FORCEPOINT

Identifying Insider Threat Through Analysis of Data-at-Rest

Forcepoint and The University of Texas at San Antonio Research

Dalwinderjeet Kular
Audra Simons
01 August 2019
Information classification label: Public

Table of Contents

Introduction

Google can predict our age, gender, and interests based on search history. Netflix recommends the next movie or shows for us, similarly, Amazon recommends items based on our purchasing habits. Data reveals quite a lot about ourselves. Therefore, we wanted to explore whether pseudonymized employee stored data, data-at-rest, could be used to detect malicious or compromised users while preserving privacy. Retailers and social media companies are paying a closer look at their users' data to facilitate their customers' need, similarly, we believe we can use the data stored on our machines to save ourselves from malicious or compromised users as well.

The University of Texas at San Antonio (UTSA) research staff and faculty members Dr. Nicole Beebe, Eric Bachura (Ph.D. student), and Dr. DJ Ko, joined Forcepoint Innovation Labs in this effort as our sponsored research partner. Using the sponsored research agreement, they identified the primary goal of the project as the development of an analytical risk scoring algorithm that leveraged the anonymized data provided to identify the risk represented by each user within an organization. The research agreement emphasized preserving user privacy and anonymity.

The purpose of this research initiative was to develop and deliver analytically, empirically derived algorithms to provide risk profiles of user data storage. Using pseudonymized file storage hierarchical and file metadata, pseudonymized user data, DLP rule violation data, and document classification tagging data, all provided by Forcepoint, along with UTSA derived graph measures and rule violation metrics, we created a risk scoring algorithm to quantify the risk associated with user behavior. This algorithm can be used for both overall risk and per-incident (situational) risk[1]. The overall risk is a user-level consideration considering the nature of incidents that the user is involved with.

Data-at-Rest

Data-at-Rest is the data that resides on hard drives, USBs, laptops or shared drives. This data doesn't move between devices or networks frequently. This data is not accessed or modified regularly. Data-at-rest is very valuable. It is valuable because it is well organized, properly labeled (labeled according to the data in the file), contains information that has a higher retention period, such as employees' social security number (SSN), bank account and routing numbers, customers contract, or intellectual property.

Data-at-rest is ***highly valuable*** for attackers. Attackers can gain inside knowledge (SSN, bank accounts, intellectual property) and exploit it to their advantage. This makes it more attractive than data obtained by sniffing every single packet over the network to steal information (it's like asking a food lover *"what will you prefer a pie vs a slice of a pie?"*). Every company takes appropriate measures to keep data-at-rest safe and secure by enabling firewalls, antivirus programs, encrypting hard drives, etc. (but there are some slip-ups too).

Is Data-at-Rest only valuable for attackers? Can we use this data to identify our shortcomings, such as data over sharing, insider threat, or detect and stop a masquerade[2]attempt? Analysis of data-at-rest can be highly beneficial for any company to protect themselves from insiders as well as compromised users. In this article, we discuss the value of data-at-rest in the detection of malicious or compromised users and with examples of how that detection would appear.

Usefulness of Data-at-Rest

In our opinion, data-at-rest can tell us ***"what a user does"*** in a company, i.e., a user is in software development or an HR employee. For example, we would assume that on an engineer's hard-drive there will be a higher number of source code and design document files. Similarly, on an HR employee's hard

[1] An incident is a file with one or more DLP rule violations.
[2] Masquerade: pretending to be someone else.

drive there will more files on employees' bonuses and hiring. Therefore, we can infer, if a user has a vast majority of source code files then they are part of an engineering team.

Do all the users in the engineering department have the same files? No. We believe based on a user's job role the file collection will change. Also, research has shown that different users have different collections of documents on their machines. While some of these differences are natural and innocuous (for example, different roles require different content), some are more revealing (for example, a sales representative who has detailed customer histories for thousands of customers).

Can stored files and their storage patterns help in identifying a malicious user? Yes. In some cases, this data aggregation can be a precursor to data theft, either by the employee or by an attacker who has assumed the identity of a particular employee.
For example,

- ***Leaver or disgruntled employee***
    - It may be very normal for an HR employee to contain multiple documents on their system classified as resumes, but it may be an indicator of 'risk of employment termination' for a resume document to be found on an engineering employee's system

- ***Data over-sharing or data theft***
    - It may be normal for HR employee to have files containing employees SSN and banking number, but it may be a precursor of "Data Theft" for this file to be found on marketing employee's system.

Detecting a malicious or a compromised user

In this section, we have discussed data, features and the method we used to detect malicious or a compromised user.

Data

For this project, we collected stored file data for multiple users in multiple departments. These files were run through the Forcepoint Data Loss Protection (DLP) product. This provided us with the DLP classifiers that fired on the files. Along with the classifier we used the following file data:

- Files
    - types
    - path
    - timestamps

- DLP

- DLP classifiers fired on collected files
    - DLP classifier e.g.,
        - Proprietary and Confidential Footer
        - Self CV/Resume Distribution

- Document Classification
    - File contextual information
        - Tag: Tag represents file features such as kind of file (Type), retention period (Retention), data importance (Critical), etc.
        - Tag Value: File feature values, for example, a tag “Type” can have the following values
            - a contract file, a resume, etc.

Features

UTSA used the above data provided by the Forcepoint to extract the following features:

- Number of DLP classifiers
- Number of unique DLP classifiers
- Tag ratios
- Tag value ratios
- File system graph metrics
- Group detection metrics

Method

We built a user’s data storage risk profile using the features mentioned above and compared it against its peers. The proposed method is as follow:

- Convert file system topology into a graph
- Compute graph characteristics
    - Local graph properties
    - Individual node and edge values
    - Connected components values
    - Community values
- Map graph data to DLP and document classification data
- Calculate group membership based on graph
    - Use community and connected components values
- Compute risk score at the user level
- Aggregate user risk scores
    - Summation
    - Average

- The merged rank of summation and average score rankings
- Resulting risk scores can then be used to inform security decisions

The rationale is that users in the same department with similar job roles should have similar DLP classifiers firing on files, file types, etc., therefore, their data storage risk profiles should be similar to one another. Likewise, the functional nature of department delineation should result in data storage profiles that are similar to one another. Graph measures provide an opportunity to quantify the characteristics of file system profiles that often become obvious visually. These quantified measures can then be used in an automated risk scoring approach that can help to identify high risk user profiles as well as potential masquerading attempts. This rationale served as the basis for the development of a risk score.

Results

The resulting algorithm was able to detect multiple masqueraders. To check the creditability of the algorithm Forcepoint manipulated the data. Forcepoint swapped several users data between multiple departments and added additional folders from a different department to various users, to simulate masquerading behavior in the data. This was done without the knowledge of our academic research partners, to determine if the existing approach would detect these users as high risk. The results indicated that this was the case. Initially, the academic research partners thought it was an issue with the data preparation and collection process, due to the high scores and misidentified departments for the swapped users. Forcepoint then notified the academic partners of the true nature of the swapped users, indicating that the risk score had successfully identified and properly scored the data. A visual representation of what was expected versus what was detected can be found in the figures below.

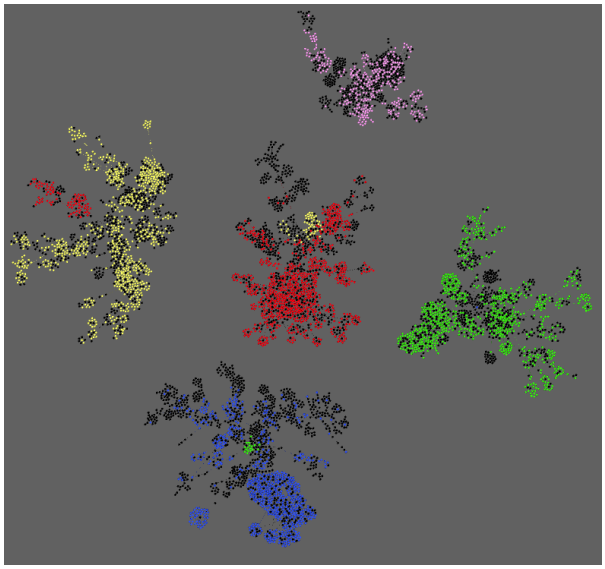

Figure 1. Visualization of assigned groups.

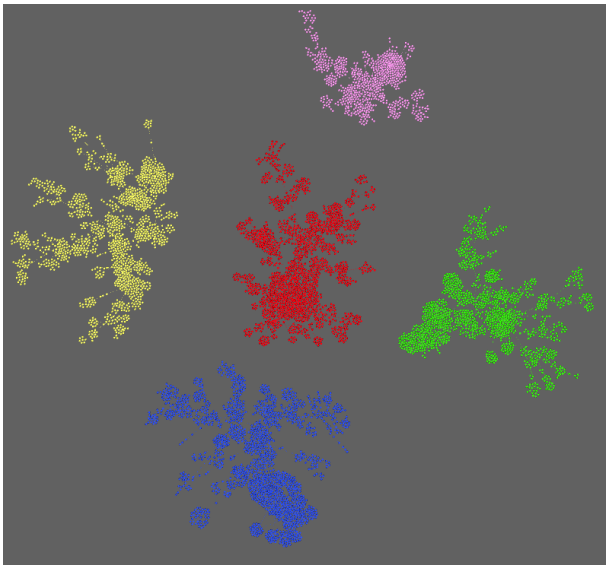

Figure 2. Visualization of detected groups.

In both figures, the data represent five departments, each color-coded. Figure 1 has colored the data according to the department that was indicated in the dataset by Forcepoint. Figure 2 has colored the

data according to the graph metrics. Note that figure one has color-coded some data black to indicate that it is a container that has no user or department information associated with it due to the lack of DLP related data. A manual review of these two graphs indicates that there are users swapped between groups. The quantification of this information via graph measures and probability values derived from those measures is what facilitated and informed the risk score that identified the masquerading users. This highlights the key outcome of this research partnership: the development of a risk scoring algorithm that utilizes easily collected data to measure and identify risk represented by user activity on the computer systems within an organization while preserving user privacy.

About the Project and Privacy

This report presents the insider threat detection project undertaken in collaboration and coordination among Forcepoint personnel and UTSA research staff. We have taken extra precautions while sharing the data because privacy and protecting the identity of an individual is a high priority for us. Forcepoint has provided UTSA researchers with DLP classifiers fired on internal users files, and the pseudonymized and salted hash of file names, file paths. Thus, we believe there is no way that leakage of this data could allow for any identification of individuals or insight into the actual content of files. Even if UTSA were to lose this information, it would not allow an attacker to learn anything of use about the users involved, including their identities or the contents of their files.

Conclusion

Data-at-rest is one of the key elements for detecting masquerading or data theft attempts. For example, uploading an unusual amount of data to an unusual location or unusual stored data files as compared to peers. In this project, Forcepoint and UTSA collaborated to detect malicious or compromised users using data-at-rest. For detection, graph measures were used. These measures were able to detect multiple malicious or compromised users.

With this project, we are proposing to use analysis of data-at-rest to shield ourselves and our customers from data misuse, data theft, and masquerading.