Detecting Anomalies with Unsupervised Learning

Julina Zhang, Tianye Song, Kerry Jones
University of Virginia, DSI

Abstract

Networks are extremely vulnerable to cyber-attacks. Many institutions depend on networks to facilitate the exchange of private and public information. Within any given network, there are diverse set of user types (students, IT professionals, business owners), and user activities (browsing web pages, Facebook messaging, video streaming, and file transferring). This wide range of user and user activities makes verifying communication on a network nearly insurmountable. Thus leaving personal and sensitive information held within these networks vulnerable to attack.

Leveraging off previous research utilizing unsupervised methods, our approach aims to use information derived from IPv4 packet headers to cluster network activity and provide baseline of normality.

Network Intrusion Detection Background

The first virus appeared in 1971. Since then, cyber-attacks have become a growing nuisance to companies and institutions around the world, particularly since we are living in a global economy heavily dependent on the internet and network connectivity [1]. Institutions both, public and private store, transfer, and process petabytes of data on a daily basis. Many, of which, require open internet to facilitate the exchange of information.

Within any given network, there are diverse set of user types (students, IT professionals, business owners), and user activities (browsing web pages, Facebook messaging, video streaming, and file transferring). This wide range of user and activities makes verifying communication on a network nearly insurmountable. Thus computer networks, particularly at large institutions, such as the University of Virginia, are vulnerable to security breaches. Hackers and other agents of nefarious activity are constantly trying to breach UVA's network to obtain valuable sensitive and personal information.

Networks within organizations need to be open to facilitate the free exchange of information, solving the problem can also bring extra value for the organization. The recent attack on the central bank of Bangladesh using the SWIFT network illustrates these vulnerabilities. Until this attack the SWIFT network was considered among the most secure in the world[2] .

Intrusion detections systems are a common solution used identify malicious activity within internal networks. For our purposes, network intrusion detection is a set of activities used to compromise and detain sensitive information from a network that can be detected via suspicious abnormal network behavior [3].

In anomaly detection, abnormal activities are defined as activities that deviates from expected or "normal" behavior, or outliers. It is important to clarify that abnormal activity does not always indicate malicious activity or attack. In return, outliers can be an indication of an inadequate model. Abnormal activities can occur when a user downloads software or streams a video from a website that they have not previously visited. Given the nature of our problem, these

[1] Boutaba, Carol Fung and Raouf. *Intrusion Detection Networks*. CRC Press, 2013.

[2] Yuning Ling, Marcus Rosti, Gregory Swanson. "A Hands-off Approach to Network Intrustion Detection." IEEE Systems and Information Engineering Design Conference (SIEDS). Charlottesville : IEEE, 2016. 216-220.

[3] Ashfaq, Rana Aamir Raza, et al. "Fuzziness based semi-supervised learning approach for intrusion detection system." Information Sciences (2016).

activities would be classified as abnormal activities because the user does not typically visit these particular sites frequently. Without more information, it would be difficult to distinguish them from attacks.

This project aims to investigate whether the accumulated information of where data is going with a network can provide vital information for identifying malicious hosts. We will do this using unsupervised learning algorithms to discover behavior types. From these clusters, we will be able identify characteristics of abnormal transmissions to destination IPs.

Previous Literature in Intrusion Detection

First introduced in 1980s, Denning provided an automated model capable of detecting security violations. The model assumes it can detect malicious behavior through defined patterns of abnormality[4]. There are two mainstream approaches for classifying network behaviors: misuse detection and anomaly detection [3]. In misuse detection, the IDS attempts to classify network behaviors by matching them with a known set of signatures. Misuse detections fail due to their inability to detect new attacker behavior patterns. Anomaly detection techniques attempt to identify outliers that deviate from normal behavior. This approach suffers from high and costly false positive rates.

In 2001, Lancope developed an intrusion detection system called Stealthwatch that detects deviations from good behavior[5]. Stealthwatch is a netflow-based network security monitoring system developed by Lancope that provides in-depth network visibility for enterprises. It employs sophisticated behavioral analytics and heuristics to analyze 90+ attributes of network flow data to establish baseline normal behavior of a host[6]. If it detects deviations from the baseline, it triggers the alarm and alert the security analysts. Netflow-based security analysis allows rapid detection of novel attack vectors by leveraging behavioral analysis and pattern recognition techniques, in contrast to signature-based intrusion detection.

Additionally, Machine learning, supervised and unsupervised, approaches are increasingly being used to increase the efficacy of IDS. KDDUCUP’99, an intrusion dataset built in 1999 is a popular dataset used to assess approaches such approaches such as K-NN, Neural Networks and SVM. However, these supervised techniques fail short of identifying new pattern types. Thus most consider the dataset obsolete because many new attacks types have been developed since then [3].

Anomaly detection techniques typically implement unsupervised learning techniques such as clustering because most investigations include unlabeled data [3]. However, properties shared by the elements of a cluster can only describe the similarity between them and the dissimilarity between other clusters.

In addition, different types of clustering algorithms produce different clusters. For example, clusters formed using DBSCAN may differ than clusters formed using K-means. DBSCAN treat

[4] Denning, Dorothy E. "An intrusion-detection model." IEEE Transactions on software engineering 2 (1987).

[5] Herring, Charles. Academic Hackademic. n.d. <https://www.lancope.com/blog/academic-hackademic>

[6] Grimes, Robert. Detect network anomalies with StealthWatch. 2014. IDG. 2016. <http://www.infoworld.com/article/2848768/security/detect-network-anomalies-with-stealthwatch.html>.

some of the observation as noise, whereas K-Means puts all elements in clusters. This can be seen in Erman, Arlit and Mahanti's investigation using the two algorithms[7].

In 2006, Erman, Arlitt, and Mahanti applied these techniques to network traffic classification [7]. The techniques exploit the distinct behavior patterns of application when they communicate on a network [7]. Comparing the overall accuracy rates of K-Means and DBSCAN to that of AutoClass, a probabilistic model-based clustering technique that automatically determines the number of clusters and the cluster parameters, K-Means performed only a slightly poorer than AutoClass, and DBSCAN performed poorer than K-Means and AutoClass.

Using the K-means algorithm, it was found large values of K can lead to overfitting. Additionally, although DBSCAN as the ability to place the majority of connections into a few clusters that have a high predictive power of a single category of traffic. The technique tends to misclassify connections as noise and when epsilon is high, it combines different categories of traffic into one cluster [8].

In 2015, Rosti et al. also applied clustering to problem. Using hierarchical clustering algorithm to classify netflow (source IP to destination IP) and then clustered on source IP to differentiate between user groups [2].

In 2016, Aamir et al. designed a SSL classifier for improving the accuracy of a classifier's performance by finding the relationship between classifier's fuzziness and its misclassification on unlabeled samples.

Novel intrusion methods are constantly changing. Intruders devise attack vectors to avoid detection from existing intrusion detection systems. Thus instead of developing an approach to discover a specific type intrusion behavior. Thus our intention is to differentiate "normal" behavior from "abnormal" behavior in UVA's network. We aim to develop an approach that generalizes behavior and allows us to find behavior that deviate from generalized activity types.

Objective

This project aims to investigate whether the accumulated information from the source IP address can provide vital information for identifying malicious hosts. Our aim is to try and evaluate abnormality based on between IP addresses based on three aspects. In no particular order we would like to understand normality in respect to location, temporal and size of data transfer. By investigating location, we can get a baseline of the most common addresses being IPV4 packets are being sent to. Additionally, we can understand of what time this typical behavior occurs and the size of typical size of the packets being sent with respect to normal patterns of behavior and occurrence. Once we established these baselines, we can get idea of abnormal behavior with regards to three aspects.

Data

For our investigation are primarily interested in extracting information from IPv4 Network Packet headers. IPv4 is a common internet network protocol[8]. An IP address is used to uniquely identify end user's location on network. The IPv4 packet is a series of bytes that describes information related to the delivery of information from one location to the next[9]. They are typically length of 20 bytes, consisting of 14 fields. Most of which may not reveal be indicative of a hacker decides partake in malicious activity but can at least provide information relating to source and

[7] Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." Proceedings of the 2006 SIGCOMM workshop on Mining network data. ACM, 2016. 281-286.

[8] Fairhurst, Gorry. *The IP Address*. 2007. 2016.
<http://www.erg.abdn.ac.uk/users/gorry/course/inet-pages/ip-address.html>

destination IPs, and how large it is [2] . Figure 1, is visual depiction of an IPv4 Header. For this investigation will we focus on extracting patterns from the source and destination IPs. Source Address and Destinations are the IP addresses of the sender and the recipient of the packet, respectfully. Additionally, we will investigate the value-added from the remaining information in the netflow data through exploratory analysis.

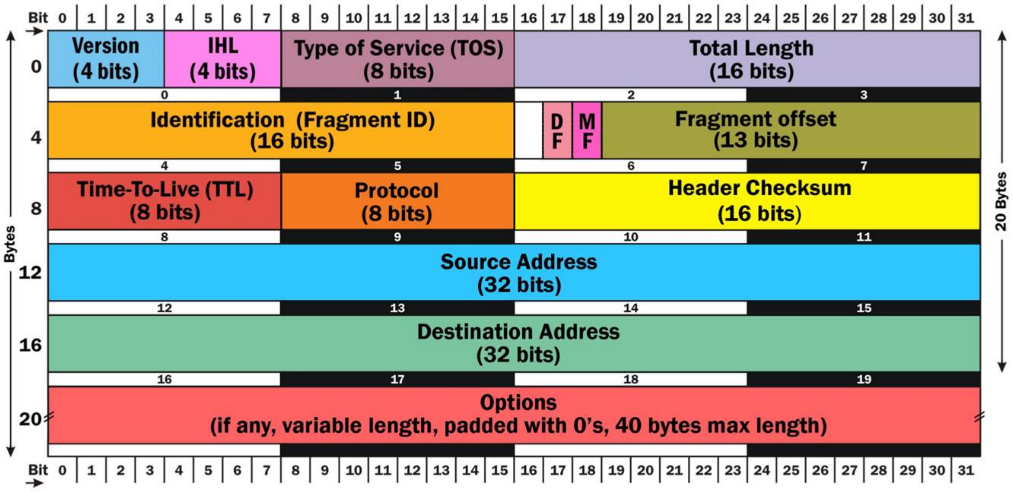

Bit 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Version (4 bits)
IHL (4 bits)
Type of Service (TOS) (8 bits)
Total Length (16 bits)
Identification (Fragment ID) (16 bits)
D F
M F
Fragment offset (13 bits)
Time-To-Live (TTL) (8 bits)
Protocol (8 bits)
Header Checksum (16 bits)
Source Address (32 bits)
Destination Address (32 bits)
Options (if any, variable length, padded with 0's, 40 bytes max length)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Bytes
0
4
8
12
16
20
20 Bytes

Figure 1-IPv4 Packet Header

The data will be provided by the UVA Information Security, Policy, and Records Office. Additionally, due to sensitivity, security, and user privacy concerns, data provided by the UVA Information Security, Policy, and Records Office will be anonymized.

Technical approach

Data Preprocessing

We are interested in understanding patterns of data that are leaving UVA network. Data leaving UVA network means the source IP in the IPv4 header will fall into the domain of UVA network host IPs. Thus, we will first subset the dataset with respect to source IP addresses. The resulting subset will include only data with source IP within the UVA network host IP domain.

We will further process subset by computing the following information with respect to the destination IP: frequency of visit, location, and the average size of the data being transmitted. This will allow us to see which destination IP addresses are frequently visited and how much data is typically being transmitted to an IP address. Furthermore, we can use destination IP addresses as primary indices for the data. Doing so will reduce the retrieval time for single data from linear to logarithmic. An optimal searching method is crucial in our investigation given the sheer size of the dataset.

Cluster Analysis

Our next step is to cluster the data and identify the characteristics of abnormal transmissions to destination IPs. We will cluster the destination IPs based on its frequency of occurrence, location, and transmission's average size. We will evaluate different clustering strategies using silhouette[9], a metric used to assess how similar the clusters are internally and

[9] Rousseeuw, Peter J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." Journal of Computational and Applied Mathematics 20 (1987): 53-65.

how dissimilar they are from surrounding clusters. We will investigate different clustering techniques using trial-and-error way. These methods include but aren't limited to: K-means, Leader clustering, and hierarchical clustering.

For illustration purposes, an example of anomaly transmission may have the following properties: 1) The destination IP is located at a location that is seldom visited by users within UVA network, e.g. a remote location in a foreign country; 2) The amount of data being transmitted to the destination IP address is abnormally large.

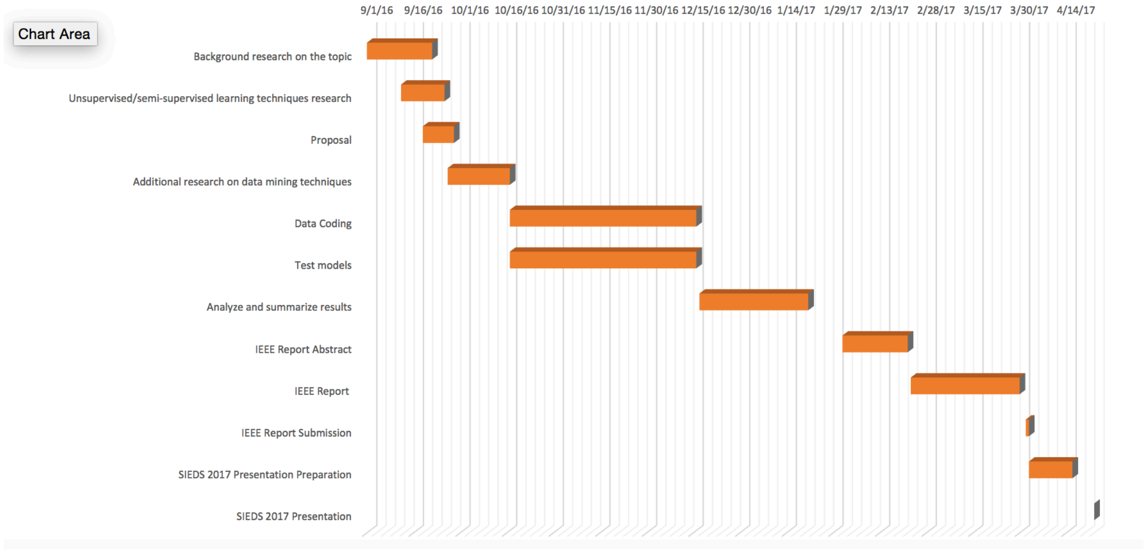
Provide deliverables

Project deliverables include: data segmentation method, clustering and time series analysis model, software application, and a paper that will suffice certain standards. All codes and models developed will be provided to UVA Information Security, Policy, and Records Office.

Budget

SIEDS Conference attendance cost
AWS usage cost

Schedule



Qualifications

Kerry Jones, M.S. Candidate in DS, University of Virginia - Over four years' experience as a Research Scientist with the Department of Defense. In the past, his research has including spatial data mining and modeling. He has experience using Python, R, and Java. He has a Dual Bachelors in Geography and Government.

Tianye Song graduated in B.A. degrees in Computer Science from the University of Virginia. Being a Computer Science major, he has experience in python and R. He will utilize his programming skillset and contribute to this project.

Julina Zhang graduated from the University of Virginia in 2015 with a Bachelor of Arts degree in Economics and Statistics. She has experience with R, Python, Java, and SAS. She will utilize these skills along with her statistical background knowledge.

Bibliography

Ashfaq, Rana Aamir Raza, et al. "Fuzziness based semi-supervised learning approach for intrusion detection system." *Information Sciences* (2016).

Boutaba, Carol Fung and Raouf. *Intrusion Detection Networks*. CRC Press, 2013.

—. *Intrusion Detection Networks: A Key to Distributed Security*. CRC Press, 2013.

Denning, Dorothy E. "An intrusion-detection model." *IEEE Transactions on software engineering 2* (1987).

Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." *Proceedings of the 2006 SIGCOMM workshop on Mining network data*. ACM, 2016. 281-286.

Fairhurst, Gorry. *The IP Address*. 2007. 2016. <http://www.erg.abdn.ac.uk/users/gorry/course/inet-pages/ip-address.html>.

Grimes, Robert. *Detect network anomalies with StealthWatch*. 2014. IDG. 2016. <http://www.infoworld.com/article/2848768/security/detect-network-anomalies-with-stealthwatch.html>.

Herring, Charles. *Academic Hackademic*. n.d. <https://www.lancope.com/blog/academic-hackademic>.

Rousseeuw, Peter J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics* 20 (1987): 53-65.

Sommer, Robin, and Vern Paxson. "Outside the closed world: On using machine learning for network intrusion detection." *2010 IEEE symposium on security and privacy* (2010).

Trevor, Hastie, Tibshirani Robert and Jerome Friedman. *The Elements of Statistical Learning*. Vol. 2. Springer, 2009.

Yuning Ling, Marcus Rosti, Gregory Swanson. "A Hands-off Approach to Network Intrustion Detection." *IEEE Systems and Information Engineering Design Conference (SIEDS)*. Charlottesville : IEEE, 2016. 216-220.