Mini project Report on

A Machine Learning Framework for Domain Generation Algorithm (DGA)-Based Malware Detection

Submitted by

Mohit Kamble: 632

Akshay Kalapgar: 631

Harsh Dobariya: 614

Guided by

Prof. A.E PATIL

MCT

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI

(Permanently Affiliated to University of Mumbai)

Juhu Versova Link Road, Andheri (West), Mumbai-53

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIVERSITY OF MUMBAI 2019-2020

CERTIFICATE

Date:__________

This is to certify that,theminiprojectworkembodiedinthisreportentitled,“**A Machine Learning Framework for Domain Generation Algorithm (DGA)-Based Malware Detection**” submitted by “***Mohit Kamble*** bearing Roll No. 632” , “***Akshay Kalapgar*** bearing Roll No. 631”, “***Harsh Dobariya*** bearing Roll No. 614” for the award of ***Third year in Bachelor Of Engineering (T.E.)*** degree in the subject of ***Information Technology***, is a work carried out by them under my guidance and supervision within the institute. The work described in this mini projectreportiscarriedoutbytheconcernedstudentsandhasnotbeensubmittedfortheawardof any other degree of the University of Mumbai.

Further, it is certify that the students were regular during the academic year 2019-2020 andhaveworkedundertheguidanceofconcernedfacultyuntilthesubmissionofthisminiproject work at ***Rajiv Gandhi Institute of Technology, Mumbai***.

Prof. A.E Patil

Mini Project Guide

Dr. SunilB.Wankhade

Dr. Sanjay U.Bokade

HeadofDepartment **Principal**

CERTIFICATE OF APPROVAL

This mini project report entitled

Submitted by:

MOHIT KAMBLE	**632**
AKSHAY KALAPGAR	**631**
HARSH DOBARIYA	**614**

In partial fulfilment of the requirements of the degree of **Third year in Bachelor of Engineering** in **Information Technology** is approved.

Internal Examiner

____

SEAL OF

INSTITUTE

External Examiner

____

ABSTRACT

Attackers usually use a Command and Control (C2) server to manipulate the communication. In order to perform an attack, threat actors often employ a Domain Generation Algorithm (DGA), which can allow malware to communicate with C2 by generating a variety of network locations. Traditional malware control methods, such as blacklisting, are insufficient to handle DGA threats. In this paper, we propose a machine learning framework for identifying and detecting DGA domains to alleviate the threat. We collect real-time threat data from the real-life traffic over a one-year period. We also propose a deep learning model to classify a large number of DGA domains. The proposed machine learning framework consists of a twolevel model and a prediction model. In the two-level model, we first classify the DGA domains apart from normal domains and then use the clustering method to identify the algorithms that generate those DGA domains. In the prediction model, a time-series model is constructed to predict incoming domain features based on the Hidden Markov Model (HMM). Furthermore, we build a Deep Neural Network (DNN) model to enhance the proposed machine learning framework by handling the huge dataset we gradually collected. Our extensive experimental results demonstrate the accuracy of the proposed framework and the DNN model. To be precise, we achieve an accuracy of 95.89% for the classification in the framework and 97.79% in the DNN model, 92.45% for the second-level clustering, and 95.21% for the HMM prediction in the framework.

Table of Contents

CHAPTER 1

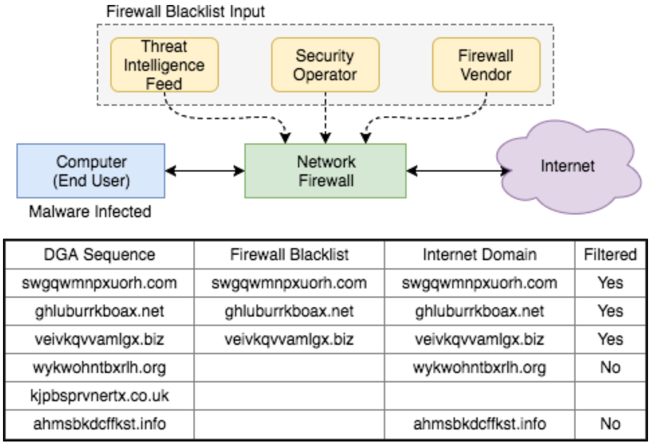
INTRODUCTION

Malware attackers attempt to infiltrate layers of protection and defensive solutions, resulting in threats on a computer network and its assets. Anti-malware software have beenwidelyusedinenterprisesforalongtimesincetheycan providesomelevelofsecurityoncomputernetworksandsystems to detect and mitigate malware attacks. However, many anti-malwaresolutionstypicallyutilizestaticstringmatching approaches, hashing schemes, or network communication whitelisting. These solutions are too simple to resolve sophisticatemalwareattacks,whichcanhidecommunication channels to bypass most detection schemes by purposely integrating evasive techniques. The issue has posed a serious threat to the security of an enterprise and it is also a grand challenge that needs to beaddressed.Some of the sophisticate malware attackers use either a static or dynamic method to communicate with a centraized server to service a Command and Control (C2). In a static method, everything is fixed. For example, the malware has both a fixed IP address and a fixed domain name permanently (i.e., its domain name will not change throughout its lifespan). Thus, as long this malware hasbeen identified as a threat, a simple rule can be applied to resolve this malware threat issue. In a dynamic method, Domain Generation Algorithm (DGA) has been commonly used to communicate back to a variety of servers. The DGA is a sequencing algorithm that is used to periodically generate a largenumberofdomainnames,whichareoftenusedbymal- waretoevadedomain-basedfirewallcontrols.Thegenerated domain names give malicious actors the opportunity tohidetheir C2 servers so that it is hard for the enterprise to identify the DGA.

EXISTING SYSTEM:

Threat models: Multiple conditions for a DGA to function in a network environment where filtering results in a firewall that protects the communication and an empty cell in an Internet domain that results in NXDOMAIN error. Each HMM date record represents a series of domain observations. First a sequence of domain name are processed by a feature extractor and each of these feature vectors is used as a training record. Then, similar sequences are clustered as a group of DGA domain names with certain outcomes. After the training process, if a sequence does not have an HMM sequence representation (or it is not presented in the training data but the test data), the HMM model then generates the future predicted results. Otherwise, we will use an existing HMM sequence presentation. Figure 5 (cp shows the HMM prediction workflow. Once the model has been trained, a set of features is formed by a series of DGAlayer and each connected line has a weight. DNNs can model complex non-linear relationships by using activation functions. The optimization algorithms are used i DNNs to reduce loss. To process large dataset. we build a deep learning model to classify the DGA domains and normal domains and compare our deep learning model with our machine learning method. Our deep learning model.

PROBLEM STATEMENT:



DGA Sequence	Firewall Blacklist	Internet Domain	Filtered
swgqwmnpxuorh.com	swgqwmnpxuorh.com	swgqwmnpxuorh.com	Yes
ghluburrkboax.net	ghluburrkboax.net	ghluburrkboax.net	Yes
veivkqvvamlgx.biz	veivkqvvamlgx.biz	veivkqvvamlgx.biz	Yes
wykwohntbxrlh.org		wykwohntbxrlh.org	No
kjpbsprvnertx.co.uk			
ahmsbkdcffkst.info		ahmsbkdcffkst.info	No

Threatmodels:MultipleconditionsforaDGAto function in a network environment where filtering results in a firewall that protects the communication and an empty cell in an Internet domain that results in an NXDOMAIN error. Note that the domains listed in the figure belong to existing live threats.Firewall blacklisting constantly expands as the multiple sources of inputs expand filtering rules. However, sequences inaDGAmaynotbeknowntotheseinputspromptly.

Our research problem is to accurately identify and cluster domains that originate from known DGA-based techniques where we target to develop a security approach that autonomously mitigates network communications to unknown threats in a sequence.

CHAPTER 2

AIM: To solve the problem of detecting DGA sequences using machine learning techniques derived from observations in a network.

OBJECTIVES

- In DBSCAN algorithm, we use the features described above to calculate the domain distance and to group the domains that are generated by the same DGA together according to their domain feature difference.
- Distinguish the model from training and prediction stages.
- The nodes in each layer are fully connected to the nodes in the next will not miss any local minima, but it will take a long time to converge.

CHAPTER 3

LITERATURE SURVEY

ADVANTAGES:

- In the second-level clustering we apply the DBSCAN algorithm. Only the DGA domains obtained from the first-level classification will be used for clustering.
- In DBSCAN algorithm, we use the features described above to calculate the domain distance and to group the domains that are generated by the same DGA together according to their domain feature difference.
- Distinguish the model from training and prediction stages.
- The nodes in each layer are fully connected to the nodes in the next will not miss any local minima, but it will take a long time to converge.

DISADVANTAGES:

- Our research problem is to accurately identify and cluster domains that originate from known DGA-based techniques where we target to develop a security approach that autonomously mitigates network communications to unknown threats in a sequence.
- Our proposed machine learning framework aims to solve the problem of detecting DGA sequences using machine learning techniques derived from observations in a network.
- The rest of the paper is organized as follows. Section II gives the problem statement. Section III discusses related work. Section IV presents data collection. the proposed machine learning framework. and the deep learning model. Then, Section V discusses the evaluation of the machine and that such features are built over our observations.
- The main reason behind our implementation is that many threat intelligence feeds and heuristic data often provide signatures to malware that has plagued a network or public Internet.

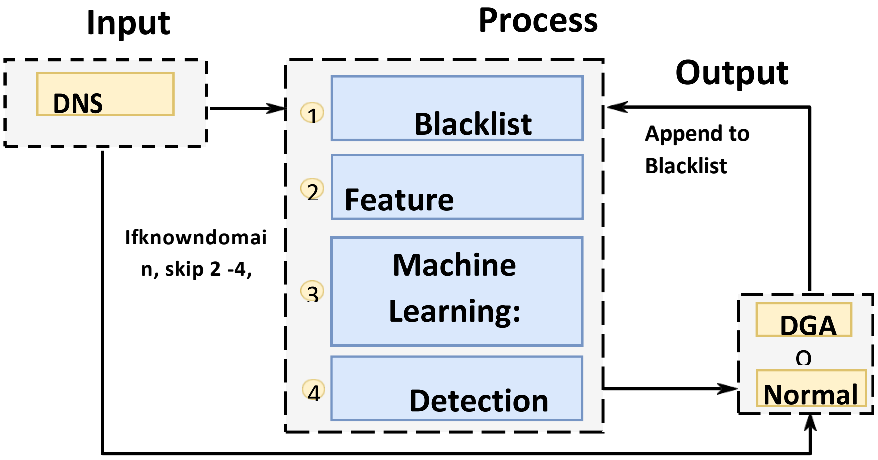
CHAPTER 4

PROPOSED SYSTEM

The important components in this research are: (1) domains extracted from DGAs; (2) a machine learning framework that encompasses multiple feature extraction techniques and the models to classify the DGA domains from normal domains, cluster the DGA domains, and predict a DGA domain. (3) a deep learning model to handle large datasets. Multiple on- line sources from simple Google searching provide example codes for a DGA construction. However, a majority of these techniques are trivial and fundamental at best. Online threat intelligence feeds give an approach to examining current and live threats in real-world environment.

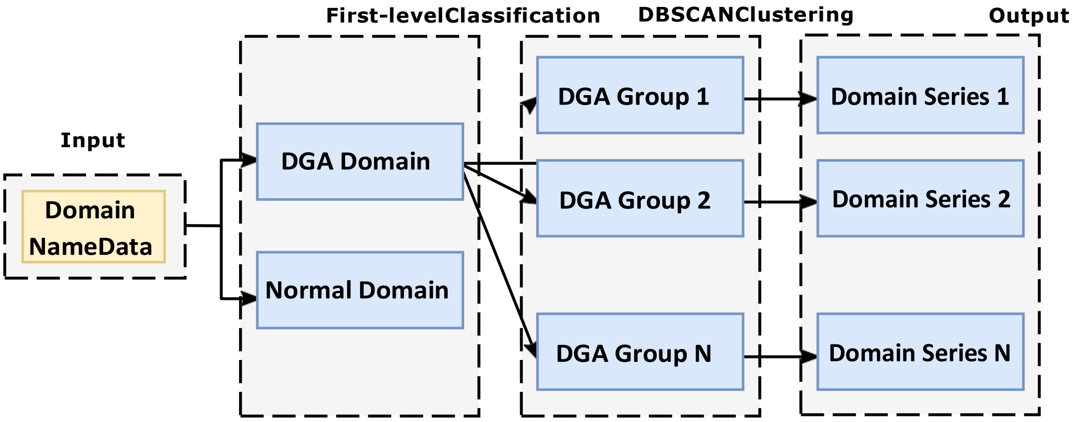
4.1 THREAT INTELLIGENCE FEED AND ONGOING THREAT DATA

DGAs are plentiful through multiple online examples that are found from Google searching and Github repositories. However, sophisticated threat actors purposely create tailored DGA to evaluate current detection systems. Using real-time active malicious domains derived from DGAs on the public Internet measures the accuracy of the proposed approach. Specifically, threat intelligence feeds collected from Bambenek Consulting over a period of one year were obtained through daily manual querying demonstrated trends of ongoing threats. The structure of the data is presented in a CSV format of domain names, originating malware, and DGA membership with the daily file size of approximate 110MB, we have collected 64GB in total. Figure 2 demonstrates an example feed from the collected data.

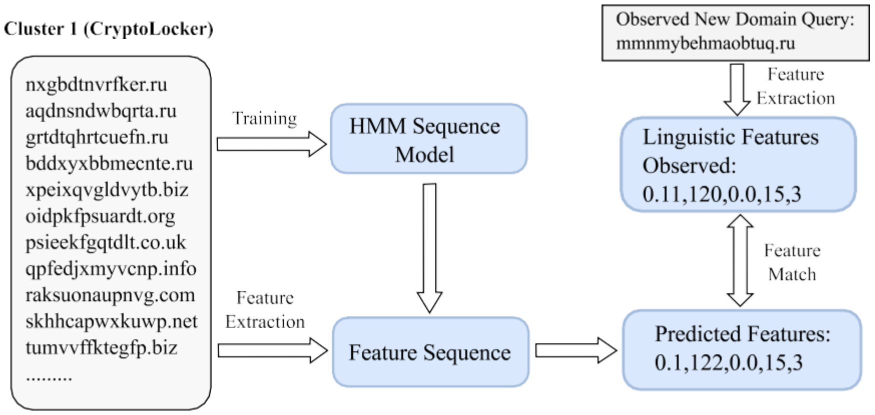


4.2 The Machine Learning Framework

We propose a machine learning framework that consists of threeimportantsteps,asshowninFigurebelow.Wefirsthavethe DNS queries with the payload as the input. Then, the DNS queries will be passed to our process step



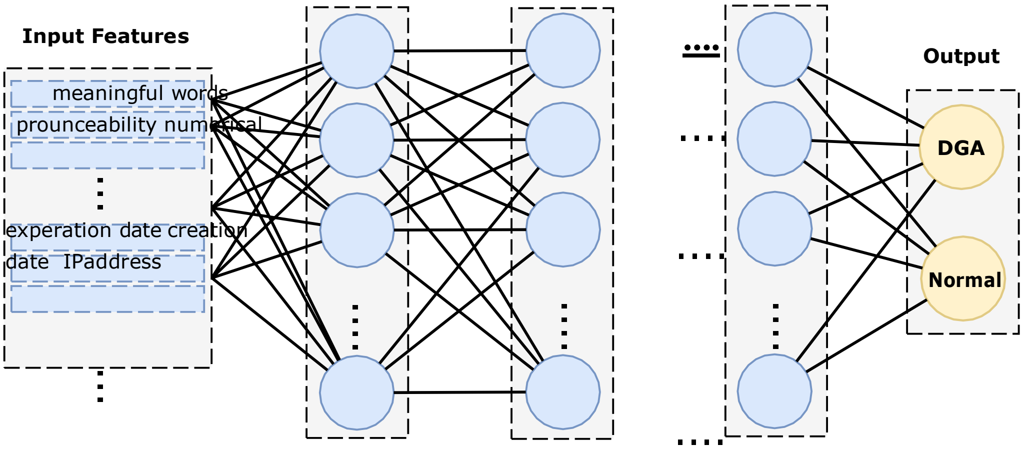
4.3 A representative example of a clustering (B) HMM training procedure (C) Workflow of the HMM model prediction



An example of the HMM model prediction

4.4 A TIME SERIES PREDICTOR

To analyze the clustering result, we build an HMM-based time-series prediction model to predict incoming DGA do- main features. Figure 5 (A) shows an example dataset derived from clustering. We use every domain cluster to train a separate HMM model. We distinguish the model from training and prediction stages. Figure 5 (B) shows the HMM training step. Each HMM data record represents a series of domain observations. First, a sequence of domain names are processed by a feature extractor and each of these feature vectors is used as a training record. Then, similar sequences are clustered as a group of DGA domain names with certain outcomes. After the training process, if a sequence does not haveanHMMsequencerepresentation(oritisnotpresented in the training data but the test data), the HMM model then generates the future predicted results. Otherwise, we will use an existing HMM sequence representation. Figure 5 (C) shows the HMM prediction workflow. Once the model has beentrained,asetoffeaturesisformedbyaseriesofDGA



The proposed deep learning model

4.5 CONCLUSION

Detecting DGAs is a grand challenge in security areas. Blacklisting is good for handling static methods. However, DGAs are usually used by an attacker to communicate with variety of servers. They are dynamic, so simply using the blacklisting is not sufficient for detecting a DGA. In this research, we have proposed the machine learning framework with the development of a deep learning model to handle DGA threats. The proposed machine learning framework consists of a dynamic blacklist, a feature extractor, a two- level machine learning model for classification and cluster- ing, and a prediction model. We have collected a real-time threat intelligence feed over a one-year period where all domains live threats on the Internet.

REFERENCE

1. K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, “Learning and classification of malware behavior,” in International Conference on Detec- tion of Intrusions and Malware, and Vulnerability Assessment. Springer, 2008, pp. 108–125.

2. T. Chin, K. Xiong, and M. Rahouti, “SDN-based kernel modular coun- termeasure for intrusion detection,” in Proceedings of 13rd EAI Interna- tional Conference on Security and Privacy in Communication Networks. Springer, 2017.

3. U. Ghosh, P. Chatterjee, D. Tosh, S. Shetty, K. Xiong, and C. Kamhoua, “An SDN based framework for guaranteeing security and performance in information-centric cloud networks,” in Proceedings of the 11th IEEE International Conference on Cloud Computing (IEEE Cloud), 2017.

4. C. Khancome, V. Boonjing, and P. Chanvarasuth, “A two-hashing table multiple string pattern matching algorithm,” in Tenth International Con- ference on Information Technology: New Generations (ITNG). IEEE, 2013, pp. 696–701.

5. S. Schiavoni, F. Maggi, L. Cavallaro, and S. Zanero, “Phoenix: DGA-based botnet tracking and intelligence,” in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, 2014, pp. 192–211.