Machine learning on network analysis

“Software written today differs from that written in the past in two main ways: it is more complex and handles more data than ever before. When Windows 95 was released it contained 11 million lines of code, in contrast six years later Windows XP had 40 million lines [1]. This gives rise to a greater need for systematic testing

Abstract

This paper attempts to incorporate machine learning techniques to improve network security. Is this really true? I thought my aim was to test IDS.

Introduction (this does not fit abstract, write that at the end)

“There are **4.39 billion** internet users in 2019, an increase of 366 million (9 percent) versus January 2018.” [1] <https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates>

Stats $6 trilion. 600% increase on Iot 69% anit virus

[2] <https://www.varonis.com/blog/cybersecurity-statistics/>

Smart water bottle

[3] <https://gizmodo.com/15-idiotic-internet-of-things-devices-nobody-asked-for-1794330999>

[4] <https://ieeexplore.ieee.org/abstract/document/6978614> internet of things issue

[5] <http://www3.weforum.org/docs/WEF_GRR18_Report.pdf> offensive cyber capabilities….

[6] <https://www.calyptix.com/top-threats/top-8-network-attacks-type-2017/> network attack types

With the rapid expansion of the Internet, it has become an essential part of our lives with over half the population connected [1]. However, this results in an increasingly complex and fragile network. Many systems are left vulnerable, waiting to be exploited. By 2021, it is predicted that cost of cyber-attacks will reach $6 trillion [2]. The importance of a good and secure security system is far too crucial.

“Offensive cyber capabilities are developing more rapidly than our ability to deal with hostile incidents” [5]. Attacks are becoming smarter, polymorphic viruses and obscured malwares are passing through current systems. Over a third of organizations believe that the threats they are facing cannot be blocked by their anti-virus [2].

Due to our rapid growth, we have left many openings for an attack, one of them is through the network. Our need for constantly being connected is causing a major gap in security. In 2017, 8 different network attacks dominated the market. [6]

1. Browser attacks - malicious users target vulnerable websites to infect, infecting new genuine users.
2. Brute force attack - attempting to guess your way through to the system.
3. Denial of service (DoS) or Distributed Denial of Service (DDoS) – flooding a service by creating many requests in order to slow or crash the system.
4. Worm attacks – self propagating program that spreads through local system through exploitable vulnerabilities.
5. Malware attacks – programs that can take many forms, however their purpose is always malicious.
6. Web attacks – exploiting vulnerabilities found in the website such as SQL injection.
7. Scan attacks – indirect attack to gain knowledge of any vulnerabilities that exist such as an open port.
8. Other attacks – attacks that were out of scope, such as physically attempting to steal device.

Fortunately, methods such as Intrusion detection system (IDS) exist to deter most of these attacks. IDS constantly scan the network for any anomalous activity in the network. Some are even capable of stopping the attack completely rather than just alerting the user.

However, IDS face many issues such as explaining what an anomaly is in the first place. Robustness and accuracy also come into question. How often does an IDS system report false negatives or how many different types of malwares can they detect?

By using machine learning it is possible to overcome these problems. Its ability of learning patterns and understanding different classifications can assist IDS. Todo add more here about machine learning.

I aim to create a system that can detect malwares on a network, and also test its robustness and accuracy rates.

Intrusion detection systems (IDS) is one of many methods of securing against cyber-attacks. IDS constantly scan for any anomalous activity, with some even performing actions to deter such activity.

However, defining what is an anomaly is difficult. Machine learning as a whole, excels at such problems. Capable of learning patterns and understanding different classifications can help solve this problem.

Networks, which are simply a connection of two or more nodes capable of sharing resources, are incredibly varied yet engrained in our everyday lives. Along with the boom of Internet of Things, devices capable of communication and storing valuable information, networks are left more vulnerable [4].

A network is a connection of two or more nodes which can share resources. The Internet happens to be the largest network out there. There are constant attacks targeting networks with new attack vectors being discovered.

Especially with the boom of Internet of Things. Smart phones, smart watches and even smart bottles [3] are increasing in popularity. In 2017, attacks on these IoT devices increased by 600% [2].

Intrusion detection systems (IDS) are one method of creating a secure network environment. IDS constantly scan for any anomalous activity, with some even performing actions to deter such activity. However, defining what is anomalous is difficult.

Machine learning has exploded in many areas

There are many vulnerabilities in a network such as … machine learning is good at learning patterns of these attacks …

The increase in Internet of Things (IoT) such as smart phones, smart watches and even smart bottle [3] has caught the attention of attackers. In 2017, attacks on IoT have increased by 600% [2]. IoT devices have the ability to connect to the Internet, creating a passage

They create a vulnerable passage on

Everything nowadays is “smart”. Smart phones, smart watches, smart lightbulbs and even smart water bottle [3]. The increase of these devices known as Internet of Things (IoT) have created a huge network, containing valuable information however are left with massive security flaws [4]. In 2017, the attacks on these IoT devices have increased by 600% [2]. Malicious users are able to constantly find new attack vulnerable vectors.

Intrusion detection systems (IDS) are on the forefront of tackling against network attacks. IDS are constantly scanning

A network is a connection of two or more nodes which can share resources. The Internet happens to be the largest network out there. There are constant attacks targeting networks with new attack vectors being discovered.

<https://www.calyptix.com/top-threats/top-8-network-attacks-type-2017/>

Types of network attacks

1. DoS – flood traffic to services in order to overwhelm them, causes slower service response or shut it down completely.
2. Malware – types of malware constantly change with constant new zero-days. Malware can take many forms but its purpose is always malicious.
3. Web attacks – almost everything has a public web-based interface which requires a database. There are several attack vectors such as SQL injection or cross-site scripting.

“Offensive cyber capabilities are developing more rapidly than our ability to deal with hostile incidents” <http://www3.weforum.org/docs/WEF_GRR18_Report.pdf>

One such method is Intrusion detection systems (IDS). IDS constantly monitors network traffic scanning for any suspicious activity.

Machine learning comes hand in hand with this as it is able to learn patterns well.

Intrusion detection systems (IDS) stands at a forefront as a defense against attackers.

<https://ieeexplore.ieee.org/document/6560100>

IOT increasing attacks

<https://us.norton.com/internetsecurity-emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html>

<https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/813599/Cyber_Security_Breaches_Survey_2019_-_Main_Report.pdf>

Explain network and types of attack

Explain why machine learning

Testing robustness

to detect/learn different types of network malwares

synthesized attacks will get flagged – bank, non-malicious anomaly

One area known as Intrusion Detection System (IDS) is on the rise to detect anomalous activity real time. However, it faces many challenges. Defining abnormal activity from the norm is problematic.

Aim

Classify different malwares in dataset – to detect/learn malwares

Synthesize network packets using classifier network // Attempt to generate my own attacks and check with classifier

Test available IDS with synthesized malware packets – bank example analogy where person comes in with mask with water gun not really stealing but gets flagged as suspicious.

[1] <https://ieeexplore.ieee.org/document/6169141>

Aims

This project aims to create a system that can detect malware on a network by incorporating machine learning to enforce security on the network. There are multiple steps that are required with possible extension.

Data drives machine learning. Immense amounts of data have been collected, ready to be analysed however it is lacking for cyber security, especially for network analysis. Creating a network dataset that is unbiased and realistic proves to be a difficult challenge [1]. Unfortunately, the initial step would be finding a suitable dataset. Despite the issues with collecting network data, there are still many public resources available.

A note to consider is that I have already chosen a dataset which I will expand more under initial work. This is because the entire project depends on the dataset therefore, I had to choose a dataset before the project started.

There are several objectives that I would like to fulfill with this project.

1. Develop a system capable of classifying different malwares using machine learning.

All IDS systems can detect anomalous activity. This would attempt to create an IDS system using machine learning techniques. The malwares would be anomalies that need to be detected.

1. Create synthetic malware attack patterns.

This depends heavily on the results of step 1. If the system can detect malwares, then it must also learn the pattern of what makes a malware. This would allow me to create synthetic malware attack patterns. Also, able to retrain classifier in step 1 to learn patterns of fake attacks creating a more secure IDS system.

* 1. Replicate environment and create my own attacks

Another possibility would also be to create actual malware attacks, not just synthetic. This would allow me to test the classifier on a live network rather than just on pre-collected data.

1. Test robustness of available IDS with synthetic data.

If synthetic malware data is realistic enough, it would be able to fool IDS systems into thinking an attack has happened. What is the point of this?

* 1. Create genetic adversarial networks to create even more realistic synthetic malware data.

By creating GANNs, more realistic synthetic malware data can be generated. Thus, a loop of constantly testing robustness of IDS in step 3. Again, what is the point of this?

The main aim of this paper is to implement machine learning with network security in order to improve it. Specifically, to detect network malwares. Depending on the possible Multiple steps are required to achieve this with possible extension.

The aim of this paper is to test the robustness of network detection systems by feeding synthesized fake network packets.

In order to do so, a neural network will be trained on amazon’s dataset to create a classifier that can detect certain malwares. Once pattern is learnt, it will attempt to synthesize such patterns which will then be fed into a network detection service. The result should be that the network packets are fake.

Initial work

Four datasets were chosen as candidates however there were many others available at insert link here

1. GT Malware netflow https://www.impactcybertrust.org/dataset\_view?idDataset=1143

This is an ongoing project by Georgia Tech (GT) where they are collecting a daily network feed in an isolated environment. Specific programs are executed for short period of time and recorded.

The issue with this dataset is that