Phishing Classification using Lexical and Statistical Frequencies of URLs

Sergio Villegas, Alejandro Correa Bahnsen, and Javier Vargas
Easy Solutions Research
{svillegas,acorrea,jvargas}@easysol.net

Abstract. Phishing attacks have been a growing problem worldwide. According to the Anti-Phishing Working Group, during 2014 the number of unique phishing sites in the world reached an all time high of 247,713 [1]. Phishing, by definition, is the act of defrauding an online user in order to obtain personal information by posing as a trustworthy institution or entity [2]. Users usually have a hard time differentiating between legitimate and malicious sites because they are made to look exactly the same [3]. Therefore, there is a need to create better tools to combat attackers.

In this study, we combine statistical analysis of a URL and a Random Forest classifier to accurately classify phishing websites based only on the URL. We used a sample of 1.2 million phishing URLs extracted from Phishtank and 1.2 million ham URLs from the CommonCrawl corpus to train the model. Classification based on URLs facilitates a defense against all phishing attacks due to the feature they all share, a URL. We estimate a total of 35 features based on analysing the structure of the URL, for example by estimating Kullback-Leibler Divergence between the normalized character frequency of the English language and the URL [4]. Other features include the character frequencies, the number of @ and - symbols, the number of top-level domains in the URL, whether the URL is an IP address, the length and the number of suspicious words in the URL.

Our results confirm that a simple defense vector as this has shown great technical results due to its simplicity and excellent statistical measures of performance. The resulting model had a F_1-Score of 0.94, an Accuracy of over 95% and showed great stability in the holdout set.

References

1.APWG, "Global Phishing Survey : Trends and Domain Name Use in 2H2014," Tech. Rep. May, 2015.
2.S. Roopak and T. Thomas, "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity," in *2014 Fourth International Conference on Advances in Computing and Communications*, 2014, pp. 167–170.
3.R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," in *SIGCHI Conference on Human Factors in Computing Systems*, 2006, pp. 581–590.
4.R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," in *ACM Conference on Data and Application Security and Privacy*, 2015, pp. 111–121.