Privacy Preserved Data Mining for the Norton Core

Yining Wang, Richard Messina, Satish Agrawal, Peter Holoubek, Sahil Verma

Abstract - Many companies nowadays mine the data of their users. They accomplish this by collecting data and applying learning algorithms to this data in order to extract information. The value of these learnings is used to improve products, increase the quality of the user experience, and automate/optimize points of pain. However, storing large sets of personal user data adds additional risk for the user. What if this data leaks? What if an internal employee uses this data with malicious intent. In this paper, we explore several options for how this data mining can be performed without adding additional privacy risks for the user. We analyze such methods in the context of the Norton Core router.

Keywords - Norton Core, Internet of Things, Security and Privacy, Threats, Security Layer, Privacy Preserved Data Mining (PPDM)

I. Introduction

In the recent years, the size of the internet has been growing exponentially, along with the amount of data that is generated by users. Many companies are hungry for this data because there are so many uses for the information which can be learned from this data. For example, companies can study trends in user history and determine what their needs are and improve. One major use of data right now is for the advertising industry. Targeted ads can be created for users to induce a higher selling rate. Another major use for data is in Machine Learning. By using a large amount of data, researchers can design algorithms and solution models to address and fix existing problems or creating more complex ones.

As a result, data is becoming more valuable and competitive to companies. Companies relentlessly acquire data by data mining, data sharing, or data purchasing. For example, Norton's security-enabled router, the Norton Core, is used to collect user information. This information is supposedly used for security purposes as well as improvements to the Core. However, this brings up a major concern to the public and consumers: how is our privacy being protected? In the recent scandals with Facebook and other companies collecting users' personal data and selling them, the European Union is placing harsh restrictions on data transfers. Even without companies purposely trading or selling data, there can be data leaks, and the aftermath of such leaks is quite dangerous.

Regardless of the situation, collecting and storing user's personal data can always pose a threat to our security and privacy. Therefore, this kind of data stored should be minimized when possible. At the same time, data is necessary for technological breakthroughs and improvements. So, in this paper, we are going to explore the idea of Privacy Preserved Data Mining (PPDM) – acquiring data and at the same time protecting users' privacy.

II. Background

Intensive data mining poses threats to Users' privacy and security, because data may contain sensitive information. As a result, the need to Privacy Preserving Data Mining is on the rise. The Norton Core will be used as a primary example on to data mine while preserving privacy.

A simple method for attempting to preserve the privacy of users while collecting their data is to anonymize this data. This would involve removing any personally identifiable information from the data collected. Sadly, this is not a fool-proof mechanism. Even without the name, username, or email of a user, it may still be possible to figure out what data belongs to who given a leaked set of data. In this paper we will be exploring more secure methods to apply in addition to anonymization in an effort to best protect the privacy of users whose data is collected.

III. Threat Model

The Norton Core is responsible for handling the entire home network and as a result, over a period of time, it will gather a lot of data. This data comprises of user passwords, device usage per device, user preferences, contacts, shipping information, and so on. To understand the privacy implications of data mining results, we first need to

understand how data mining results can be used (and misused).

As stated previously, all this data stored in the Norton Core is a culmination of data generated by a number of devices and users, therefore, it makes it even more crucial to ensure that it is privacy preserved. If the data that Norton collects is compromised, credit fraud, account hijacking, and even cyberbullying would be possible.

Below is an attack tree (Figure 1) along with a table (Table 1) showing how our proposed solution could minify the risks of such attacks.

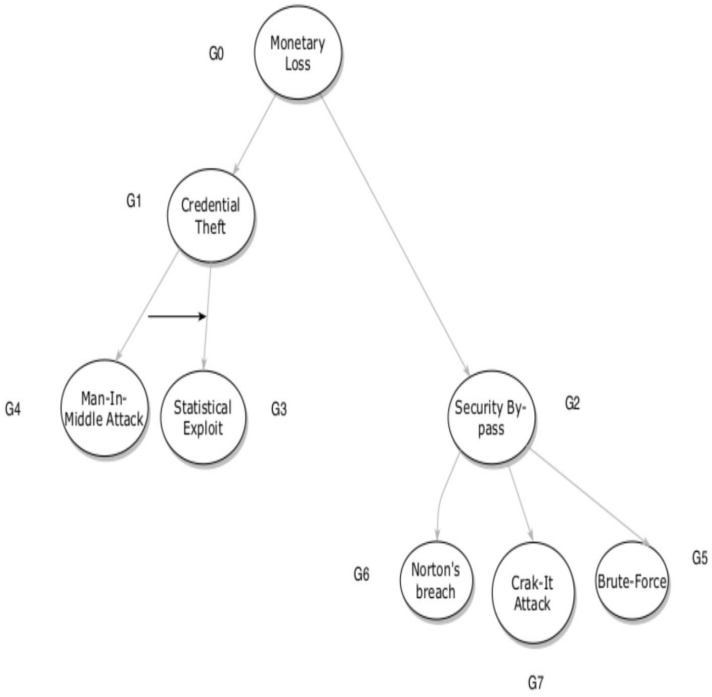

Figure 1. Attack Tree

Threat	**Solution Proposed by Designers**
Credential theft	Anonymization: remove personal identifiable data to create a privacy preserving layer. This is done using 2 techniques: Condensation and Randomization
Statistical exploit	The solution provided is effective in handling this because we will first generate statistically equivalent small clusters of data which share aggregate distributions.
Security Bypass (Norton Breach/ Brute-force)	Proposed solution describes a technique of randomization which changes the original values of the data in a way that the numerical information computed remains consistent, therefore making brute force attacks ineffective.

Table 1. Threat Analysis

IV. Design and Implementation

The goal of privacy preserved data mining is to define methods for extracting data from users and learn from this data without introducing additional risks for the users. Ideally, it would be as if the users' data was never collected. For the case of the Norton Core, we explored methods for transforming user data into anonymized but still usable shapes. The anonymization goes far beyond simply removing personally identifiable information. This is because even with the removal of a user's name and device ID, it may still be possible for attackers to trace back certain data points to specific users using their network history, for example.

We will discuss two different techniques to apply on top of anonymization in order to accomplish the goal of privacy preserved data mining. These techniques include the following:

1. Condensation [8]
2. Randomization [7][8][9]

These techniques will be applied during a Privacy Preserving Layer which ensures that the majority of data that Norton handles is effectively anonymized. However, in order to generate the most reliable anonymous data, the condensation technique needs to be applied to large sets of raw data. Statistical approaches can ensure that the original distribution of data is preserved. Thus, anonymized raw data will still need to be stored for some period of time. The amount of time that it is stored for before being run through the Privacy Preserving Layer is variable and can be decided by weighing the benefits and risks. A short amount of time (1 month) would ensure that potentially traceable information on users is removed quickly, but the distribution of the data may have a large number of outliers and may not be the best. A long amount of time (1 year) would keep potentially traceable information on users stored

for a more dangerous amount of time but the data learnings garnered would be based on a more accurate distribution with relatively less noise.

The architecture of a system which includes a Privacy Preserving Layer would likely appear as shown in Figure 2.

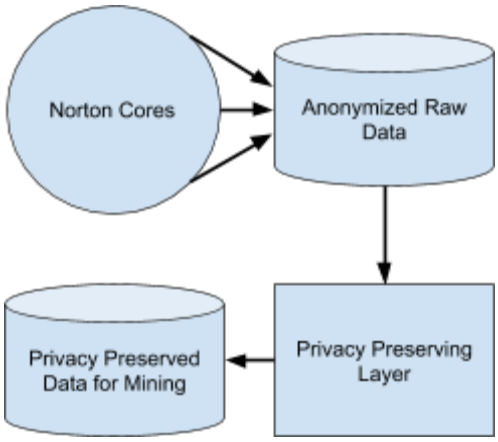

Figure 2. PPDM Architecture

It is important to note that this architecture applies to Privacy Preserving Layers which utilize the condensation technique. If only the randomization technique is used, the Anonymized Raw Data does not need to be stored and the data sent from the Norton Cores can be immediately run through the Privacy Preserving Layer [8].

V. Implementation

The first step in implementing this design will be to remove personally identifiable information from the data that is collected. For the case of the Norton Core, this would involve removing device names, usernames, Norton accounts, contact information, and shipping information from the list of items that Norton collects through the core [4].

Next, the Privacy Preserving Layer needs to be defined. There are a few techniques which can be used during this layer, as previously mentioned. For the condensation technique, this layer would process a set of raw anonymous data for long-term use in data mining practices. Condensation can be performed on this data set in a way that produces a dataset ready for PPDM which has a nearly identical distribution.

Condensation describes the practice of generating statistically equivalent clusters of data to work off of [8]. It is a simple but effective idea. Basically, sets of smaller size are created based on sets of raw data which share aggregate distributions. For the Norton Core, this could be done by taking the mean of certain network traffic statistics. If 100 users spend an average of x number of seconds on a page, then this mean could be recorded and treated as a single user.

Another technique which could be used instead of condensation is randomization. This is a type of perturbation which could be applied immediately, therefore bypassing the need to store Anonymized Raw Data, which condensation requires. The goal of randomization is to change the original values of the data in a way that the numerical information computed remains consistent. This is normally accomplished by adding noise from a known distribution. This way, the original distribution may be reconstructed [6].

Figure 3. Randomization and Reconstruction

One simple method of randomization would be additive perturbation. Let X be the original data distribution and Y be a defined noise distribution. Therefore, if $Z = X + Y$, then Z is the result of the randomization. In order to operate correctly on this randomized data, the original distribution must be reconstructed. For our current example, this can be accomplished with the following equation: $X = Z - Y$ [6]. Note that we are talking about *distributions* here. The Z distribution in the reconstruction equation would only be an estimate, and thus the original values should be impossible to retrieve.

Once again, for the Norton Core, this randomization could be applied to the network traffic statistics. A constant perturbation can be applied to the page visits being counted, for example, and the original distribution can eventually be reconstructed using the inverse process.

VI. Evaluation

Like Conventional Data Mining (CDM) approaches, PPDM faces the same requirements when it comes to the efficient Performance. Unlike CDM, PPDM has to also provide sufficient Data Utility and minimal Potential Privacy Loss. [10] [11]

When we look into Data Utility and Potential Privacy Loss we face a trade-off. One extreme is to fully minimize Privacy Loss by non-deterministic obfuscation. Even though an adversary won't be able to learn information, except by an always present coincidence, neither will the intended party.

On the other end, we can maximize Data-utility by choosing CDM technique and achieve maximal correctness of Information. [10]

One school of thoughts on the Privacy vs. Utility trade-off, claims that "even modest privacy gains require almost complete destruction of the data-mining utility." [12] Simplified reasoning which concludes the above-mentioned statement is as follows:

Utility gain is obtained when the data is used towards a good cause, such as proper public policy. Privacy loss, on the other hand, is incurred when an adversary learns a subset of identifying data which can map to a specific entity, such as a person's insurance learning medical record. It follows, that on average, any gain obtained by a good cause will be offset privacy loss. [12]

The shortcoming of the above reasoning is the symmetric gain/loss of the trade-off. Intuitively, learning incorrect data for a good cause might be fatal as much as a privacy loss. However, when the PPDM technique is exploited and an adversary learns even incomplete or incorrect data, privacy loss was still incurred to the user. It follows that due to asymmetric nature, comparing averages is an inadequate measure. [10] [11]

An alternative view is to measure worst case privacy loss vs. aggregate utility. Furthermore, the Utility vs. Privacy Loss should be measured against the trivially-anonymized data and utility should be measured against the original data as the baseline.[10] The latter point stems from the asymmetric nature of Privacy Loss vs. Utility. More specifically, to gain utility from the dataset, we need only correct information but false or correct information obtained by adversary incurs Privacy Loss. [10]

Where does Norton Stand?

As we mentioned previously, Norton Core fails several principles by obtaining client's data. The use of these data is intended to provide a greater utility to its client due to dynamic security updates. On the other hand, Norton poses risk to the community by misusing the data.

Given the sensitive nature of information passing through Norton Core, the company builds immense data wealth. In the corporate world of profit-seeking, this could potentially create a dilemma for the company and ultimate mistrust for the community.

The critical nature of both utility and privacy, due to enhancing cyber-security mechanisms and the sensitive content of the data, demands low Privacy Loss and high Data utility. Yet, it's important to realize that Norton Core exists in order to secure our household from cyber threat. It implies that Data Utility for a greater good is somewhat more important than Privacy Loss.

With the above-stated reasoning, we will briefly evaluate two proposed PPDM techniques: Randomization and Condensation.

Randomization - The biggest drawback of randomization is the potential weakness in outlier data. The solution to remedy for this exposure is to inject useless noise. This in return reduces somewhat data-utility because dataset after reconstruction might yield a different distribution. [8]

The advantage of this technique is its simplicity and very efficient and high threshold against Privacy-Loss. [8] From the technical standpoint the device doesn't require knowing the distribution of the data. Hence it performs faster than other PPDM techniques and can be even performed in run-time. Lastly, Randomization doesn't require trusted server where anonymization is applied.

Condensation - Condensation is highly recommended whenever the concern is streamed data. The technique uses clusters in order to create pseudo-data. The pseudo-data will preserve the original data format but won't reveal any defining data. This in result achieves minimum Privacy Loss.

Unfortunately, Condensation lacks the performance of Randomization. Since the pseudo-data is released in blocks, there might be time gaps which might affect the performance.

Also, the very fact that we use pseudo-data affects substantially data-utility. It's hard for the

intended party to learn specific information other than distributions of the dataset.

VII. Related Solutions

A. *Cryptography Based PPDM*

Cryptographic techniques are ideally meant for distributed computing scenarios where multiple parties collaborate to compute results or share non-sensitive mining results and thereby avoiding disclosure of sensitive information. Cryptographic techniques offer a well-defined model for privacy that includes methods for proving and quantifying it. The data may be distributed among different collaborators vertically or horizontally. All these methods are almost based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining consists of a set of secure sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product. Although cryptographic techniques ensure that the transformed data is exact and secure, but this approach fails to deliver when more than a few parties are involved.

B. Secure Multi-Party Computation

In this approach every part of private data is validly known to one or more parties. The problem arises when the private information is revealed to some other third parties. It uses specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data, participate in a protocol that generates the data mining results, that guarantees no data items is revealed to other parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy.

C. High Utility Sequential Patterns Mining

Sequential pattern hiding method is necessary to conceal sensitive patterns that can otherwise be extracted from published data, without seriously affecting the data and the non-sensitive interesting patterns. High Utility Sequential Patterns mining discovers all sequential patterns in sequence database whose utility values are equal to or greater than a given minimum utility. The expansion algorithm of USpan is used to obtain all HUSPs and collect the input data for proposed hiding algorithms. Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

VIII. Conclusions

Information privacy is particularly critical with ubiquitous information systems capable of gathering data from several sources, therefore raising privacy concerns with respect to the disclosure of such data. Norton Core should implement the PPDM approach discussed above to mitigate its privacy concerns. Most of the privacy attacks can be efficiently prevented by either condensation and or randomization techniques. Still, we face a trade-off between Data Utility and Potential Privacy Loss, but that is the case with every technique. Different algorithms may perform better than other whens when looking at certain criteria. Norton Core has posed its users with unneeded security and privacy risks which could be minimized.

REFERENCES

[1]https://web.stanford.edu/group/mmds/slides/mcsherry-mmds.pdf

[2]http://didawiki.cli.di.unipi.it/lib/exe/fetch.php/dm/ppdm.08.05.08.pdf

[3]https://www.symantec.com/privacy/norton-privacy-english

[4]https://arxiv.org/pdf/1804.04250.pdf

[5]https://ieeexplore.ieee.org/document/7950921

[6]https://ijcsmc.com/docs/papers/October2013/V2I10201344.pdf

[7]https://pdfs.semanticscholar.org/2802/2aac5e21fdb56469ba9bb3b1218aaa7481ab.pdf

[8]https://pdfs.semanticscholar.org/48a7/42320cdf046020d7f7dfd9000714cce57223.pdf

[9]https://www.cs.purdue.edu/homes/ninghui/papers/privacy_utility_kdd09.pdf

[10]http://www.stat.cmu.edu/~jiashun/Research/Year/KDD04.pdf

[11]https://www.cs.cornell.edu/~shmat/shmat_kdd08.pdf