Visual Similarity based Zero-Day Phishing Websites detection

Olaru Gabriel Iulian, University Politehnica of Bucharest

***Abstract*— Phishing attacks have been rising concerns more and more often in the past years. According to IC3, in 2019 phishing attacks affected 241,342 victims. In 2020, 6.95 million new phishing and scam pages were detected, and the attacks targeted 251,342 people. The numbers just kept rising in 2021, with 260.642 victims. Not only this has prompted the users to be more careful about their online activity, but also the experts to try to come up with preemptive detection solution to such phishing pages. As a response to the rising threat, a team of such experts from the CISPA Helmholtz Center for Information Security has produced a tool called *VisualPhishNet* that will increase the detection rate of malicious pages. This paper aims to explore the team's solution as well as present other available options when it comes to protect against phishing.**

I. Introduction

In order to understand the importance of detection algorithms such as the one developed by "CISPA Helmholtz Center for Information Security" team Sahar Abdelnabi, Katharina Krombholz and Mario Fritz have developed, one must understand first understand that the main threat that phishing attacks rise is the theft of the user's confidential data. This exposes the victim to a range of risks varying from financial threats to identity theft. This are very high risk threats that can deviate the whole trajectory of someone's life. With the increasing in the occurrences of such attacks, facilitated by the rising number of phishing kits (sets of tools that allow bad actors to replicate a legitimate website's appearance with ease), counter measures such as preemptive protection are becoming more and more necessary.

There are two ways of protecting against phishing: the use of a browser and the use of a toolbar. The first is by installing a browser plug-in, while the second is by clicking on the "Protect" button. Aside from being able to prevent phishing, it also has to be considered how effective the various features of a browser can be in preventing users from visiting fraudulent website.

II. Protection Strategies

As backed by a review article published by Ankit Kumar Jain and B. B. Gupta from the National Institute of Technology, Kurukshetra, there are many approaches to implementing tools meant to defend against online phishing.

In the current ecosystem, some solutions that protect the users against scam websites are based on blacklisting sites that have been reported as malicious. This approach has the drawback tat is reactive instead of proactive, leaving users prone to zero-days attacks. The bad actors' websites must first be detected by the maintainer of the software. This works as a basic layer of security, but leaves the user exposed to the novel variety of phishing scams than appear every day.

Another popular approach is to employ the use of heuristics in order to detect the fraudulent websites. This approach, however, struggles to keep up with the rising number of fake sites, since each heuristic needs to be updated for every new approach to creating a fake website.

There are also content-based phishing detection solutions, which compute the lexical signature of a site. They then look this up. The odds are high that the original site will appear in the query. One such solution is designed by a team of researchers composed of Eric Medvet, from the University of Trieste, Engin Kirds, from Eurecom and Christopher Kruegel, from the university of California. They define the signature of a site as a collection of data about the text and visuals of a page, extracted from each node of the HTML DOM tree. Not only that they take into account the layout of the page when constructing the signature, but also the source, color histogram size and compression rate of each image. They then compare the signature one against another by grouping the elements of a website based on types and comparing each element with each other.

Although this particular tools is able to achieve 0.% false positive rate and a 7.4% false negative rate, the drawback of these approaches are that they are based on a web search engine. Web pages that the engine did not register in the database before the search can appear in the search as well, making this method of protection also vulnerable to the rapid pace of the developing of fake websites.

III. Related Work

The authors from the Helmholtz Institute have contributed with a solution that relies on the construction of a threat model using data about malicious websites gathered form trusted sources. Based on this model, they trained an AI solution based on a triplet Convolutional Neural Network that promises to deliver great results when it comes to detecting new phishing sites. It this fast growing ecosystem, an approach based on preemptive actions, instead of reactions, might just provide the insurance that the user need in order to trust a website with their personal and confidential data. The researchers did not stop here tough, and also gathered a data set called *VisualPhish*, that is now publicly available for other models to train on.

IV. The phishing data set

After conducting some research on the existing phishing data sets, the authors have concluded that they lack a lot in size. Another limitation is most of the phishing detection data sets are not available for public consumption. However, DeltaPhish is one of the few that can provide a good overview of how phishing pages are distributed on the web, although the validation methods used in this data set only apply to the websites that are hosting the phishing pages; not to mention, the screenshots are duplicates. In order to overcome this shortcomings, the researchers have increased the size of a the trusted list, reduced the number of duplicates, added pages that targeted the whole website as well as new legitimate pages and phishing pages and finally have gathered a sample sized that limits bias. Also, since the most well-known web pages provide the attackers with be best results, they are motivated to copy those. This is way the data was gathered mostly from the most popular sites. The result was a well balanced test set that could be used to solve the similarity learning problem that the team was facing.

V. How it works

The neural network is trained in two stages: first, with a random list from the trusted sites screenshots, second, by looking for out of the ordinary examples, going from checkpoint to checkpoint. The model also does image, signature and character recognition, combining most of the strategies described above. The main analysis is performed on an image that serves as an anchor, one that has a similar identity and one with a different identity, respecting the "Triple Networks" model. The data is also sampled in three stages, first uniformly, then fine tuned. This allows the model to determine the identity of a website and the match it against the database, based on the distance between the phishing site and another. The success is measured by the deviance between the identities of websites and the percentage of overall matches of a phishing site.

VI. Results

The tool manages to pull 81.03% accuracy when detecting malicious pages in a set of 155 websites, with a 93% true positive rate and only a 4% false positive rate.

The only drawback of this solution is the high deployment cost, given by the large data set and the high computational power required for training the model and for making prediction for a large number of sites. The domain names also present an issue of maintenance, since they need to be kept updated in order to maintain the accuracy of the model.

VII. Similar solutions

Shuichiro Haruta, Hiromu Ashina and Iwao Sasase from the Keio University propose another AI based solution that analyses the CSS properties of a page. The model is trained of the 500 most popular sites from the Alexa dataset. This works by maintaining a white list of HTML and CSS templates and computing the identity of a page based on this, making it prone to errors when trying to detect sites with slightly different layouts, or with vastly different code bases that result in the same layout.

Masanori Hara, Akira Yamada and Yutaka Miyake also propose a solution that allows detection of malicious sites without any information about the victim page. They pull this off only by comparing URL and the images hosted on a page and then classifying a site as undetermined, fake or legitimate, delivering an 80% detection rate, although the number of false positive is high for a small enough data set. The initial database that they rely on is a set of screenshots from an assortment of legitimate websites.

The last solution covered is proposed by Jiann-Liang Chen, Yi-Wei Ma and Kuan-Lung Huang. Their system consists of three big components: a screenshot matcher, split into a contur matcher and a color matcher, a logo finder and a cache of white and black lists. The screenshot first goes trough the color and contour matchers then trough the logo detection, resulting in a prediction after a cache lookup. Since their approach is solely based on image recognition, though, is prone to all the failures described above. Although it manages to score impressive performances (97% accuracy) it is only for a very selective and small (just a handful) sites.

VIII. Conclusion

In conclusion, this paper presented a variety of solutions when it comes to protecting against phishing, focusing on the importance of preemptive approaches. The spotlight is taken by the *VisualPhishNet* tool and its creators, who after a thorough analysis of the existing solutions and data sets have managed to come up with an innovative solution that give users reassurance when they are browsing websites. Not only this, but the creation of the large *VisualPhish* test set will facilitate further development in the area of threat response and protection for future developers and researchers.

References

[1] https://arxiv.org/abs/1909.00300
[2] https://downloads.hindawi.com/journals/scn/2017/5421046.pdf
[3] https://www.hindawi.com/journals/scn/2017/5421046/
[4] https://ieeexplore.ieee.org/document/4925087
[5] https://www.researchgate.net/publication/228906286_Visual-Similarity-Based_Phishing_Detection
[6] https://www.mdpi.com/2073-8994/12/10/1681/htm
[7] https://www.researchgate.net/publication/350371788_Detecting_Phishing_Sites_--_An_Overview
[8] https://www.tessian.com/blog/phishing-statistics-2020/
[9] https://www.csoonline.com/article/3634869/top-cybersecurity-statistics-trends-and-facts.html#:~:text=In__2020%2C__the__key drivers,in one month of 206%2C310.