Analyzing Social and Stylometric Features to Identify Spear phishing Emails

Prateek Dewan*, Anand Kashyap†, Ponnurangam Kumaraguru*
*Indraprastha Institute of Information Technology, Delhi
*Cybersecurity Education and Research Centre (CERC), IIIT-Delhi
†Symantec Research Labs
*{prateekd,pk}@iiitd.ac.in, †anand_kashyap@symantec.com

ABSTRACT

Spear phishing is a complex targeted attack in which, an attacker harvests information about the victim prior to the attack. This information is then used to create sophisticated, genuine-looking attack vectors, drawing the victim to compromise confidential information. What makes spear phishing different, and more powerful than normal phishing, is this contextual information about the victim. Online social networks can be one such source for gathering vital information about an individual. In this paper, we present a machine learning model to detect spear phishing emails sent to employees of 14 international organizations by combining email features with *social* features extracted from LinkedIn. We obtained a true positive dataset of spear phishing, spam, and phishing emails from Symantec's enterprise email scanning service. Features extracted from these emails were combined with features extracted from publicly available information on recipients' LinkedIn profiles. We applied various machine learning algorithms on this data and achieved a maximum accuracy of 97.76% in identifying spear phishing emails. The same model achieved a slightly better accuracy of 98.28% without the *social* features, signifying that *social* features extracted from LinkedIn did not help in identifying spear phishing. To the best of our knowledge, this is one of the first attempts to make use of a combination of stylometric features extracted from emails, and *social* features extracted from an online social network to detect spear phishing emails.

1. INTRODUCTION

There have been numerous reports of spear phishing attacks causing losses of millions of dollars in the recent past. [1] [2] Although there exist antivirus, and other similar protection software to mitigate such attacks, it is always better to stop such vectors at the entry level itself [4]. This requires sophisticated techniques to deter spear phishing attacks, and identify malicious emails at a very early stage itself. Spear phishing emails usually contain victim-specific context instead of general content. Since it is targeted, a spear phishing attack looks much more realistic, and thus, harder to detect [3].

[1] http://businesstech.co.za/news/internet/56731/south-africas-3-billion-phishing-bill/

[2] http://www.scmagazine.com/stolen-certificates-used-to-deliver-trojans-in-spear-phishing-campaign/article/345626/

In this work, we attempt to identify spear phishing emails by leveraging victim-specific information from the LinkedIn social network. We attained a dataset of spear phishing emails, spam and phishing emails from the Symantec's enterprise email scanning service, which is deployed at multiple international organizations around the world. We extracted the most frequently targeted organizations from these emails, and filtered out 14 organizations for which, the recipients' first name and last name could be derived from the email address. We also used a random sample of 6,601 benign emails from the Enron email dataset [1]. We extracted 18 stylometric features from these emails and combined them with 9 social features from LinkedIn profiles of the recipients of these emails. These features vectors were subjected to four classification algorithms. We achieved a maximum accuracy of 97.76% for classifying spear phishing, and non spear phishing emails using a combination of *email* features, and *social* features. However, without the *social* features, we were able to achieve a slightly higher accuracy of 98.28% for classifying these emails. We then looked at the most informative features, and found that *email* features performed better than *social* features at differentiating targeted spear phishing emails from non targeted spam / phishing emails, and benign Enron emails.

2. DATA COLLECTION

Symantec collects data regarding targeted attacks that consist of emails with malicious attachments. These emails are identified from the vast majority of non-targeted malware by evidence of there being prior research and selection of the recipient, with the malware being of high sophistication and low copy number. The corpus may omit some attacks, and most likely also includes some non-targeted attacks, but nevertheless it represents a large number of sophisticated targeted attacks compiled according to a consistent set of criteria which render it a very useful dataset to study.

We identified and extracted the most attacked organizations (excluding free email providing services like Gmail, Yahoo, Hotmail etc.) from the domain names of the victims' email addresses, and picked 14 most frequently attacked organizations. We were however, restricted to pick only those organizations, where the first names and last names were easily extractable from the email addresses. The first name and last name were required to obtain the corresponding LinkedIn profiles of these victims using LinkedIn's People Search API. This restriction, in addition to removal of duplicates, left us with a total of 4,742 targeted spear phishing

emails sent to 2,434 unique victims (referred to as *SPEAR* in the rest of the paper); 9,353 non targeted attack emails sent to 5,912 unique non victims (referred to as *SPAM* in the rest of the paper), and 6,601 benign emails from the Enron dataset, sent to 1,240 unique Enron employees (referred to as *BENIGN* in the rest of the paper).

3. METHODOLOGY

We performed classification using a) *email* features [3]; b) *social* features, and c) combination of these features. We compared these three accuracy scores across a combination of datasets viz. SPEAR versus SPAM emails from Symantec's email scanning service, SPEAR versus benign emails from BENIGN dataset, and SPEAR versus a mixture of emails from BENIGN, and SPAM from the Symantec dataset. Not all *email* features were available for all the three email datasets. The BENIGN dataset did not have attachment related features, and the *body* field was missing in the SPAM email dataset. We thus used only those features for classification, which were available in both the targeted, and non targeted emails. In all, we used four machine learning algorithms, and a total of 27 features; 18 stylometric, and 9 *social* for our analysis. The entire analysis and classification tasks were performed using the Weka data mining software [2]. We applied 10-fold cross validation to validate our classification results.

4. RESULTS

Table 1 presents the detailed results of our analysis where we subjected SPEAR and BENIGN + SPAM emails from Symantec to machine learning classification algorithms. We found that two out of the four classifiers performed better with a combination of email and social features, while two classifiers performed better with only *email* features. However, the overall maximum accuracy was achieved using a combination of *email* and *social* features (89.86% using Random Forest classifier). This result was in contradiction with our analysis of SPEAR versus SPAM, and SPEAR versus BENIGN separately, where *email* features always performed better independently, than a combination of *email* and *social* features. [4] Our overall maximum accuracy, however, dropped to 89.86% because of the absence of *attachment* features in this dataset. Although the *attachment* features were available in the SPAM dataset, their unavailability in BENIGN forced us to remove this feature for the current classification task. Eventually, merging the SPAM email dataset with BENIGN reduced our email dataset to only 7 features, all based on the email "subject".

As mentioned earlier, combining the SPAM email dataset with BENIGN largely reduced our *email* feature set. We were left with 7 out of a total of 18 email features. Understandably, due to this depleted *email* feature set, we found that the email features did not perform as good as *social* features in this classification task. Despite being fewer in number, the *subject* features, viz. *Subject_richness* (information gain: 0.1829) and *Subject_numChars* (information gain: 0.1050) were found to be two of the most informative features. However, the information gain value associated with both these features was fairly low. This shows that even being the best features, the *Subject_richness* and *Subject_numChars* were not highly distinctive features amongst spear phishing, and non spear phishing emails.

Feature set	Classifier	Random Forest	J48 Decision Tree	Naive Bayes	Decision Table
Subject (7)	Acc. (%)	86.48	86.35	**77.99**	**85.46**
	FP rate	0.333	0.352	0.681	0.341
Social (9)	Acc. (%)	88.04	84.69	74.46	80.61
	FP rate	0.241	0.371	0.454	0.432
Email + Social (16)	Acc. (%)	**89.86**	**88.38**	73.97	84.14
	FP rate	0.202	0.248	0.381	0.250

Table 1: Accuracy and false positive rates for SPEAR emails versus SPAM + BENIGN emails.

5. DISCUSSION

There can be multiple reasons for our results being non-intuitive. Firstly, the amount of social information we were able to gather from LinkedIn, was very limited. It is likely that in a real-world scenario, an attacker may be able to gain much more information about a victim prior to the attack. This could include looking for the victim's profile on other social networks like Facebook, Twitter etc., looking for the victim's presence on the Internet in general, using search engines (Google, Bing etc.), and profiling websites like Pipl [5], Yasni [6] etc. The process of data collection by automating this behavior was a time consuming process, and we were not able to take this approach due to time constraints. Secondly, it was not clear that which all aspect(s) of a user's social profiles were most likely to be used by attackers against them. It is possible that none of the features we used in our analysis were used by attackers to target their victims. We have no way to verify that the spear phishing emails in our dataset were even crafted using features from social profiles of the victims. These reasons, however, only help us in better understanding the concept of using social features in spear phishing emails.

6. ACKNOWLEDGEMENT

We would like to thank Symantec for providing us with the email data that we used for this work. We would also like to thank all members of CERC@IIITD for their support.

7. REFERENCES

[1] W. W. Cohen. Enron email dataset. *Internet: www. cs. cmu. edu/enron/,(Date Last Accessed on May 25, 2008)*, 2009.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[3] M. Jakobsson and S. Myers. *Phishing and countermeasures: understanding the increasing problem of electronic identity theft*. John Wiley & Sons, 2006.

[4] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):7, 2010.

[3] We further split email features into *subject*, *body*, and *attachment* features for analysis, wherever available.

[4] These results have been omitted from this paper because of space constraints. Please refer to the full paper for details.

[5] https://pipl.com/

[6] http://www.yasni.com/