Apply Data analysis in network security

Contents

[1. Set the target 1](#_Toc393104842)

[2. Gather the evidence 1](#_Toc393104843)

[3. Understand the data 2](#_Toc393104844)

[4. Communicate to data 3](#_Toc393104845)

[5. References: 4](#_Toc393104846)

# 1. Set the target

Use data analysis to build up a proactive security solutions. Traditionally, proactive security solutions (Canetti et al., 1997; Barak et al., 1999) are designed to maintain the overall security of a system before any system/components have been compromised by an attack. Here we will discuss PPDM(privacy preserving data mining)and scan detection.

# 2. Gather the evidence

We first plan how and where to collect the original data and prepare data for analysis.

Effective information monitoring builds on data collected from multiple sensors that

generate different kinds of data and are created by many different people for many

different purposes. A sensor can be anything from a network tap to a firewall log; Building up a useful sensor system requires balancing its completeness and its redundancy. A perfect sensor system would be complete while being nonredundant: complete in the sense that every event is meaningfully described, and nonredundant in that the sensors don’t replicate information about events.

No single type of sensor can do everything. Network-based sensors provide extensive

coverage but can be deceived by traffic engineering, can’t describe encrypted traffic, and can only approximate the activity at a host. Host-based sensors provide more extensive and accurate information for phenomena they’re instrumented to describe. Here are several common log formats in network analysis, including HTTP ELF and CLF, SMTP log messages, Microsoft exchange message tracking logs(MTL),syslog.

There are two ways we sample the data:

The objective of using a supervised algorithm is to obtain the highest classification accuracy. The most popular supervised machine-learning methods include artificial neural network (ANN), support vector machine (SVM), decision trees, Bayesian networks (BNs), k-nearest neighbor (KNN), and the hidden Markov model (HMM). Here we will use k-nearest neighbor approach.

The unsupervised machine-learning methods we will use is k-means clustering. Clustering is the assignment of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters. The sameness of the objects is usually determined by the distance between the objects over multiple dimensions of the data set. k-Means clustering partitions the given data points X into k clusters, in which each data point is more similar to its cluster centroid than to the other cluster centroids.**③**

The k-means clustering algorithm generally consists of the steps described as follows:

Step 1. Select the k initial cluster centroids, c1, c2, c3 , ck.

Step 2. Assign each instance x in S to the cluster that has a centroid nearest to x.

Step 3. Recompute each clusters centroid based on which elements are contained in it.

Step 4. Repeat Steps 2 through 3 until convergence is achieved.

# 3. Understand the data

We first discuss scan detection. Strictly speaking, a scan is not a real attack but is the precursor to an attack. Scans find vulnerabilities in cyber infrastructures that they can use to infiltrate systems easily and successfully. Thus, we consider scan detection a preventive process that is different from the classical IDSs that are designed to detect malicious patterns demonstrated during cyber attacks. As the precursors of attacks, scans have noticeable characteristics, and the detections require a variety of techniques to combat the special challenges. Here we show how to use threshold random walk in Horizontal scan detection. Jung et al. (2004) developed a threshold random walk (TRW)**②**, an online detection algorithm, based on sequential hypothesis testing to detect malicious remote scanners while maintaining promptness and high accuracy. This method is based on the observation that benign remote sources have more precise knowledge about the targeted hosts and services than scanners, such that their successful connection rate is higher than the scan rate. They used detection rate TP and FAR in the framework. Here, TP refers to the conditional probability describing the cases in which the connection is really launched by scanners and the sequential hypothesis is correct. FP refers to the conditional probability, which denotes the cases in which benign users launch the connection but the hypothesis is that scanners launched the connection.

The most employed methods for detecting malicious code or behavior use rule-based or statistical models to identify threats in real-time, using adaptive threat detection with temporal data modeling and missing data. One of the limitations underlining most of the rule-based approaches is that they treat each event as a separate activity without considering the context of the events. A rule relies on the signature of a packet based on a set of elements such as protocol. Because a subset of a packet tracking a malicious user and a signature of a normal user may be matched to activate the rule, rule-based misuse detection systems are known for false alarms. In a false alarm, an alarm may be raised after detecting an activity that may be part of an attack, whereas the activity is actually legitimate network traffic. One way to avoid this problem is to perform inference between rules using BN. Using Bayesian statistics to specify the causal relationships between subsets of variables. The naïve Bayes (NB) classifier makes the assumption of class conditional independence. Given a data sample, its features are assumed conditionally independent of each other. This assumption is different from the assumption in BN that dependencies exist between features. In this sense, NB is a special and simple case of BN. In Schultz et al. (2001), naïve Bayes was used to detect new, previously unseen malicious executables accurately and automatically. The method was compared with a traditional signature-based method, and it more than doubled the detection rates for new malicious executables.

Secondly, we will discuss the PPDM(privacy preserving data mining). Preserving privacy is nearly ubiquitous in various informatics disciplines, including but not limited to bioinformatics, homeland security, and financial analysis. It influences cybersecurity significantly with the recent development of information collection and dissemination technologies. The unlimited explosion of new information through the Internet and other media have inaugurated a new era of research where data-mining algorithms should be considered from the viewpoint of privacy preservation, called privacy-preserving data mining (PPDM). The Online Security and Privacy Study of 2009 conducted among 2385 U.S. adults showed a 78% increase from 2007 respondents who choose to log on Internet browsers that protect private information, and a 62% increase in respondents who choose servers that provide built-in security (Online Security and Privacy Study, 2009).\*

The ubiquitous applications of data analysis algorithms allow malicious users to employ data mining to obtain private information and, hence, raises the following questions: will data mining compromise privacy and should data mining be limited in some applications. This concern can be addressed from two aspects: ethical and technological. Legitimate use of private data would benefit the data-mining users and private owners. Many countries have produced regulations and legislation to protect the data owners,control the dissemination of private data, and regulate the accuracy of a database.

From technology perspective, we use PPDM to prevent unauthorized users from accessing private information, such as private data-mining or machine-learning results. Privacy preservation and data mining worked in parallel, until Aggrawal et al. defined the specific research area in data mining concerning privacy protection in 2000. In PPDM, researchers adopt a large number of privacy preservation techniques in data-mining and machine-learning algorithms to preserve knowledge security. The complexity in PPDM algorithms raises several research topics other than privacy preservation and data mining. Verykios et al. (2004a) classified the existing PPDM techniques by considering five views: horizontal or vertical data distribution, data modification methods, data-mining algorithms, rule confusion, and privacy preservation. Most data distributions are horizontal or vertical. We classify the PPDM methods into three groups according to data resource distribution: heuristic privacy preservation, cryptographic privacy preservation, and randomization/perturbation/reconstruction-based privacy preservation. The first group of techniques includes centralized data resources; the second group includes distributed computation, or SMC; and the third group includes both data resources. Centralized data-based PPDM refers to data-mining or machine-learning algorithms that perform applications on the data resources collected at a single central repository. The data storage system can be violated for privacy easily from the inference or learned rules. The proposed solutions employ perturbation-based or blocking-based rule confusion methods to downgrade the rule mining results so that sensitive rules are buried in insensitive rules or insensitive information blurs sensitive data. Heuristic privacy preservation is also called downgrading privacy preservation, because the data modification leads to the downgrading of rules or the ineffectiveness of machine-learning classifiers. Data-mining or machine-learning algorithms must balance between mining or learning efficiency and privacy preserving. We explain evaluation criteria in Section 8.2. Another solution to the centralized data system is to distribute data among multiple repositories and minimize the information leaks by applying SMC. SMC takes advantage of the proliferation of Internet technologies and cooperative computations on private data resources. Researchers have investigated various data-mining algorithms in isolation of each other. Among them, the most important privacy preservation methods have been proposed for a number of data-mining algorithms, like support vector machines (SVM) classification (Yu et al., 2006), association rule mining algorithms (Evfimievski et al., 2004), K-means clustering (Vaidya and Clifton, 2003), decision tree inducers (Agrawal and Srikant, 2000), BN (Wright and Yang, 2004), KNN (Kantarcioglu and Clifton, 2004), ANN (Barni et al., 2006), and other statistical methods (Du et al., 2004).

# 4. Communicate to data

After collecting/analyze the data, we need to transform security to more specific and measurable as an approach to inform the decisions or actions in context. I would like to suggest dashboard with interactive visualizations.

There is an inherent “call to action” nature to dashboards, with each element being either quantitative (has a value) or categorical (a list of items). Most of us have a great deal of readily available quantitative data related to information security ranging from lost assets, to security incidents, to SIEM events-per-second, to firewall/IPS operational data. In order for this data to be useful in the context of a dashboard, these quantitative measures must be able to answer two questions:

● What’s going on?

● So what?

For the categorical measures, you are usually identifying a set of elements that:

● Provide useful information—such as “which incident handlers are primary for the day?”;

● Require the most attention—such as “which Payment Card Industry Data Security Standard (PCI DSS) controls are slipping?”; or

● Need follow up—such as “what are the top expedited firewall port open requests?”

In order to help people to understand and learn from data, effective communication is the primary goal of visual/dashboard creations. Previous rule of developing simple and successful fixed dashboard probably can meet 95 percent of communication request. There are situations when static views are either insufficient or just not practical, requiring the move to a more dynamic medium. Here we will discuss some basic rules to help move from static to interactive.**①**

**Augmentation**—If adding interactive capabilities helps speed up or automate tasks consumers would normally perform manually, going interactive is definitely the right thing to do. There are many repetitive, time-consuming, data-driven tasks in information security. Logs must be collected and correlated, alerts must be received and attended to, and anomalies must be investigated. These actions often involve running a variety of utilities over individual pieces of data or sets of data elements to determine whether there truly is an issue on network. A research team led by Robert Erbacher worked to understand both the problem domain and how incident responders think and process information. This resulted in the creation o f VisAlert(http://digital.cs.usu.edu/~erbacher/publications/visalert.pdf).

**Exploration**—If the number of dimensions and size/diversity of the data set grow sufficiently large, it may be better to enable consumers to explore the relationships and outcomes on their own rather than trying to guess which set of static graphics will be most useful. There is an innovative open source tool released in 2013 by John Goodall called the Nessus Vulnerability Explorer (NV) (http://ornl-sava.github.io/nv/#). NV allows you to take an export from your Nessus scans, drag thefile right into your browser, and begin exploring the vulnerabilities contained within.

**Illumination**—If a topic is complex enough, it may help to provide a well-executed, interactive visualization that provides a user-friendly interface for directed/constrained navigation around the data you’ve chosen to present.

# 5. References:

1. Reference**:** ***Interactive Data Visualization for theWeb***
2. Reference: ***http://www.cert.org/flocon/2008/presentations/flocon08-mchugh-vagi.pdf***
3. Reference: ***Naked statistics: stripping the dread from data***
4. Reference**:**  ***Data Smart: Using Data Science to Transform Information into Insight***
5. Reference**: *Data Mining and Machine Learning in Cybersecurity***