

Project Exhibition at ECML/PKDD 2009

„AntiPhish - Machine Learning for Phishing Detection“.

Gerhard Paaß, André Bergholz
Knowledge Discovery Department
Fraunhofer IAIS
St. Augustin, Germany

Description of the Project AntiPhish

EU Strep: 2006-2009.

Partners: Fraunhofer IAIS, Symantec IRL, Tiscali (IT), Nortel (FR), K.U. Leuven (BE)

<http://www.antiphishresearch.org/>

Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website where she/he is asked to enter a password or other confidential information. Most phishing emails aim at withdrawing money from financial institutions or getting access to private information. Phishing has increased enormously over the last years and is a serious threat to global security and economy. There are a number of possible countermeasures to phishing. These range from communication-oriented approaches like authentication protocols over blacklisting to content-based filtering approaches.

We argue that the first two approaches are currently not broadly implemented or exhibit deficits. Therefore content-based phishing filters are necessary and widely used to increase communication security. A number of features are extracted capturing the content and structural properties of the email. Subsequently a statistical classifier is trained using these features on a training set of emails labeled as ham (legitimate), spam or phishing. This classifier may then be applied to an email stream to estimate the classes of new incoming emails.

The primary goal of the EU AntiPhish project is to develop a prototype system that is highly accurate in the detection of phishing email messages. In the first phase of the AntiPhish project a machine learning prototype called AntiPhish Filter System (APS) was developed and evaluated on public benchmark data. It combines a number of novel features that are particularly well-suited to identify phishing emails. In particular we investigated Latent Dirichlet Allocation (LDA) topic models to capture words that frequently co-occur in email messages. We developed a special version called the latent Class-Topic Model (CLTOM), which is an extension of latent Dirichlet allocation (LDA) in such a way that it incorporates category information of emails during the model inference for topic extraction [BCP+08]. In addition we derived an optimized version of Dynamic Markov Chain (DMC) models, which generates far smaller models without sacrificing performance [BCP+08]. We combined these features with other standard and image features and trained a classifier using feature selection. In experiments our methods increase the f-value for classifying phishing emails on public benchmark data from 97.6% [FST07] to 99.5% [BDG+09].

Common tricks of spammers known as message salting are the inclusion of random strings and diverse combinatorial variations of spacing, word spelling, word order, etc. Some salting techniques called hidden salting cause messages to visually appear the same to the human eye, even if the machine-readable forms are very different. We rendered the email image to detect the appearance and overlap of characters. Using this evidence we developed specific

classifiers for identifying hidden salting of new types using outlier detection methods [BPR+08].

In the last phase of the project we applied the APS to real-world spam and phishing detection. In the first field experiment APS was applied to the real email stream at an Internet Service Provider. Starting with an initial labeled sample to estimate a starting classifier for ham vs. non-ham (spam + phishing) we applied active learning to select new emails for classification. These selected emails have to be labelled, which can be done by volunteer customers. The APS achieved good results as stand alone filter with 0.34% false positives (ham classified as non-ham) and 7.1% false negatives. The combination with a commercial spam filter (which evaluates blacklists) could improve the performance of both filters yielding 0.33% false positives and 5.4% false negatives.

The second field experiment was devoted to the analysis of known spam and phishing emails from a honey pot network. The task is to separate phishing emails from spam emails in a constant message stream. This is especially important as most phishing scams exist only a few hours. To update commercial phishing filters immediately it is important to capture phishing emails which are not covered by the current phishing signatures. From the emails new rules may be created and forwarded to customers.

In a sliding window approach (4 weeks training, one week prediction) the emails were monitored for 6 months. The results show that the APS has a good performance with low errors: 0.18 % Spam was classified as Phishing, 4.9% Phishing classified as Spam. More importantly APS detected a large number of emails which were not captured by current filtering rules. Hence APS permits prioritization of phishing-filter updating, which is most important because of the high damage caused by phishing emails..

References

- [BCP⁺08] A. Bergholz, J.-H. Chang, G. Paaß, F. Reichartz, and S. Strobel. Improved phishing detection using model-based features. In *Proc. Conference on Email an AntiSpam CEAS 2008*, 2008.
- [BDG⁺09] Andre Bergholz, Jan De Beer, Sebastian Glahn, Marie-Francine Moens, Gerhard Paass, and Siehyun Strobel. New filtering approaches for phishing email. *Journal of Computer Security*, accepted for publication, 2009.
- [BPR⁺08] A. Bergholz, G. Paaß, F. Reichartz, S. Strobel, M.-F. Moens, and B. Witten. Detecting known and new salting tricks in unwanted emails. In *Proc. Conference on Email an AntiSpam CEAS 2008*, 2008. Submitted for publication.
- [FST07] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 649–656, 2007.