Information Sciences 278 (2014) 488–497

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests

Kamaldeep Singh [a,1], Sharath Chandra Guntuku [b,*,1], Abhishek Thakur [a], Chittaranjan Hota [a]

[a] Department of Computer Science, BITS – Pilani, Hyderabad Campus, AP 500078, India
[b] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Drive, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form 9 February 2014
Accepted 11 March 2014
Available online 25 March 2014

Keywords:
Hadoop
Mahout
Peer-to-Peer
Botnet detection
Machine learning
Network security

ABSTRACT

Network traffic monitoring and analysis-related research has struggled to scale for massive amounts of data in real time. Some of the vertical scaling solutions provide good implementation of signature based detection. Unfortunately these approaches treat network flows across different subnets and cannot apply anomaly-based classification if attacks originate from multiple machines at a lower speed, like the scenario of Peer-to-Peer Botnets.

In this paper the authors build up on the progress of open source tools like Hadoop, Hive and Mahout to provide a scalable implementation of quasi-real-time intrusion detection system. The implementation is used to detect Peer-to-Peer Botnet attacks using machine learning approach. The contributions of this paper are as follows: (1) Building a distributed framework using Hive for sniffing and processing network traces enabling extraction of dynamic network features; (2) Using the parallel processing power of Mahout to build Random Forest based Decision Tree model which is applied to the problem of Peer-to-Peer Botnet detection in quasi-real-time. The implementation setup and performance metrics are presented as initial observations and future extensions are proposed.

1. Introduction

Botnet attacks are one of the biggest challenge that security researchers and analysts face today on an International scale. The economic losses caused by breach of computer networks step up to several billions of dollars. Just a few months prior to this writing, a massive DDoS attack had targeted the admin accounts of numerous WordPress users. Online security communities and blogs suspect it to be a Peer-to-Peer (P2P) based Botnet [43]. It was revealed that at least 90,000 unique IPs were utilized to achieve the attack [40]. It is suspected that this attack was part of a bigger plan, where the attacker wanted to compromise numerous WordPress servers and then use the army of bots to launch an even larger DDoS attack. Hence, even though P2P Botnets have been on the rise since one decade now, detecting and mitigating their attacks is still a challenge.

For detecting and mitigating such attacks, network traces and packet captures are amongst the most valuable resources for network analysts and security researchers. With such attacks increasing every day the magnitude of network traces handled is expanding exponentially. However, computer systems lack the hardware and are fundamentally limited by the device

[*] Corresponding author.

E-mail addresses: mailkamaldeep@gmail.com (K. Singh), sharathc001@e.ntu.edu.sg (S.C. Guntuku), abhishek@hyderabad.bits-pilani.ac.in (A. Thakur), hota@hyderabad.bits-pilani.ac.in (C. Hota).

[1] Authors 1 and 2 have equal contribution.

http://dx.doi.org/10.1016/j.ins.2014.03.066

fabrication restrictions to accommodate the gigantic size of the datasets. This has led to an upsurge of interest in distributed algorithms, taking advantage of the multi-core architectures and distributed computing.

In the past, researchers have used various techniques like Signature and Anomaly-based Intrusion Detection Systems (IDS) to deal with the issue of network security threat detection. But these solutions have scalability issues while dealing with the large datasets that have been discussed above. Though there were kernel scaling methods proposed [28], there were issues with datasets having high variance. It was found that when dataset has high variance, the larger the dataset, the better the training accuracy of the model [31]. In a high variance scenario, which happens when data is over-fit, the training error will be low and the cross validation error will be much higher than the training error. The intuition behind this is that training is over-fitting the data, the model will memorize the data. But then the model will generalize poorly for new observations, leading to a much higher cross-validation error. This can be corrected by increasing the training data set size which would lower the cross-validation error, in the scenarios where variance is high. Thus data sets demand the use of Big Data Technologies. Several large datasets containing the malicious activity of various bots have been captured and released by CAIDA and other organisations. One of these traces, which was captured by UCSD, of 40 GB, [48] was used for this research and this necessitated the existence of a scalable framework to train the classification module.

Hence, in this research a scalable and distributed intrusion detection system has been proposed which can handle heavy network bandwidths. This framework is built on top of Hadoop, which is an open-source software framework that supports data-intensive distributed applications and leverages the libraries that are designed to use the power of clustered commodity machines. This framework utilizes Apache Mahout which has machine learning algorithms to build predictive models.

The rest of the paper is organized as follows. Section 2 describes the related work in the realm of P2P Botnet detection which are based on machine learning algorithms and the application of Hadoop in this area. Section 3 describes the experimental setup and the methodology used for the framework to achieve security threat detection in real-time. Section 4 deals with the specific application of Peer-to-Peer Botnet detection using this framework and Section 5 concludes with results and scope for future work.

2. Related work

Over the last few years, several researchers have proposed machine learning based solutions for mitigating security threats. Research has systematically drifted from signature-based methods to more semantic-based methods like developing Ontological models to handle web application attacks [36] and Hidden Markov Models for spam detection [37]. In [5] researchers have proposed security aware agent based systems which can manage their own security at runtime. There has also been research on using large network traces for the mitigating security threats [24] where authors proposed a Hadoop based DDoS attack detection method. They have developed a novel traffic monitoring system that performs NetFlow analysis on multi-terabytes of Internet traffic in a scalable manner [23]. They have devised a MapReduce algorithm with a new input format capable of manipulating *libpcap* files in parallel. But their approach hardcodes the features that can be extracted from the *libpcap* files and thereby the user is not presented with the flexibility to decide on the feature set based on the problem instance. As per the current knowledge of the authors, the area of network security analytics severely lacks prior research in addressing the issue of Big Data.

On the contrary, there has been significant research in the area of security threat detection using machine learning techniques. Authors [52] differentiate the network flow records based on certain features related to traffic volume and categorize them as malicious and benign. They also present how the plotters could change their behavior to evade their detection technique, which was observed in Nugache, which is known to randomly change its behavior. Authors in [17] show that P2P Botnet detection can be achieved with high accuracy based on a novel Bayesian Regularized Neural Network.

In [9] authors provide an overview on the adversarial attacks against IDSs and highlight the most promising research directions for the design of adversary-aware, harder to defeat IDS solutions. Researchers [42] have also worked on building anomaly detection systems which are unsupervised in nature. Markovian IDS has been developed to shield wireless sensor networks based on mining the attack patterns [20].

Researchers [39] report their work on detecting Sinit and Nugache that these bots communicate on the same port. They also indicate that a large number of destination unreachable error messages and connection reset error messages were observed. But bots which encrypt their payloads will disable the authors technique from working as their method depends heavily on the payload signatures. Researchers [11] explain the challenges and features when dealing with Nugache P2P Botnet. Authors in [45] conclude that it is impossible for a static Intrusion Detection System (IDS) to detect Nugache traffic.

Researchers [38] worked on understanding the role of reputation and trust in the P2P domain. There have also been works [21] trying to classify applications based on the traffic generated to characterize them. There have also been work [7] on developing privacy preserving packet anonymization frameworks for transforming packets to suite research purposes. Authors analysis of Storm in [45] concludes that the bot can be detected by configuring an IDS to find the configuration file used by the bot. But it is difficult to distinguish between legitimate P2P communication and Storm bot's communication as the behavior of Storm can be camouflaged to look like legitimate P2P communication. To address this issue, authors in [19] develop ways to mitigate the Storm worm and introduce an active measurement technique to enumerate the number of infected hosts by reverse engineering of the bot's binary executable, in order to identify the function which generates the key that is used for searching other infected machines and bots.