Countering Advanced Phishing

Phishing attacks form the primary vector used by malicious hackers to penetrate corporate and government network – in fact, according to Deloitte, 91% of cyberattacks are launched through phishing. The initial waves of phishing attacks were launched through emails, attachments with enticing names, and websites designed to lure the unwary.

In recent years, social engineering by malicious hackers have reached new levels. These include methods to phish through citizens’ Twitter, Facebook, LinkedIn and other social media feeds. They include methods to set up fake accounts on such platforms and connect with users in an attempt to target both the user and his/her friends.

Researchers at the Northwestern Security & AI Lab have developed sophisticated methods to detect and counter some types of advanced phishing methods. We developed MALT\textsuperscript{p}, a machine learning based method to detect tweets which lead users to phishing URLs. More recently, in collaboration with the University of Liechtenstein, we developed the largest phishing dataset, LNU-Phish which we will be releasing in 2022. We also developed a framework called POC to robustly detect a wide variety of phishing attacks even when the adversary attempts to evade the phishing detector.

MALT\textsuperscript{p}: Malicious Tweets In Parallel

Embedded URLs and multimodal content (images, video, and sound recordings) in tweets are increasingly used to seduce users into a “wrong click”, leading to malware infection. We have developed methods to predict whether a tweet is malicious or not by examining five classes of features: (i) textual content including sentiment, (ii) paths emanating from a URL mentioned in the tweet, (iii) attributes associated with URLs, and (iv) multimodal content in the tweet. A fifth class of features first constructs a novel “tweet graph” and then defines features by analyzing “metapaths” contained in the tweet graph.

Our MALT\textsuperscript{p} classification algorithm merges together tweet graphs, metapaths, and collective classification proposed previously in the literature. We conduct detailed experiments using two data sets — the Warningbird (WB) dataset and another dataset called KBA.

We show that our metapath-based approach outperforms past efforts at identifying malicious tweets and further show that metapath-based features in conjunction with Alexa ranks and features from KBA yield very high predictive accuracy — AUCs over 0.98 on KBA and over 0.94 on KBA, outperforming past work.

More significantly, metapath features alone generate a predictive accuracy (AUC) of 0.977 and 0.923, respectively, on the KBA and WB data sets, significantly outperforming the other methods in isolation. We performed a further analysis to identify the most important features. Surprisingly, our results show that the presence of multimodal content is not a major factor and that metapath-based features dominate in separating malicious from benign tweets.

**The LNU (Lichtenstein-Northwestern University) Phishing dataset Is the largest dataset appropriate for reproducible research on phishing URLs.**

Although machine learning based algorithms have been extensively used for detecting phishing websites, there has been relatively little work on how adversaries may attack such “phishing detectors” (PDs for short).

In this joint research with the University of Liechtenstein, we propose a set of Gray-Box attacks on PDs. In Gray box attacks, the adversary neither knows everything about the defenses (white box) or nothing (black box). Instead, the adversary’s knowledge of the defenses (i.e. phishing detectors used and their parameters) may vary considerably. We show that our proposed Gray-Box attacks severely degrade the effectiveness of several existing PDs.

We propose the concept of operation chains that iteratively map an original set of features to a new set of features and develop the “Protective Operation Chain” (POC for short) algorithm. POC leverages the combination of random feature selection and feature mappings in order to increase the attacker’s uncertainty about the target PD. Using 3 existing publicly available datasets plus a fourth that we have created and will release in late 2022, we show that POC is more robust to these attacks than past competing work, while preserving predictive performance when no adversarial attacks are present. Moreover, POC is robust to attacks on 13 different classifiers, not just one. These results are shown to be statistically significant at the p < 0.001 level.
The LNU-Phish Dataset

Past publicly available datasets about phishing URLs had several problems. First, many of the URLs referenced are dead and hence cannot be studied any more. Second, some past datasets only contain feature vectors and don’t mention the actual webpages and this means that new features can no longer be extracted from those URLs. Third, different datasets offer different sets of features and code is not often available to compute the same set of features across all datasets. Fourth, public datasets such as PhishTank are continuously updated, making replication a challenge as a researcher may not know what a previous researcher used. All of this means that it is very challenging to replicate and improve upon past work in a single, uniform manner.

The LNU-Phish dataset, to be released in late 2022, contains over 23,000 samples. LNU-Phish is a large and fixed dataset that can be used for future work on PDs. It contains complete information on each sample, such as the URL, the DNS records, as well as the underlying HTML code and a screenshot of the webpage. In addition, the dataset includes most important features presented in prior work, as well as several new features. Hence, even if a URL has been taken down, the data in LNU-Phish captures all information needed for reproducibility by future researchers; such information can also be augmented. LNU-Phish also provides code capable of working with the data.

Additional Information

References