Phishing Email Detection in Czech Language

Vít LISTÍK[1], *Tomáš GOGÁR*[1]

[1]Dept. of Computer Science, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

listivit@fel.cvut.cz, gogartom@fel.cvut.cz

Abstract. *Current phishing email detection methods are mainly based on url blacklisting. Goal of this work is to detect phishing emails without blacklists. State of the art techniques were evaluated and decision tree classifier based on 26 features was trained on public phishing data set. Promising results of this approach on testing data set were not proved in live traffic. Used data set is not representative, most probably because it contains old emails. New solution consists of two phases. First phase is prefiltering and second is phishing detection itself. Prefiltering phase is used to reduce heavy computations and it consists of two steps. First step is based on 30 traffic statistics features. This step directly modifies score because traffic statistics for phishing emails are not available for training. Second phase uses content features and decision tree classifier trained on gathered data set.*

The second phase detects sender domain with usage of domain specific keywords, commonly used image sources, plain links to domains and header from, at first. Secondly it detects most suspicious link and decides whether domain extracted from links is commonly linked by detected domain or not. The final decision is based on the score threshold. The threshold was set via ROC evaluation, which was built on manually classified emails with high scores. In current setup this system is capable of detecting 98% of phishing attacks with 26% of misclassifications.

Keywords

Phishing, email, machine learning, natural language processing, classification.

1. Introduction

Phishing is a type of an electronic identity theft which uses social engineering techniques. Phishing is widely used to steal personal information such as important online accounts (bank or email account) or credit card number. Attackers use these information to directly steal money, to steal the users identity or for other kinds of fraudulent activities, e.g. selling stolen email accounts which are further used as botnets.

Over time, phishing has developed two techniques worth mentioning.

- **Spear-phishing**, a targeted phishing technique in which the attacker gathers publicly known information about victim . It means that the attack is personalized [13].
- **Pharming**, a technique based on DNS spoofing which redirects victim from the legit page to a fraudulent one. Because of the same URL address, the victim does not suspect anything, and thus it is a huge problem [15].

Phishing emails often contains links leading to malicious websites and text urging to take a quick action, e.g. to change the password on the malicious page. The attacks are often successful because phishing messages and target websites look very trustworthy - in many attacks legitimate sources are copied. Phishing websites contain forms luring sensitive data from users and they are published on free web hosting or hacked websites. To maximize phishing success, emails containing links to malicious forms are often sent to many recipients.

Worldwide phishing attacks appearance raised from hundreds to tens of thousands in the last 10 years [8, 9]. In 2014 there was the highest number of phishing attacks ever [6] and September 2014, the latest measured month, brought the number of 53 661 unique phishing emails sent [8]. This number is slightly decreasing over the last years [7]. Over 75% of all phishing attacks were targeted on financial institutions [7] which led to 3.6 billion US dollars loss. 2014 Czech Republic was among the 10 countries with the highest number of hosting malicious websites in September 2014 [5, 8]. Recent research results showed that good phishing websites fooled an alarming number of 90% of all participants [10].

There are several ways how to defend against phishing. It is reasonable not only to educate users, but also to detect phishing to alert them. Those alerting defensive mechanisms can be implemented on client or server side. Client side refers to web browser, email client and anti-virus software. Client security often consists of updated blacklists and shows warning to the end user. Problem of this solution is the lack of the message metadata information in client software, so it has to rely on blacklists., but the blacklisting process is slower than the appearance of new attacks.

Server side may also check blacklists, but it has also more useable information about the traffic. But solutions that are based on machine learning may be very precise when they process a lot of positive examples, but phishing creates only tiny fraction of the traffic, which makes the problem harder. Defensive mechanisms are often based on identity and spoofing detection since the phishing is an identity theft.

Phishing is an illegal activity, not only in the Czech Republic but also in other states. It is classified as the type of a fraud. Phishing causes troubles to individuals as well as to targeted companies and email providers, which are loosing their reputation. That is the reason why Seznam.cz, the biggest Czech freemail provider, requested a phishing detection system that is described in this work.

2. Methods

At the end of 2013 an article was published, showing how highly is this area covered [16]. Many teams achieved very good results on public dataset [3]. Most successful phishing detection methods are based on information included in the messages itself or on message meta information. Big companies usually use custom solutions, but open source solutions are also available. ScamNailer uses lists of addresses from which it generates rules for SpamAssassin [2]. PhishTag is based on link blacklisting and it rewrites malicious links. PhishTag is also connected with SpamAssassin [1].

Public phishing testing state of the art data set is [18]. This phishing data set consists of 4450 emails in mbox format (described in RFC 4155 [12]) sent between 2004 and 2007. SpamAssassin data set [3] is often used as negative sample [3]. It is divided to ham and spam and was collected between 2002 and 2005. It consists of 6047 messages in eml format.

2.1. Content Based Approach

Method based only on content of the message was implemented and trained on publicly available data sets. Used features were based on HTML tags, images, java scripts, headers and mainly text in body and subject of the emails. These features were picked from state of the art solutions [11, 4, 14, 19]. Mainly links were used from HTML tags. For example total number of links, number of domains, number of IP address based links and link schemes were used. Word list gathered from the positive examples was used for text analysis. Another text features were text length, vocabulary richness and number of words in the subject. Exhaustive list of features may be found in [17]

Decision tree model, successfully used in other solutions [5], were trained on 26 features with results shown in table 1.

This promising model was tested in real traffic. The test results showed that the model classifies around 50% messages as phishing, which is not possible in normal traffic. Random samples were picked from the classificated messages and manually labeled. It showed that the model is not classifying phishing correctly. The classifier probably did not work because of the nature of used data set. The examples were pretty old and the features are not enough discriminative for the phishing detection.

2.2. Mimic Detection Based Approach

Content based approach failure led to the solution based on abnormal entity behaviour described on Fig. 1. At first, this solution does prefiltering based on traffic metadata and then the entity detection and behaviour analysis. This solution is based on score and threshold, which is commonly used for spam filtering.

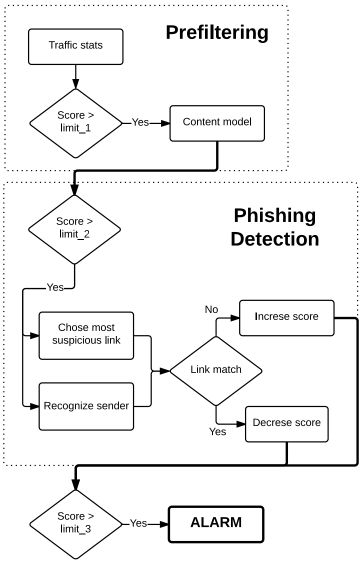

Fig. 1. Phishing detection process.

First phase of the detector is prefiltering. This phase reduces usage of computational resources by filtering of emails which are not malicious. This method is based on traffic statistics and content model.

Traffic statistics are based on history of well known senders. Those senders use DKIM, SPF and DMARC which helps with determining their true identity. It also uses historical data for analyzing their reputation which contains information like when the sender was firstly seen, how many emails were sent and how users reacted to them.

class	precision	recall	f1-score	support
Legit	0.96	0.97	0.96	1591
Phishing	0.94	0.93	0.94	936
avg / total	0.95	0.95	0.95	2527

Tab. 1. Learned decision tree test results.

Content model is based on the text of the email. This model classifies email topic to predefined classes. Some of the predefined classes, like for example advertisement or work proposals are clearly not phishing, but transactional topic emails (containing information about accounts) may indicate that they are phishing emails. This model also uses some features form the content based model (like for example number of links or phishing word list).

It is essential to recognize the entity which is mimicked. Several features like for example text, links or email headers may be used for the entity recognition. This work uses plain text urls, source urls and text for determining sender entity.

- Plain urls are often used to assure the reader that the sender is legit.
- Source urls are significant because attackers are using logos directly from mimicked servers.
- Text is the most significant feature for the entity recognition. This method is based on TF-IDF but instead of inverse document frequency it is using inverse domain frequency. This method is trying to maximize the effect of using company names. Company names, are commonly used in the domain's communication, but it is used only very few times in the communication of other domains. Gathering data for this task is not trivial. We are aggregating number of occurences for each domain-word pair and storing them to persistent storage (Aerospike). Then we are periodically calculating most significant domains for each word using Hadoop.

The final decision is mainly based on detection of using abnormal links for the entity. Links are evaluated for uncommon signs like http links claimed as https, common string changes in the detected domain names and link length. Most important sign is how common is to link to that domain for the detected entity and for other entities. The most suspicious link is compared with the entity link profile. If the link is not commonly used, the email is classified as phishing.

3. Results

Claimed domain recognition was tested on headers. Header from each tested email was extracted and taken as correct result. Our detection method correctly recognizes domain for **77.15%** of tested emails. Domain image sources correctly recognizes domain for **82.19%** of tested emails and text links correctly recognizes domain for **55.81%** of tested domains. When we use all of the methods, we gain **100%** right recognition for tested email sender domains. All three methods agreed on same solution for **23.22%** of test samples.

The system is testing around 50 million emails per day. All these emails cannot be classified by hand, especially when it is supposed that percentage of phishing emails will be very small. 3517 emails with the highest score was manually classified. Distribution of classified emails based on score showed that for score lower than 8 phishing email count is not rising so it is supposed that in lower scores there are also no phishing messages.

The classifier decision is based on threshold, true positive rate (TPR) and false positive rate (FPR) are dependent on threshold setup. How TPF and FPR change according to the score threshold setup is shown on Fig. 2. This metric considers only unique emails, because when using absolute counts, some frequently delivered messages make bigger differences, which is not desired.

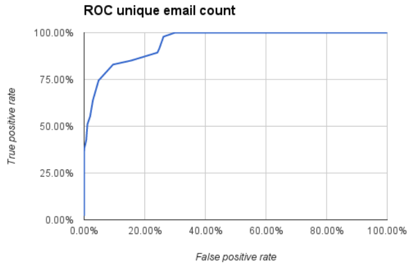

Fig. 2. ROC of phishing email classification for unique emails.

Some of the interesting threshold setups are shown in the table 2

Score	TPR	FPR
21	0.10%	38.30%
17	2.89%	63.83%
16	4.78%	74.47%
11	26.20%	97.87%
10	29.98%	100.00%

Tab. 2. Classifier statistics based on threshold setup.

4. Discussion and conclusions

Choosing score 19 will classify 87.99% of phishing messages correctly with 4.02% of misclassifications. But we decided to use threshold 11 which detects 97.87% of attacks with 26.20% FPR. This FPR means that approximately two hundred unique emails will be misclassified each day. That is quite high number but this result will be used as input for the anti-spam solution so the decision may not be final.

Goal of this work was to find method for detecting phishing emails, implement it and test it on production data. This work shows that phishing may be recognized based on detection of abnormal behaviour of pretended entity (company). Pretended entity may be recognized by several signs like plain text domains, image sources and mainly used keywords. When the entity is recognized, it's link profile is compared to links contained in the message and the class may be determined. Proposed implementation is capable to classify 98% messages correctly with 26% misclassifications. This system in incorporated into the anti-spam production environment.

Acknowledgments

Research described in the paper was supervised by Ing. Jan ediv, CSc., FEE CTU in Prague and supported by the Seznam.cz company.

References

[1] PhishTag. http://search.cpan.org/dist/Mail-SpamAssassin/lib/Mail/SpamAssassin/Plugin/PhishTag.pm. [Online; accessed 2015-02-01].

[2] ScamNailer. http://www.scamnailer.info/documentation.html. [Online; accessed 2015-02-01].

[3] Spamassassin corpus. http://spamassassin.apache.org/publiccorpus/. [Online; accessed 2015-02-01].

[4] Ammar Ali Deeb Al-Mo, Tat-Chee Wan, Karim Al-Saedi, Altyeb Altaher, Sureswaran Ramadass, Ahmad Manasrah, Loai Bani Melhiml, and Mohammad Anbar. An online model on evolving phishing e-mail detection and classification method. *Journal of Applied Sciences*, 11(18):3301–3307, dec 2011.

[5] Ammar Almomani, BB Gupta, Samer Atawneh, A Meulenberg, and Eman Almomani. A survey of phishing email filtering techniques. *Communications Surveys & Tutorials, IEEE*, 15(4):2070–2090, 2013.

[6] APWG. Phishing Activity Trends Report 1st Quarter 2014. http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf. [Online; accessed 2015-02-01].

[7] APWG. Phishing Activity Trends Report 3rd Quarter 2013. http://docs.apwg.org/reports/apwg_trends_report_q3_2013.pdf. [Online; accessed 2015-02-01].

[8] APWG. Phishing Activity Trends Report 3rd Quarter 2014. http://docs.apwg.org/reports/apwg_trends_report_q3_2014.pdf. [Online; accessed 2015-02-01].

[9] APWG. Phishing Attacks Trends Report January 2004. http://docs.apwg.org/reports/APWG.Phishing.Attack.Report.Jan2004.pdf. [Online; accessed 2015-02-01].

[10] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.

[11] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*, pages 649–656. ACM, 2007.

[12] E. Hall Network Working Group. RFC 4155. http://tools.ietf.org/html/rfc4155. [Online; accessed 2015-02-01].

[13] Jason Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[14] Rafiqul Islam and Jemal Abawajy. A multi-tier phishing detection and filtering approach. *Journal of Network and Computer Applications*, 36(1):324–335, jan 2013.

[15] Chris Karlof, Umesh Shankar, J Doug Tygar, and David Wagner. Dynamic pharming attacks and locked same-origin policies for web browsers. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 58–71. ACM, 2007.

[16] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Phishing detection: a literature survey. *Communications Surveys & Tutorials, IEEE*, 15(4):2091–2121, 2013.

[17] Vit Listik. Phishing email detection in czech language for email.cz. 2015.

[18] J. Nazario. Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus. [Online; accessed 2015-02-01].

[19] Cleber K. Olivo, Altair O. Santin, and Luiz S. Oliveira. Obtaining the threat model for e-mail phishing. *Applied Soft Computing*, 13(12):4841–4848, dec 2013.