e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

GRAPH BASED BOTNET DETECTION

Jajam Naga Sreeja[*1], G. Varsha Reddy[*2], V. Vishal[*3], V. Anitha[*4]

[*1,2,3]Student, Department of Computer Science, B V Raju Institute of Technology, Narsapur, Telangana, India.

[*4]Assistant Professor ,Department of Computer Science, B V Raju Institute of Technology, Narsapur, Telangana, India.

ABSTRACT

Bot detection using machine learning (ML), with features of network flow level, is widely read in the literature. However, existing flow-based methods often introduce high overhead calculations and do not completely capture network communication patterns, which may expose additional characteristics of malicious hosts. Recently, bot detection systems used to analyze the communication graph using ML gained attention to overcome these limitations. The graph-based approach is accurate, as graphs are a true picture of network communication. In this paper, we proposed, a two-phase bot detection system that uses checked and unchecked ML. The first phase cuts the potential athletes, while the second phase gains bot detection with high accuracy. Our proposed system prototype implementation detects many types of bots and shows resilience to zero-day attacks. It also accepts different network topologies and is suitable for large-scale data. Compared to modern, Proposed system exceeds the end-to-end system that uses streaming-based features and performs very well online.

Keywords: Detection, Features, Resilience, Communication, Attacks.

I. INTRODUCTION

Organizations are regularly under threat of security, which not only costs billions of dollars in damages and repairs, often harmful to their own reputation. The botnet-assisted attack is a well-known threat to these organizations. A botnet is a collection of bots, agents that are not at risk, and are controlled by botmasters through command-and-control channels (C2). A vicious enemy controls the bots through a botmaster, which can no longer be distributed to several agents living inside or outside the network. Therefore, bots can be used for activities ranging from distributed denial-of-service (DDoS) to bulk spam, to fraud and theft detection. While bots thrive with different evil intentions, they show the same behavior pattern when read closely. Intrusion kill-chain controls the stages where vicious agent goes through in order to achieve his goal. They establish the basis for normal behavior in a secure system and then design the decision engine. The decision engine determines and notifies any deviations from the standard as a threat. Machine learning will automatically detect the normal behavior of the system. The use of ML has enhanced the robustness and accuracy of confusing IDS. The most used learning parameters in ML include surveillance and surveillance. Supervised learning uses labeled data sets to create models. It is used to read and identify patterns in known training data. However, labeling is no small feat and often requires domain professionals to label the data sets in person. This can be difficult and prone to make mistakes, even in small databases. Unsupervised learning, on the other hand, uses labeled data sets to create models that can distinguish between data patterns.

II. METHODOLOGY

Our proposed system has 5 modules:

1. Uploading CTU -13 dataset

We upload CTU-13 dataset and this dataset contains 13 observation and each observation contains PCAP files and capture file. To generate graph we need to used capture file not PCAP file.

2. Apply KMEANS to separate Bot & Benign Data

After applying KMEANS, dataset size will be reduced as benign records will be removed.

3. Run Flow Ingestion & Graph Transformation

The system creates a graph G where Vertex is a set of nodes and Edge is a set of directed edges between those nodes. This module shows see total nodes and edges generated and time taken to calculate betweenness, alpha centrality and clustering.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

4. Feature Extraction and Normalization

Features are extracted from graphs. After features extraction from graph we will go for normalization to get mean values of each features. Normalized features will be used to train decision tree classifier and this model can be used to predict type of future requests.

5. Run Decision tree algorithm

This module divides the normalized dataset features into training and testing data, after building train model we apply test records and model detects number of malicious and benign bots present. Also calculates accuracy, precision and recall.

III. MODELING AND ANALYSIS

System consists of three major components as shown in Figure1 below. The components are data preparation and feature extraction, model training, and inference.

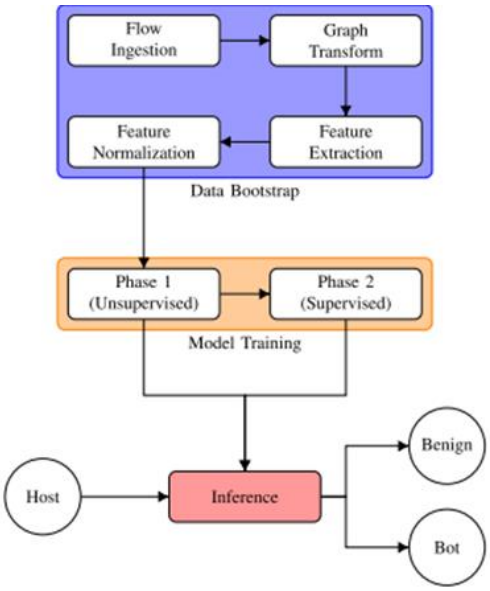

Figure 1: Components of botnet detection.

Data Preparation

1) Flow Ingestion

The contribution to the framework are bidirectional organization streams. These streams convey data of the fundamental IP layer. We want to apply decrease on stream, subsequently makes a bunch of information that contains 4-Tuple streams which incorporates source ip address, number of information bundles from source to objective, objective ip address, number of objective packets.

2) Graph Transform

The framework makes a chart where Vertices are set of hubs and Edges are set of guided edges from hub vi to hub vj with weight. The Framework makes diagram based highlights like in-degree, out-degree, in-degree weight, out-degree weight, betweenness centrality, alpha centrality for the AI models.

3) Feature Extraction

Features are intrinsic to fulfilment of model gaining knowledge that need to constitute and discriminate host behavior.

4) Feature Normalization

Normalization facilitates in bringing all functions to identical scale, which in addition facilitates in quicker convergence of learning.

Model Training

The model accepts graph-based features as input and distinguish between malicious and benign hosts. It contains 2 phases.

Phase 1: System performs K-Means algorithm to cluster the hosts, this is done observing similar patterns in data, such as sending and receiving similar number of packets. This creates 2 clusters, they are benign and bot.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

Phase 2: Phase 1 separates the dataset between nodes that are inside and outside the benign cluster. All the nodes that present outside the benign cluster are input to Phase 2 for further classification. Optimally, all the bots should be outside the benign cluster. Depending on the amount of hosts outside the benign cluster, the supervised learning (SL) classifiers used in this phase will exhibit different results.

Inference

When the models are prepared, they are conveyed in the framework to perform bot identification. In this framework, the surmising should be possible in two stages ;probable harmless hosts get sifted through in Stage 1 as they get grouped into the harmless bunch, while malignant hosts that are arranged into an alternate group are additionally ordered in Stage 2

IV. RESULTS AND DISCUSSION

We get normalized data after performing feature extraction and normalization. This data is further divided into training and testing data.

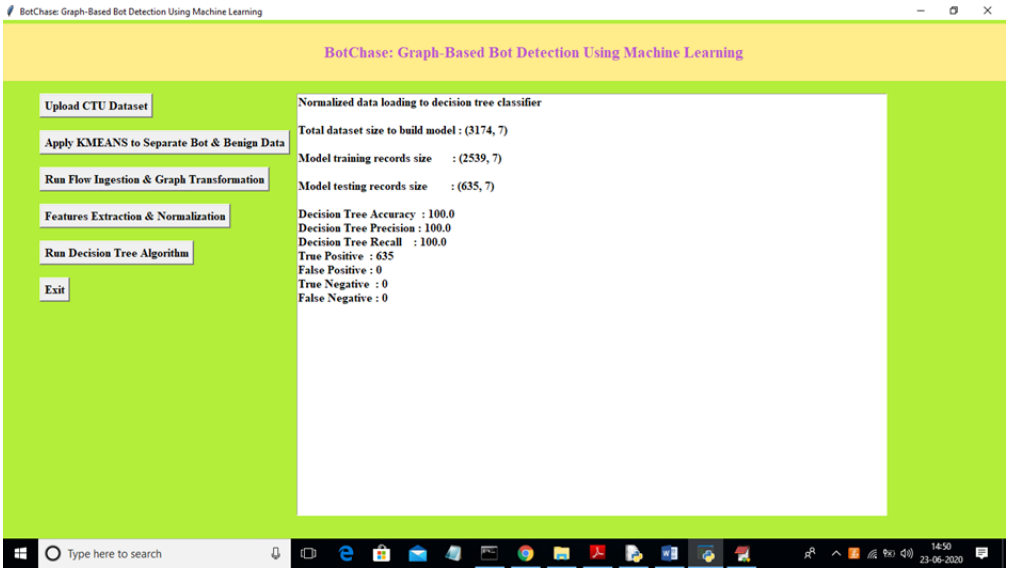

Figure 2: Output Screen

In above Figure 2 after normalization we got total records as 3174 with 7 columns (in-degree, out-degree, weight etc.) and application split total records into train size as 2539 and test size as 635. After building train model we apply test records and got accuracy as 100%.

V. CONCLUSION

The proposed system is capable of efficiently transforming network flows into a graph model. It leverages two ML phases to differentiate bots from benign hosts. In 1st phase, system uses K-Means and distinguishes bots from benign hosts. Furthermore in 2nd phase it uses decision tree algorithm to classify bots. Proposed system is able to detect bots that rely on different protocols, proves robust nature against unknown attacks and cross-network Machine Learning model training and inference.

VI. REFERENCES

[1] J. Caballero, C. Grier, C. Kreibich, and V. Paxson, “Measuring pay-perinstall: The commoditization of malware distribution,” in Proc. USENIX Security, 2011, p. 13.

[2] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, “Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains,” Inf. Warfare Security Res., vol. 1, no. 1, p. 80, 2011.

[3] R. Boutaba et al., “A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities,” J. Internet Services Appl., vol. 9, no. 1, pp. 1–99, 2018.

[4] B. Venkatesh, S. H. Choudhury, S. Nagaraja, and N. Balakrishnan, “BotSpot: Fast graph based identification of structured P2P bots,” J. Comput. Virol. Hacking Techn, vol. 11, no. 4, pp. 247–261, 2015.