INSE 6180 - DATA PRIVACY AND WEB PRIVACY TECHNIQUES

CHARLES AUGUSTIN MOSES 40084105

JAYAPRAKASH KUMAR 40083709

MANUSHREE MALLARAJU 40082236

ABSTRACT

The developing prevalence and improvement of data mining advancements carry genuine danger to the privacy of person's, private information. A rising examination subject in data mining, known as protection over querying databases, has been widely concentrated lately. The fundamental thought of privacy preservation is to alter the information in such a manner to perform data mining calculations adequately without compromising the privacy of sensitive data contained in the data set. For each user, we talk about their privacy concerns and the strategies that can be embraced to ensure delicate data sharing between different private data owners. By separating the obligations of various private data owners concerning privacy, we might want to give some helpful bits of knowledge into the investigation of preserving privacy in a large scale databases. In a general scenario, When an user wants to query the Netflix database to find the most favourite movie. The query sent to the Netflix database through Netflix server and retrieves the exact result. In the world of internet, data transmission between account holders and database are highly prone to attacks. To avoid malware attacks and unwanted intrusion into the database, we proposed different techniques that maintains data privacy, web privacy, and prevent malware attacks.

***K*eywords** Data Pseudonymization · Curator · Private Data Owner · Randomization · AES · JWT · CORS

1 Introduction

Database enables the data to be more secured and provide efficient access. Data mining is the process of analyzing data from a different perspective and summarizing it into the useful information. Data mining is a major step towards discovery of information from the database.

Privacy preservation of data mining has become progressively popular, as it allows sharing of private sensitive data for analysis purposes. Some privacy preserving data mining algorithms are developed to extract required data from huge database, where the sensitive data or information is hidden or not disclosed. There are several key definitions of web privacy and data privacy they are curator, private data owner and queries to database. The private data owners queries the database to get the data set based on their own privacy access level. Some sensitive attributes such as username, password, email address, other credentials, etc, should be hidden from the privacy intruders. Moreover there exists a large number of collection of information on vulnerabilities focusing on various subjects, for instance on different web platforms and databases. Parts of these information are made available as traditional databases. This is the motivation behind why the exploration on the reasons and properties of vulnerabilities on data set is of prime importance.

2 Background

2.1 Data Privacy

Data privacy is related to how the information are handled securely in the real world. At the most elevated level, protection is the privilege of an individual to be disregarded, or opportunity from obstruction or interruption. Information protection is the privilege of a individuals to have command over how close to personal data is gathered and utilized. Information security is a subset of protection. In data mining the primary goal of privacy protection is revises the original data and develops the appropriate data mining algorithms to protect the data from attackers. Privacy data protection mainly consists of two aspects. First, sensitive information like ID card number, name, address, etc is not revealed during the process of data application. The next is how to make these data enterprise level data applications

2.2 Web Privacy

Data mining has evolved its methods and techniques in maintaining sensitive data in terms of data analysis, validation, and publishing. This will have a high impact on preserving the privacy information on the web application. The intention behind the existing privacy is preserving data mining methods. The current privacy preserving data mining techniques are clustering, association rule, outsourced data mining and k-anonymity.Also the data owners send the query through the HTTP request, it has to face many malware and intruders in between. To safe guard the query and the owners must be authenticated and the requested query must be encrypted by appropriate algorithm.

2.3 Naive-Bayes Algorithm

A Naive Bayes classifier is a simple technique of machine learning models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The crux of the classifier is based on the Bayes theorem. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A) * \mathbb{P}(B|A)}{\mathbb{P}(B)}$$

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. The technology is more strengthened to process huge amounts of data. Data Mining acts as a bridge between extracting valuable information from raw data and which will lead to new methods and techniques. Most of the applications of Naive Bayes Classifiers are used in recommendation systems, spam filtering, sentiment analysis, etc. Although they are fast and independent, the predictors used should be independent, and ultimately this will help us to achieve a good accuracy guarantee on the classification problems, especially for our data set.

3 Overview

In our system, we run the Naive-Bayes classification on amazon data set and then we push the labeled result set into the database. By using this classifier the restricted domain access of users (email id field) is handled. In our data set, the users with gmail, yahoo and outlook domains are only considered as authorized review. The curator gives the result set for the query sent by the private data owners of the database. In our system, we have two data owners namely, reviewer, admin and curator of our database is Express JS server.

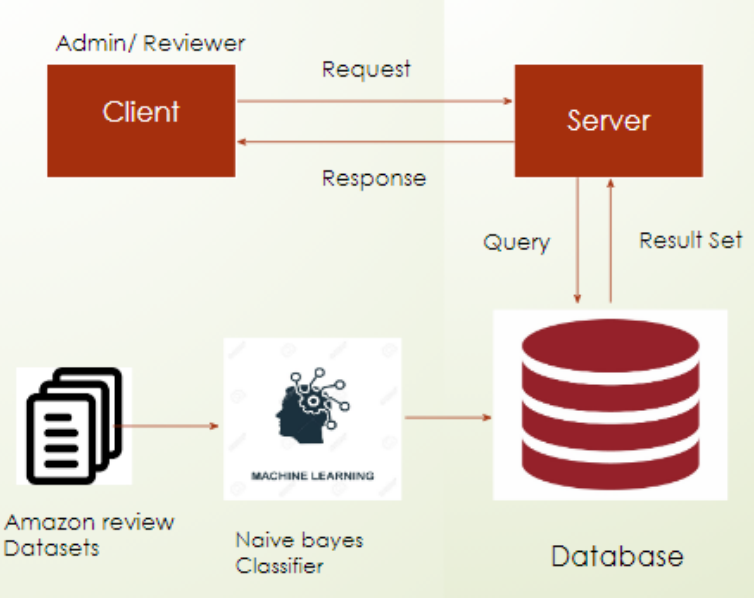

Figure 1: System Architecture

4 Data set

For our study, we took the Amazon review as our data set. It has 4 fields like username, email ID, password and reviews. We took 900 data sets and classified them into positive reviews and negative reviews by a Naive-Bayes classifier. We choose this data set because it has more sensitive information on it. The sensitive information is highly prone to privacy attacks and also highly vulnerable. We discussed some mechanisms in detail to handle the data vulnerability and privacy in below sections.

Label	username	Password	email	Reviews
__label__2	JIOlDrsw	XPWYpvaSKH	JIOlDrsw@outlook.com	Great CD: My lovely Pat has one of the GREAT voices of her
__label__2	uYVbFSAG	SjwQdZXGeR	uYVbFSAG@yahoo.com	One of the best game music soundtracks - for a game I didn'
__label__1	rFTgytbn	CZYQCLXxkd	rFTgytbn@outlook.com	Batteries died within a year ...: I bought this charger in Jul 2

Figure 2: Amazon review data set

5 Factors Influencing the data set vulnerability

5.1 Excessive Privilege

When an admin gives excessive access to the user or accidentally misuses the access, system faces privilege abuse and escalation. It may happen unintentionally or out of negligence but it causes a major privacy leakage in the database. In privilege account abuse, when an authorized user used the account inappropriately or fraudulently, maliciously or accidentally, or through willful ignorance of policy whereas in privilege escalation the attackers took advantage of vulnerabilities in database management systems. In our application, we grant proper access to our private data owners.For instance, Admin has view access to all the data set and reviewer has only to it's own data.

5.2 Unmanaged Sensitive Information

The sensitive information like password, credit card number, social insurance number, username, etc must be handled securely. If we do not secure this information then it has a high chance for a privacy leakage. In our application, we are processing data pseudonymization techniques over sensitive data.

5.3 Database Injection Attacks

In SQL injection, let us consider an unauthorised query inserting into a database or malicious statements into the input field of a web application. In both the cases an input injection attack can give an attacker unrestricted access to an entire database. It ends up with the wrong result for the requested query. It makes a data set more vulnerable. It can be avoided by allowing authorized users, using appropriate privileges to data owners, proper web security. In our system we build a proper authentication system for the users to send a query to the database and continuously monitor the current status of data owners using JWT tokens.

5.4 Malware

Unauthorised user or attackers send the query through HTTP request. It attacks the database and makes the data set vulnerable. It often occurs in the web application if the web requests are not authorised properly. Our proposed system encrypts the HTTP request used to send the query.

6 Common Attacks

6.1 Linkage Attack

Linkage attacks combine the data with another data set and tries to re-analyze individuals in an anonymised data set. These 'linking' includes indirect identifiers also called quasi-identifiers. Linkage attacks are more powerful attacks and identify the data set information by the collection of multiple source. We can handle this attack by randomization, centralization of data. In our proposed system, we use data pseudonymization to defend against linkage attack.

6.2 Web Privacy Attacks

Cookies: A cookie is a well known mechanism in the application.It allows the server to store some private information about a user on his/her own computer.The main use of the cookies is to speed up the transaction between the server and client, monitor the user identity. Unfortunately, due to insufficient security measures the cookies can be accessed by unauthorized person. An attacker can examine a cookie to determine its purpose and retrieve the user information from the website that sent the cookie.The very famous attack is Cross-site scripting. In our application we handled two levels of privacy by using JWT tokens and AES algorithm.Since the data transmission is the encrypted by adding 256-bit string, the privacy intruders don't get any information when they attack.

Cache: In computing, a cache is a small computer component that stores data so that future requests for that data can be served faster.But the time of data stored in cache can be attacked by the attackers.To avoid this issue, we are saving the data in encrypted way by AES algorithm.

7 Privacy Mechanisms

7.1 Data Pseudonymization

Before explaining the pseudonymization, let us discuss the difference between anonymisation and pseudonymization where the former irreversibly destroys the data subject odf data set and latter replaces the data subject's identity in such a way that further information is required to re-identify the data subject. Data Pseudonymization is a privacy and de-identification technique in which the data records are replaced by one or more randomly generated artificial strings(Tokenization). This randomized string makes the record less identifiable but at the same time it serves its purpose for data analysis and processing. In other words, Pseudonymization method substitutes the identifiable record with reversible and consistent information

7.2 Tokenization

It is a technique of adding noise to the original data record. It involves mapping original data to a token requires methods that make tokens unfeasible to reverse in the absence of the tokenization system such as JWT tokens (upto 255 characters in length). The tokens are usually generated by the curator. Upon authorization, data will be provided to respective data owners. For generating the tokens we used the AES algorithm.
AES uses symmetric key randomized encryption technique, which involves the use of only one secret key to cipher and decipher information. AES-256, which has a key length of 256 bits, supports the largest bit size and is practically unbreakable by brute force based on current computing power. This algorithm allows curators to preserve sensitive information in a data set from privacy intruders.

Figure 3: Pseudonymization / Tokenization

User Name	Tokenization/ Pseudonymization data	Anonymized Data
Thomas	U2FsdGVkX19cG/UuhRq1X/8h7wjDSas1JUGfSfAZ22s=	xxxxxxx
Chris Evans	U2FsdGVkX1+G3mvjhg7eRbPeNaQg5WqX5c6w3d2dQ7M=	xxxxxxxxxxx
Tony Stark	U2FsdGVkX19YMnfRrlLg2oI3rjHjs+L1oFy8RK3L4sg=	xxxxxxxxxx
Natasha	U2FsdGVkX1/2Mtbi0s8GzzUZ9gteYfsefgSrxelrrHI=	xxxxxxxx

7.3 Privacy preserving querying using added noise

The main objective of querying is to extract information from the database for analysis. As we know, a large number of queries to the database using laplace mechanism kills privacy. To overcome this problem, we used tuple based labelling, when querying the data set, the labels of corresponding user names will be used instead of original user names. This makes our mechanism (epsilon, 0) differentially private.Example shown in the table 2.

Figure 4: Privacy preserved querying

Labelled Username	Original Name
U2FsdGVkX19cG/UuhRq1X/8h7wjDSas1JUGfSfAZ22s=	Thomas
U2FsdGVkX1/2Mtbi0s8GzzUZ9gteYfsefgSrxelrrHI=	Natasha
U2FsdGVkX19YMnfRrlLg2oI3rjHjs+L1oFy8RK3L4sg=	Tony Stark

7.4 Fine grained access control

7.4.1 Access privacy : Tokens

Access privacy is a mechanism through which we can provide access to private data owners for accessing sensitive information in the dataset. In our project we introduce JWT tokens to achieve the access privacy. Here, the curator generates a randomized token when the data owner wants to access the dataset for the first time. Post verification, the owners can access the data using existing tokens. If the token is kept valid for a long period of time, there is a possibility for a privacy attack. Hence, it is important to limit the validity time of the generated token and we kept the token valid for a period of 1 hour and expires after the stipulated time. More importantly, JWT token adds a public key to the token which makes it nearly impossible for privacy invaders to manipulate the token.

Example Token

eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJfaWQiOiI1ZWUyZTY4MDIyYzQzNDYyNzJmMWI2YTciLCJ
pYXQiOjE1OTI0OTgwMDMsImV4cCI6MTU5MjUwMTYwM30.hQo8sSO03MZbOn741eAraFwOr7rbUSCe
kc7YxuedzzM

Figure 5: Access privacy token

7.4.2 Passport-hashed privacy

Before authenticating the private data owners, the private information such as username and password must not be available for querying other than private data owners. In order to achieve this property, we are not storing the original password directly. However, the curator/ server converts the password and stores it as hash and salt value. Password will not be stored in the database, it will be constructed from the hash and salt value for authenticating the private data owners.

Example of database record with hash and salt value is shown below:

Labelled Username	Password	Hashed value	Salted value
YBLvUudL	MfMUdq	d40ccafd20cfd3ec9f92651f8d349ca139e1a216fce85a21ebebfd1fc5cddcb0f8e1509ab69ea37576ccbc4a992a3ef5b00e5d948330d65da1e59117977cc318887546586 45bc4bfa092e356913282e263bbc91c66a2da4b04bfef89158fd781bdf3dba1ee04ccc9ddb94293c520f78d69d93da4a71aaaeb24606a1713c03bd88a56d1557416181f3308ceff3d	b891896edc7474be637fff44c4c442a95a17445252b8a5f0bb60179efebc5128
bJTeoUgm	tCXjpoD	cafb0a16ee01c8885f97f7069a3258ffd3154c1ce7b7100b87f93a375f2c268e1778c82699f82b734b1ceebc7b0843e8ce1cd847a0520ce817ad84ce7d29eff04d0cb30c6f2ee67b00e9d3fb4209fd919bf437f55f43a9b7ff3790e008877836c683f8d55e7d5a2cdcd1a16bf893c0c235e77e7a75c8a138f25948029e6d6db17ac5a1d51794fe23ea00f	8b609ab269ca22868621d32bbff334c924c8523daa0d200d1f8ebb004f592002

Figure 6: Passport Hashing Mechanism

7.4.3 Blocking Malicious Querying

When users other than private data owners tries to query the database, they will get nothing but useless information. Here, We are adding a noise to the result of the query which ensures that the provided data cannot be comprehensible by data intruder. Below figure shows the noise added data for un-authorized users.

Figure 7: Query result for non-private data owners

But when private data owners tries to access the database, they will get the original data set without any added noise. This is taken care of by data pseudonymization and tokenization mechanisms.

7.4.4 Web Privacy

Same Origin Policy The same-root strategy is exceptionally prohibitive. Under this strategy, an archive (i.e., like a site page) facilitated on server A can just communicate with different reports that are likewise on server A. So, the equivalent cause strategy that allows collaboration with one another that has a similar inception. Since this policy is too

strict, new policy evolved in the mean time to allow resource sharing only with authorized domains. This paved the way for origin of CORS.

CORS: Cross Origin Resource Sharing CORS is a mechanism that allows only a set of whitelisted domains to access or share the data. Here the web servers must implement some methods to handle requests, outside its origin. We implemented the same in our project to restrict set of white listed domains as shown below,

```
const whitelist = ['http://localhost:3000', 'https://localhost:3443', 'http://localhost:4200'];

var corsOptionDelegate = (req, cb) => {
    var corsOptions;

    if(whitelist.indexOf(req.header('Origin')) !== -1){
        corsOptions = {origin: true};
    }
    else {
        corsOptions = {origin: false};
    }
    cb(null, corsOptions);
};
```

Figure 8: Cross Origin Resource sharing

7.5 General Release Mechanism

This mechanism restricts the number of possible queries. This mechanism answers a normalized counting query $Q(x)$ by returning a (random) smaller synthetic version x' of x. The query is then answered by computing $Q(x')$. If x is a database with over one million rows, and the true answer is $Q(x)$, then the mechanism allows us to compute $Q(x')$, which hopefully gives a good estimate of $Q(x)$.

$$Q(x) = \frac{1}{|x|} \sum_{x_i \in x} 1_{[g(x_i)=1]}.$$

In the below figure, we can see the table of amazon review data set. Here, we use a query input field to get the subquery of dataset by applying query value on the large dataset fetched from the database. The label field with the red indicator denotes negative review and green indicator denotes positive review which is classified using Naive-Bayes algorithm discussed before and the status field indicates authorized or malicious depending upon review submitted by users from authorized domain or not.

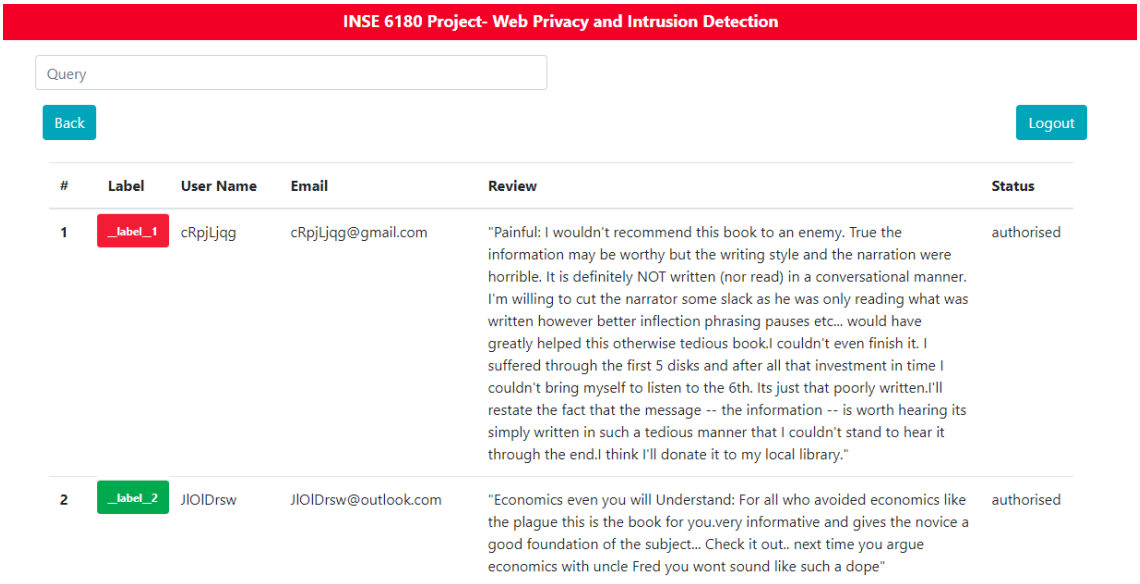

Figure 9: General Release Mechanism

8 Conclusion

In the advancement of modern browser and complexity in the web, data privacy becomes a difficult goal to achieve. In this report, we summarize the general privacy problems in the data privacy world and proposed the different techniques to protect the data from privacy attacks. We also discussed some common attacks in the database and highlighted the most favourable solution by using data pseudonymization techniques. Moreover, the hazards of saving the private information in cookies and caches. The usage of AES algorithm helps us to prevent leakage of confidential information from web cookies. To achieve privacy and performance guarantee, we implemented the general release mechanism which retrieves a small subset of the entire data set so that, the curator does not need to send the data again and again. In future, we would perform a case study on different privacy mechanisms to mitigate the privacy leakage in large datasets.

References

[1] Protecting Browser State from Web Privacy Attacks https://www.abortz.net/papers/sameorigin.pdf

[2] Data Mining for Malicious Code Detection and Security Applications https://ieeexplore.ieee.org/document/6061180

[3] Modern Intrusion Detection, Data Mining, and Degrees of Attack Guilt https://link.springer.com/chapter/10.1007/978-1-4615-0953-0_1

[4] Preventative measures for SQL injections. https://www.esecurityplanet.com/threats/how-to-prevent-sql-injection-attacks.html

[5] VulinOSS: A Dataset of Security Vulnerabilities in Open-source Systems https://antonisgkortzis.github.io/files/GMS_MSR_18.pdf

[6] Pseudonymization vs. Anonymization https://www.protegrity.com/blog/pseudonymization-vs-anonymization-help-gdpr

[7] Cache Attacks and Countermeasures: the Case of AES https://www.cs.tau.ac.il/~tromer/papers/cache.pdf

[8] Privacy-Preserving Database Systems http://people.cs.aau.dk/~simas/teaching/privacy/275agh936b763r6r.pdf

[9] https://www.trustwave.com/en-us/resources/blogs/trustwave-blog/the-3-biggest-database-threats-and-what-your-security-plan-should-look-like/

[10] Database security attacks; http://index-of.es/z0ro-Repository-3/Top_Ten_Database_Threats.pdf

[11] JSON web tokens; https://en.wikipedia.org/wiki/JSON_Web_Token

[12] https://www.securetrust.com/challenges/by-mandate/data-privacy/?gclid=Cj0KCQjwoaz3BRDnARIsAF1RfLeQ6qBpXfBldZyZcpcGyJqe1ibRXkUGYxR-c7OrFrIJWsqAa39-HW8aAubgEALw_wcB