# Poster: User-Centric Phishing Threat Detection

Lung-Hao Lee (Student), Kuei-Ching Lee (Student),
Yu-Yun Liu (Student), Hsin-Hsi Chen (Faculty)
Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{d01922005, p00922002, r99922102, hhchen}@ntu.edu.tw

Yuen-Hsien Tseng (Faculty)
Information Technology Center
National Taiwan Normal University
Taipei, Taiwan
samtseng@ntnu.edu.tw

***Abstract*—This paper presents a context-aware phishing threat detection model from users' behavioral perspectives. The context of users' information accesses is investigated to explore the users' browsing behaviors that confront phishing situations. Large-scale experiments show that our approach achieves an accuracy of 0.9973 and an F1 score of 0.9311 for predicting the phishing threats of users' next accesses without intelligent content analysis. Error analysis indicates that our proposed model results in a favorably low false positive rate of 0.0006. In practice, our proposed model is complementary to the existing anti-phishing techniques for cost-effectively blocking phishing threats with wisdom of the crowds.**

***Keywords—collaborative filtering*; *browsing behaviors*; *collective intelligence*; *context-aware category prediction***

## I. Introduction

Phishing crimes are significant security threats involving fraudulent web pages that masquerade as trustworthy ones for tricking users into revealing private and sensitive information, *e.g.*, bank account numbers, passwords, personal identification numbers, and credit card numbers. In the past, a content-based method has adopted Robust Hyperlinks for anti-phishing [1]. Content-based lexcial features has been extracted to detect phishing URLs using online learning [2]. A hybrid approach has been proposed to detect phishing web pages by identity discovery and keywords retrieval [3]. Several heuristics has been introduced to expand existing backlists that defend against phishing threats [4].

The previous approaches, which formulate the discriminative patterns from phishing web pages themselves, suffer from those security threats resulting from unknown phishing patterns. In contrast, we study the phishing context which users will fall into from users' points of view. We aim at exploiting collective intelligence embedded in users' browsing behaviors to detect phishing threats without the help of crawling web pages for intelligent content analysis.

## II. Behavioral Maximum Entropy Markov Model

Criminals usually create phishing web pages by exactly copying the legitimate ones or slightly modifying their page content for redirecting users' valuable information to the criminals rather than the legitimate sites. Users' browsing behaviors on the web result in users' click-through trails, which are defined as access sequences during web surfing. The browsing context of users' information accesses is explored to understand how users fall into phishing states. The problem statement in this study is described as follows. Let $u_1u_2\ldots u_{(n-1)}u_n$ be a user's access sequence, where $u_i$ is the $i^{\text{th}}$ clicked URL in the click-through trail. We focus on determining the category of a user's next access $u_n$, *i.e.*, phishing or legitimate, based on the previous accesses $u_1u_2\ldots u_{(n-1)}$ and their contextual information from users' behavioral perspectives.

Behavioral features of each clicked URL in a user's access sequence, which are extracted to capture contextual information, are classified into the following 3 types. (1) Hostname: phishing URLs tend to look like the original legitimate ones, so users are usually not conscious of them easily. For example, the hostname "faecbook.com" was verified as a phishing website of well-known social networking service Facebook. We identify the hostnames of clicked URLs as hints for phishing threat detection. (2) IP Address: phishing criminals usually create and maintain a large number of hosts or redirections to pretend legitimate URLs. These suspected URLs may be hosted in the same suspicious IP address. We also look up the referring IPs of clicked URLs as features. (3) Port Number: Secure Socket Layer (SSL) is a cryptographic protocol that provides communication security on the web. The port number is usually defined as 443 for accomplishing this secure connection. In addition, some content providers use specific ports to achieve their specific purposes. We also identify the port numbers of clicked URLs for anti-phishing.

We employ the Maximum Entropy Markov Model (MEMM) by learning users' browsing behaviors for predicting the category of a user's next access. A user's access is regarded as a state in our behavioral MEMM. Given an observation and its previous states, which are in terms of the above features, the probability of reaching a state is trained via maximum entropy. In testing phase, the proposed MEMM reports the category with the largest probability as the predicted result.

## III. Experiments And Performance

The data sets came from click-through data in the Trend Micro research laboratory. They consist of web browsing behaviors from 76,943 anonymous worldwide users. After manually checking the candidate categories proposed by analyzing content signature of each clicked URL, the category of a user's access is determined to provide secured web surfing. User click-through trails were divided into two distinct data sets shown as follows. (1) Training set: 99,249 clicked URLs from November 1st to December 31st 2010 were rated as phishing accesses. A phishing trail is denoted as $u_1u_2...u_{(n-1)}u_n$ where the previous accesses $u_1u_2...u_{(n-1)}$ are legitimate and the target URL $u_n$ is phishing. For balanced learning consideration, we selected the same number of legitimate trails $u_1u_2...u_{(m-1)}u_m$ in which all the accesses are legitimate and the hostname of $u_m$ has edit distance less than 4 with at least one phishing target, because the phishing URLs are usually similar to their legitimate URLs that want to pretend. A total of 198,498 users' access trails were used for training. (2) Test set: 134,432 phishing trails from January 1st to March 15th 2011 were used for testing. All of legitimate access trails from the same time period were used to reflect real-life users' browsing behaviors. In total, there are 6,496,860 legitimate trails.

The following two anti-phishing approaches based on click-through data were compared to demonstrate their performance. (1) Maximum Entropy: this model is a context-less method, which only focuses on the features extracted from the target URLs themselves. (2) Behavioral MEMM: this model is the proposed approach for context-aware phishing threat detection. Besides the target URL $u_n$, the previous accesses $u_1u_2...u_{(n-1)}$ is also considered.

Table 1 shows the experimental results. The performance difference between the two models was statistically significant ($p$<0.01), no matter which metric was adopted. The *Maximum Entropy* model slightly performed better than the *Behavioral MEMM* model when recall was concerned. This implies that considering the features selected from the target URL itself only has the effect on detecting the phishing accesses. The *Behavioral MEMM* model has better precision than the *Maximum Entropy* model. This reveals that exploring the collective intelligence embedded in browsing behaviors is effective on predicting the categories of users' next accesses. The proposed model accomplished the best accuracy of 0.9973 and F1 score of 0.9311. These results show that contextual information extracted from users' behavioral perspectives has strong impact on detecting phishing threats effectively.

TABLE I. PERFORMANCE EVALUATION ON PHISHING DETECTION

| **Models** | **Evaluation Metrics** | | | |
|---|---|---|---|---|
| | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1*** |
| **Maximum Entropy** | 0.9866 | 0.6099 | 0.9357 | 0.7385 |
| **Behavioral MEMM** | 0.9973 | 0.9681 | 0.8968 | 0.9311 |

Table 2 shows the confusion matrix of using the proposed *Behavioral MEMM* model for phishing threat detection. Experimental result indicated that our model maintained a favorably low false positive rate of 0.0006 (*i.e.*, 3967/(6492893 +3972)), which is the proportion of legitimate accesses that are incorrectly predicted as phishing. We found that most of false positive cases are related to some specific web sites, *e.g.*, the error cases containing the hostname "ads.web.aol.com". This can be solved with except-lists, which contain legitimate hostnames to avoid being incorrectly predicted. Phishing URLs which were not correctly detected result in false negative cases. We found that some of these cases only exist in our test set. This implies that collecting users' access sequences as many as possible is needed for reflecting real-life users' browsing behaviors during web surfing.

We also analyze the data sets to understand the major categories of previous accesses that will result in phishing threats. Empirical findings indicate many users visiting web pages rating as the "Economy," "Shopping," or "Auction" category, which are all involved in personally financial payments or investments. It confirms the guideline that users should be more careful to visit such kinds of web pages to have more secured web surfing.

TABLE II. CONFUSION MATRIX USING BEHAVIORAL MEMM.

| **Behavioral MEMM** | | **Predicted Results** | |
|---|---|---|---|
| | | ***Positive*** | ***Negative*** |
| **Ground Trugh** | ***Positive*** | 120,552 | 13,880 |
| | ***Negative*** | 3,967 | 6,492,893 |

## IV. CONCLUSIONS AND FUTURE WORK

This paper proposes a user-centric model that exploits users' browsing behaviors only for context-aware phishing threat detection. Experimental results show that our behavioral MEMM model, which explores browsing contexts of users' previous accesses, yield favorable results in the large-scale experiments. In practice, our cost-effective approach is a lightweight process compared to the existing content-based analysis for blocking phishing threats.

This work is our first exploration to adopt URL information alone for anti-phishing. More discriminative features from users' behavioral perspectives will be investigated in the future to further improve real-time filtering performance.

## ACKNOWLEDGMENT

This research was partially supported by National Science Council, Taiwan under grant NSC99-2221-E-002-167-MY3, and the "Aim for the Top University Project" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan. We are also grateful to Trend Micro research laboratory for the support of click-through data.

## REFERENCES

[1] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A content-based approach to detecting phishing web sites," In *Proceedings of the 16th International World Wide Web Conference*, pp. 639-648, 2007.

[2] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing URL detection using online learning," In *Proceedings of the 3rd CCS Workshop on Artifical Intelligence and Security*, pp.54-60, 2010.

[3] G. Xiang, and J. Hong, "A hybrid phish detection approach by identity discovery and keyword retrieval," In *Proceedings of the 18th International World Wide Web Conference*, pp. 571-580, 2009.

[4] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: predictive blacklisting to detect phishing attacks," In *Proceedings of the 29th IEEE Conference on Computer Communications*, 2010.