Data Mining in Intrusion Detection

Manh Phung
October 24, 2000

Where are we today in intrusion detection?

In today's world where nearly every company is dependent on the Internet to survive, it is not surprising that the role of network intrusion detection has grown so rapidly. While there may still be some argument as to what is the best way to protect a companies networks (i.e. firewalls, patches, intrusion detection, training, …) it is certain that the intrusion detection system (IDS) will likely maintain an important role in providing for a secure network architecture.

That being said, what does current intrusion detection technology provide us? For the analyst who sits down in front of an IDS, the ideal system would identify all intrusions (or attempted intrusions), and take or recommend the necessary actions to stop an attack.

Unfortunately, the marketplace for IDS is still quite young and a "silver bullet" solution to detect all attacks does not appear to be on the horizon or necessarily even plausible. So what is the "next step", albeit the "next phase" for intrusion detection? A strong case could be made for the use of data mining techniques to improve the current state of intrusion detection.

What is data-mining?

According to R.L. Grossman in "Data Mining: Challenges and Opportunities for Data Mining During the Next Decade", he defines data mining as being "concerned with uncovering patterns, associations, changes, anomalies, and statistically significant structures and events in data." Simply put it is the ability to take data and pull from it patterns or deviations which may not be seen easily to the naked eye. Another term sometimes used is knowledge discovery.

While they will not be discussed in detail in this report, there exist many different types of data mining algorithms to include link analysis, clustering, association, rule abduction, deviation analysis, and sequence analysis.

How do current IDS detect intrusions?

In order for us to determine how data mining can help advance intrusion detection it is important to understand how current IDS work to identify an intrusion. There are two different approaches to intrusion detection: misuse detection and anomaly detection. Misuse detection is the ability to identify intrusions based on a known pattern for the malicious activity. These known patterns are referred to as signatures. The second approach, anomaly detection, is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns. Most, if not all, IDS which can be purchased today are based on misuse detection. Current IDS products come with a large set of signatures which have been identified as unique to a particular vulnerability or exploit. Most IDS vendors also provide regular signature updates in an attempt to keep pace with the rapid appearance of new vulnerabilities and exploits.

Shortfalls with current IDS.

While the ability to develop and use signatures to detect attacks is a useful and viable approach there are shortfalls to only using this approach which should be addressed.

- *Variants*. As stated previously signatures are developed in response to new vulnerabilities or exploits which have been posted or released. Integral to the success of a signature, it must be unique enough to only alert on malicious traffic and rarely on valid network traffic. The difficulty here is that exploit code can often be easily changed. It is not uncommon for an exploit tool to be released and then have its defaults changed shortly thereafter by the hacker community.
- *False positives*. A common complaint is the amount of false positives an IDS will generate. Developing unique signatures is a difficult task and often times the vendors will err on the side of alerting too often rather than not enough. This is analogous to the story of the boy who cried

wolf. It is much more difficult to pick out a valid intrusion attempt if a signature also alerts regularly on valid network activity. A difficult problem that arises from this is how much can be filtered out without potentially missing an attack.

- *False negatives* …detecting attacks for which there are no known signatures. This leads to the other concept of false negatives where an IDS does not generate an alert when an intrusion is actually taking place. Simply put if a signature has not been written for a particular exploit there is an extremely good chance that the IDS will not detect it.
- *Data overload*. Another aspect which does not relate directly to misuse detection but is extremely important is how much data can an analyst effectively an efficiently analyze. That being said the amount of data he/she needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

How can data mining help?

Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. By identifying bounds for valid network activity, data mining will aid an analyst in his/her ability to distinguish attack activity from common everyday traffic on the network.

- *Variants*. Since anomaly detection is not based on pre-defined signatures the concern with variants in the code of an exploit are not as great since we are looking for abnormal activity versus a unique signature. An example might be a Remote Procedure Call (RPC) buffer overflow exploit whose code has been modified slightly to evade an IDS using signatures. With anomaly detection, the activity would be flagged since the destination machine has never seen an RPC connection attempt and the source IP was never seen connecting to the network.
- *False positives*. In regards to false positives there has been some work to determine if data mining can be used to identify recurring sequences of alarms in order to help identify valid network activity which can be filtered out.
- *False negatives* …detecting attacks for which there are no known signatures. By attempting to establish patterns for normal activity and identifying that activity which lies outside identified bounds, attacks for which signatures have not been developed might be detected. An extremely simple example of how this would work would be to take a web server and develop a profile of the network activity seen to and from the system. Let us say the web server is locked down and only connections to ports 80 and 443 are ever seen to the server. Thus, whenever a connection to a port other than 80 or 443 is seen the IDS should identify that as an anomaly. While this example is quite simple this could be extended to profiling not only individual hosts, but entire networks, users, traffic based on days of the week or hours in a day, and the list goes on.
- *Data overload*. The area where data mining is sure to play a vital role is in the area of data reduction. With current data mining algorithms there exists the capability to identify or extract data which is most relevant and provide analysts with different "views" of the data to aid in their analysis.

Difficulties when it comes to data-mining in intrusion detection.

The concept of data mining has been around for years. Despite this data mining in intrusion detection is a relatively new concept. Thus there will likely be obstacles in developing an effective solution. One is the fact that even though the concept of data mining has been around for some time the amount of data to be analyzed and its complexity is increasing dramatically. As stated previously, it is possible for a company to collect millions of records per day which need to be analyzed for malicious activity. With this amount of data to analyze one can guess that data mining will become quite computationally expensive. Unfortunately, for some processing power or memory is not always cheap or available. Of course there may be the argument that you only need samples of the data in order to generate profiles, but there will also be the argument that analyzing anything, especially network traffic, without all the data could lead to false conclusions. Another obstacle will be tailoring data mining algorithms and processes to fit intrusion detection. An effort to identify how the data needs to be looked at in order to provide us with a better picture is surely integral in providing accurate and effective results.

Conclusion.

Obviously data mining and anomaly detection is not a silver bullet for intrusion detection, nor should it be a replacement for misuse detection. The goal should be to effectively integrate anomaly detection and misuse detection to create an IDS which will allow an analyst to more accurately and quickly identify an attack or intrusion on their network.

Bibliography

Bass, Time. "IDS Data Mining." 4 Mar 1999. URL: http://www.silkroad.com/papers/html/ids/node4.html (10 Oct 00).

Gordeev, Mikhail. "Intrusion Detection: Techniques and Approaches." URL: http://www.infosys.tuwien.ac.at/Teaching/Courses/AK2/vor99/t13 (10 Oct 00).

Grossman, R.L. "Data Mining: Challenges and Opportunities for Data Mining During the Next Decade." May 1997. URL: http://www.lac.uic.edu/grossman-v3.htm (10 Oct 00).

Lee, Wenke and Stolfo, Salvatore. "Data Mining Approaches for Intrusion Detection." URL: http://www.cs.columbia.edu/~wenke/papers/usenix/usenix.html (12 Oct 00).

Rothleder, Neal. "Data Mining for Intrusion Detection." The Edge Newsletter. Aug 2000. URL: http://www.mitre.org/pubs/edge/august_00/rothleder.htm (9 Oct 00)