffer from these drawbacks.

Key distribution and identity based digital signatures have been proposed to make e-e-mail messages trustworthy. S/MIME, PGP [9] and GPG [12] are popularly adopted standards for digitally signing e-mail messages which are supported by most of the GUI mail clients. As these methods encrypt the outgoing e-mails along with the sender's identity, it makes them resilient to e-mail spoofing. However at this point not all web based mail clients like Yahoo! Mail, Hotmail, Gmail support S/MIME. In the case of PGP/GPG schemes, as there is no central authority server which could verify the e-mails, a phisher may infiltrate the web of trust and digitally sign his e-mails. Also, another drawback of this approach is that it necessitates that both the sender and receiver have the compatible infrastructure to support digital signing and verification. Other techniques like smartcards, one- time passwords [10] are used to prevent phishing attacks. However, these are beyond the scope of comparison.

VI. STRUCTURE OF PHISHING E-MAIL

In this section we discuss the common structure used across all the phishing e-mails with the intention of identifying a set of generic features to be used for classification. Drake et al. [2] also provide a concrete summary of the anatomy of current day phishing e-mails.

1. Spoofing of online banks and retailers: Since phishing e-mails must resemble online banking and retailers to gain the trust of the users in divulging their information, the phishers in the e-mails mimic the appearance of a reputable company. The companies spoofed most often are Citibank, eBay, and PayPal. The most targeted industry is financial services. Internet retailers and Internet service providers are also targeted. The audacity of phishers was also evident from the recent phishing attacks impersonating Internal Revenue Service (IRS), appearing to return the tax refund via the Internet. This is done by primarily using the company's image and through links referring to the company's website in the fake e-mail.

2. Link in the text is different from the destination: In spoofed e-mail messages, the link text seen in the e- mail is usually different from the actual link destination. In the following example, though the e-mail refers to "http://www.chase.com," it redirects the user discretely to the site http://www.climagro.com.ar/agro/chase.htm <a class="m1" target="_blank"title="LOGIN"href="http://www.climagro.com.ar/agro/chase.ht m>http://account.earthlink.com</a>.

3. **Using IP addresses instead of URLs:** Frequently, phishers attempt to conceal the destination website by obscuring the URL. One method of concealing the destination is to use the IP address of the Web site, rather than the hostname. An example of an IP address used in a fraudulent e-mail message's URL is "http://210.14.228.66/sr/." Also, the URL can be hidden through representation in DWORD, Octal, or Hexadecimal format.

4. Generalization in addressing recipients As the success of e-mail based phishing attacks rely on the law of large numbers, most of the phishing e-mails do not contain personalized content while addressing their recipients. Also, unlike legitimate business communication, they do not address the customers using their names for identifiers, and lack embedded scrambled information such as 'last four digits of account information', which is used to establish authenticity. Although, it might be possible for a phisher to include these information, by employing social engineering and other malpractices, the success rate of such attacks are limited; it is hard to target wide range of users.

5. Usage of well-defined situational contexts to lure victims: As discussed early in Section 1, most of the phishing e- mails use the underlying contexts such as invoking a sense of false urgency, threat, wheedle, and concern to deceit the users in clicking on the visited hyperlink. Therefore it is important to build such context graphic models for detection.

.

VII. PROPOSED METHODOLOGY

To accomplish the proposed solution the following methodology will be adapted:

1. We apply simulated annealing for feature selection, which is a well suited approximation measure for locating global optimum in a large search space.

2. Once the appropriate strong feature set has been selected, ranking of features is applied to categorize the chosen features based on the relevance.

3. We apply concepts from information theory, such as information gain (IG) to rank these selected features.

4. We adopted SVM as our underlying algorithm because it has been widely used in text classification applications, and especially in the field of computer security in the context of spam detection, hidden e-mail construction, authorship attribution and masquerade detection

VIII. CONCLUSION

This research is centered on "Identifying Phishing E-Mail Based on Structural Properties". The outcome of the proposed work will likely to yield expected result and fulfill the following objective:

1. Secure email access.

2. Prevent email phishing attack.

REFERENCES

1. M. *Chandrasekaran*, R. Chinchani and S. Upadhyaya, PHONEY: Mimicking user response to detect phishing attacks, To appear at TSPUC 2005 Workshop, affiliated with IEEE WoWMoM.
2. Christine E. Drake, Jonathan J. Oliver, and Eugene J. Koontz, *Anatomy of Phishing E-mail First Conference on E-mail and Anti-Spam*, 2004.
3. CNET News, *Phishing attacks skyrocket in 2004*, 2004.
4. Harris Drucker, Donghui Wu, and Vladimir N. Vapnik, Support vector machines for Spam categorization, IEEE-NN, 10 (1999), pp. 1048--1054.
5. Debuse, JCW and VJ Rayward-Smith, *Feature subset selection within a simulated annealing data mining algorithm*, Journal of Intelligent Information Systems, (1997).
6. T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proc. 10th European Conference on Machine Learning {ECML}-98, 1998, pp. 137-142.
7. C. Neil, L. Robert, T. Yuka and C. M. John, Client- Side Defense Against Web-Based Identity Theft, 2004.
8. Olivier de Vel, Alison Anderson, Malcolm Corney and George Mohay, Mining Email Content for Author Identification Forensics., SIGMOD: Special Section on Data Mining for Intrusion Detection and Threat Analysis, (2001).
9. S/MIME and OpenPGP.
10. M. Schwartz, Putting Next-Generation Smart Cards to Work, (2005).
11. Ke Wang and Sal Stolfo, One Class Training for Masquerade Detection, ICDM Workshop on Data Mining for Computer Security (DMSEC 03), 2003.
12. The GNU Privacy Gaurd, http://www.gnupg.org.
13. Netcraft. toolbar, http://toolbar.netcraft.com
14. Spoof-stick. toolbar, http://www.corestreet.com/spoofstick.
15. V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

16. Gregory L. Wittel and S. Felix Wu, On Attacking Statistical Spam Filters, First Conference on E-mail and Anti-Spam, 2004
17. “Phishing activity trends report,” Anti-Phishing Working Group, Tech. Rep., Jan. 2005.
[Online]. Available: http://www.antiphishing.org/reports/apwg report jan 2006.pdf
18. N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, “Client-side defense against
web-based identity theft.” in NDSS, 2004. [Online]. Available: http://www.isoc.org/isoc/
conferences/ndss/04/proceedings/Papers/Chou.pdf
19. “Netcraft toolbar,” 2006. [Online]. Available: http://toolbar.netcraft.com/
20. A. Alsaid and C. J. Mitchell, “Installing fake root keys in a pc.” in EuroPKI, 2005, pp.227–239. [Online]. Available: http://dx.doi.org/10.1007/11533733 16
21. “Mozilla thunderbird,” 2006. [Online]. Available: http://www.mozilla.com/thunderbird/
22. B. Leiba and N. Borenstein, “A multifaceted approach to spam reduction,” in Proceedings
of the First Conference on Email and Anti-Spam (CEAS), 2004. [Online]. Available:
http://www.ceas.cc/papers-2004/127.pdf
23. W. Cohen, “Learning to classify English text with ILP methods,” in Advances in Inductive Logic Programming, L. De Raedt, Ed. IOS Press, 1996, pp. 124–143. [Online]. Available: citeseer.ist.psu.edu/cohen96learning.html
24. P. Graham, “Better bayesian filtering,” in Proceedings of the 2003 Spam Conference, Jan 2003. [Online]. Available: http://www.paulgraham.com/better.html
25. M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A bayesian approach to filtering
junk e-mail,” in Learning for Text Categorization: Papers from the 1998 Workshop. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998. [Online]. Available: http://robotics.stanford.edu/users/sahami/papers-dir/spam.ps
26. I. Rigoutsos and T. Huynh, “Chung-kwei: a pattern-discovery-based system for the automatic identification of unsolicited e-mail messages (spam),” in Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004. [Online]. Available: http://www.ceas.cc/papers-2004/153.pdf