Technische Universität Berlin

Exposé

model2regex: Detecting DGAs with Regular Expressions Generated by a Language Model

Eric Schneider
Matr. No. 365800

Chair of Machine Learning and Security
Prof. Dr. Konrad Rieck

supervised by
Alexander Warnecke, Tammo Krüger

April 1, 2024

1 Introduction

Domain Generating Algorithms (DGAs) are increasingly used in botnets as part of command and control (C&C) communication. Malware creators use these algorithms to generate multiple possible domains each day and then have their malware contact a small portion of them to obfuscate the real server they are getting their instructions and updates from. This tactic gives the attacker a huge advantage, because to protect against it, means taking control of possible thousands of domains, while the attacker only needs to control a short lived domain executing their attack. Better protection may come from blocking botnets at the source by recognizing communication with specific domains as fraudulent or rather generated by a specific DGA family. DGAs however are generated randomly and use different seeds from either specific dates, twitter trends, hashes or word lists. Therefore static blocklists may not be able to keep up with blocking the communication at a network level. Deep Learning approaches have shown great promises and are currently state of the art in detecting Algorithmically-Generated Domains (AGD). Machine learned models can be used to filter network traffic, but there are not simple setups to put them into the filter pipeline, also is it difficult to know what the model exactly is filtering. This thesis will therefore trying to attempt to use the help of deep learning approaches to learn the structure of domains generated by a DGA and use this information to then generate a regular expression (RegEx) which will match possible domains of that DGA. That approach could potential solve the two problems of understanding what the structure the model sees and make the setup for filters much easier by providing a standard way to set up a filter with a RegEx.

2 Methodology

The main methodology of this thesis will be applied research. Using currently established solutions from the field of language processing. As data source of the learning process is a mix of real domains [2] from domain lists and self-generated domains using reverse engineered DGAs. [1] The main focus of the research will be answering the two main questions. First, is it possible to generate Regular Expressions in this setting? Second, do these regular expressions perform as well as the current state of the art?

3 Approach

I plan to split this thesis into three phases. First a research phase to expand on what I explored during the development of the first prototype. I need to explore what methods for regular expression generation can be used to extract them out of the RNN learned on the dataset. Also it should be part of the research to figure out how to best simplify or make the resulting regular expressions more efficient and capture the structure learned by the model. Currently the prototype is doing rather well on the simple banjori algorithm during training. Once the possible approaches have been explored I will need to implement them and do initial testing on how well they perform. This may be the biggest phase of the thesis work. After implementing the model and training it, I will need to evaluate its performance and compare it to the state of the art.

4 Evaluation

I will evaluate the result of this thesis by testing how well the generated regular expressions will capture the learned structures of DGAs. I will compare how well these expressions will detect AGDs. It is imperative that however the amount of false positives (benign domains detected as AGDs) should be very low to not block valid connections with the generated filter. The evaluation should compare how well the generated regular expression performs compared to the language model and also how well it compares to other deep learning approaches shown in other scientific papers. It should also be evaluated how well the language model and the regular expression performs on different kinds of DGAs, or if it is only able to detect specific kinds. The degree of success should be measured through how close the solution is performing compared to the state of the art [5] and well established solutions in detection scores and false positive rates.

5 Scope

The main scope of this work is determining the possibility of using the generated regular expressions for filtering and how well it works compared to trained models and the state of the art in the field for Detecting DGAs. Part of the necessary work is training a multi-task criterion for the language model and the classification of DGA and benign domains. If possible the resulting regular expressions should be simplified to make them more readable and efficient, however this is not a necessary requirement.

6 Related Work

The current state of the art in this field are hybrid solutions using convolutional and recurrent neural networks. Two promising approaches were developed by Zhang [6] and Liu [3], both utilizing SMOTE for balancing the datasets and both using the hybrid solution of a bidirectional LSTM and a CNN. These approaches showed high accuracy and f1-scores when detecting multiple DGA families at the same time, and experimented with different text extraction techniques. Rayhan [4] also did a experimental analysis of 13 state-of-the-art classification techniques.

References

[1] Johannes Bader. *baderj/domain_generation_algorithms*. original-date: 2015-08-31T09:14:32Z. Mar. 2024. URL: https://github.com/baderj/domain_generation_algorithms (visited on 03/28/2024).

[2] *Cisco Popularity List*. URL: https://s3-us-west-1.amazonaws.com/umbrella-static/index.html (visited on 04/01/2024).

[3] Zhanghui Liu et al. “Detection of Algorithmically Generated Domain Names Using the Recurrent Convolutional Neural Network with Spatial Pyramid Pooling”. en. In: *Entropy* 22.9 (Sept. 2020). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 1058. ISSN: 1099-4300. DOI: 10.3390/e22091058. URL: https://www.mdpi.com/1099-4300/22/9/1058 (visited on 04/01/2024).

[4] Md Maruf Rayhan and Md. Ahsan Ayub. “An Experimental Analysis of Classification Techniques for Domain Generating Algorithms (DGA) based Malicious Domains Detection”. en. In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. DHAKA, Bangladesh: IEEE, Dec. 2020, pp. 1–5. ISBN: 978-1-66542-244-4. DOI: 10.1109/ICCIT51783.2020.9392701. URL: https://ieeexplore.ieee.org/document/9392701/ (visited on 04/01/2024).

[5] Bin Yu et al. “Character Level based Detection of DGA Domain Names”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. ISSN: 2161-4407. July 2018, pp. 1–8. DOI: 10.1109/IJCNN.2018.8489147. URL: https://ieeexplore.ieee.org/abstract/document/8489147 (visited on 03/30/2024).

[6] Yudong Zhang et al. “Detection of Algorithmically Generated Domain Names Using SMOTE and Hybrid Neural Network”. en. In: *Computer Supported Cooperative Work and Social Computing*. Ed. by Yuqing Sun et al. Singapore: Springer, 2019, pp. 738–751. ISBN: 9789811513770. DOI: 10.1007/978-981-15-1377-0_57.