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 24, 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, setting them up for a filter pipeline may not be very simple and also can be a black box on determining what exactly the model is filtering. Using algorithms from the field of language processing this thesis will attempt to learn the structure of different DGA families trying to learn their structure and using that information to generate regular expressions (RegEx). Resulting in an ease in implementing filters in existing security architecture and showing a more human readable result to help understanding the structure of learned DGAs.

2 Methodology

The main methodology of this thesis will be applied research. Using currently established solutions from the field of language processing, I will train a language model that learns the structure of DGAs and classifies them into their DGA families or into benign domain names. Once the struc-

ture has been learned the research will focus on if it is possible to extract information from the language model and turn it into a RegEx. If this is a success then the work will focus on evaluating if the generated RegEx can compete with the language model itself and solutions of the current state of the art. Again in an iterative approach of researching and testing strategies for extracting and optimizing the resulting RegEx.

3 Approach

As mentioned before I will iteratively explore the possibilities of turning a learned model into a regular expression, so the beginning of the research will be working around exploring these possibilities and iterating on implementations. Currently a small prototype exists that can learn the banjori algorithm very reliably. The language model will have the following architecture:

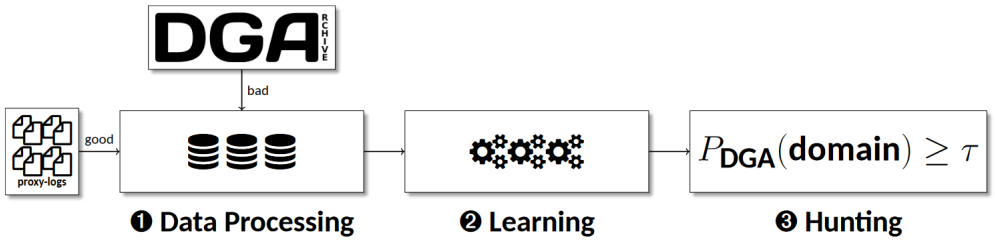

Figure 1: Graphic provided by Tammo Krüger

The learning part is already solved and does work with a decent accuracy, the last step (3) Hunting is now what the main focus of the thesis will be. The majority of these 6 months I will work on this part and at least 2 months will be used for setting up the evaluation described and on writing down my results into the thesis.

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 and well established solutions in detection scores and false positive rates. Specifically I will compare the language model, the state of the art and the regular expressions on the usual metrics used in other papers in the machine learning field. I will compare the true and false positive rate (TPR/FPR), the accuracy and the F-Score. Added to that I will compare the ROC curve. The experiment will be set up the following way:
Using our training dataset all models, my own and the ones from the following paper [2] will be trained. Both the language model and the character level detection models from Yu et al. are going to be tested on a test dataset, if possible that test dataset will be provided from a real world data collection. Once these tests are through the performance will be compared directly.

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 lan-

guage 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 the field are Convolutional Neural Networks and Recurrent Neural Networks which are also commonly used in Natural Language Processing. Yu et al. [2] showed that all these approaches achieved similar accuracy and performance in detecting DGAs. The paper used the Bambenek dataset for AGDs and the Alexa Dataset for benign domains. The models were two pure RNN based Architectures (Endgame, CMU), two CNN based architectures (NYU, Invincea) and one hybrid CNN/RNN based architecture (MIT). In their own study Rayhan et al. [1] also tested 13 different machine learning algorithms to compare how they perform in the setting of detecting the dual class problem of benign or malicious domains. The test was performed on uni-, bi- and trigrams of the domains sourced from the DGArchive, the paper shows that even some simpler approaches like K-Nearest Neighbor, or Support Vector Machines were able to get decent results.

References

[1] 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).

[2] 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).