International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 08, August 2016
www.ijates.com
ijates
ISSN 2348 - 7550

AN ITERATIVE APPROACH TO EXPLORATORY THE USEFULNESS OF DATA SANITIZATION

P.Ugamani[1], Dr.M.Mohammed Ismail[2], P. Rizwan Ahmed[3]

[1]*Research Scholar, Computer Science, Mazharul Uloom College, Ambur, Tamil Nadu, (India)*

[2]*Associate Professor & Head, Pg & Research Department Of Computer Science*

[3]*Assistant Professor & Head, Department Of Computer Application (Ug & Pg)*

ABSTRACT

When data is shared and/or published, the need for revealing data must be balanced with the need for sanitizing it. This is because some information considered "sensitive", if revealed may cause damaging consequences, for example, privacy violations, legal and financial liabilities, embarrassment, national security risks, and loss of reputation. Although many techniques for sanitizing data have been developed and used over the years, attackers have still managed to de-sanitize data. One of the reasons for this problem is the tremendous growth of publicly available information. Data like telephone numbers, date of birth, movie ratings, personal preferences like favorite movies and favorite food recipes, property records, real-time geolocation information through social media content and photo metadata can now be easily found on the Internet. This has enabled attackers to gather an immense amount of information about a user or a group of users and correlate it with sanitized datasets. Such correlations can lead to many methods of inferring sensitive information. In this dissertation, we show a method by which data can be evaluated to see if it can be sanitized effectively, while maintaining needed utility from the data, and if so, how can it be done optimally.

Keywords: Data sanitization, data mining

I. THE DATA SANITIZATION PROBLEM

Data sanitization is the process of adding, modifying, and/or removing information from a set of data that contains sensitive information, which enables that data to be used for analysis while attempting to maintain user privacy. The typical goal of data sanitization is to conceal some aspect of the dataset, for example, personally identifiable information, while still enabling the data to be useful in some way. In the rest of this dissertation, we will refer to these two interconnected goals as "data privacy" and "data utility." However, these goals can often turn out to have fundamental conflicts. This is because while data privacy aims at concealing information, data utility requires revealing it. The degree of privacy and utility is governed by policies that are defined by stakeholders, who have both privacy and utility requirements that must be fulfilled. However, sometimes it may be impossible to satisfy these requirements.

Another problem that remains with data sanitization is the presence of information external to the dataset. This may provide opportunities for inferring something about the sanitized data because of some connections or patterns in the external data that can also be found within the sanitized data. Therefore, while evaluating whether

particular data should be released or not, one must consider data within the dataset as well as the information which may exist outside this dataset. Although many techniques can be used to attempt to solve these key problems in data sanitization most of these techniques share, some common drawbacks. First, many sanitization techniques are highly focused in specific domains and are generally not applicable to other types of data. Hence, extrapolation to a general model of data sanitization is cumbersome and in some cases not possible, Due to domain-specific assumptions. For example, consider a dataset that contains names of movies, their corresponding ratings and the times when these ratings were made by a set of users. Since this dataset has no user names or pseudonyms, there are no personally identifying attributes that can identify a user. A dataset like this may still be vulnerable to inference attacks that could expose information that was intended to remain hidden, if similar data is found in publicly available movie rating websites and correlated with the given dataset. So a sanitizer might add noise (in the form of fake ratings) to make these correlations nebulous. However, the same technique of adding noise to sanitize medical records will not work, as it will fail to comply with HIPAA, which requires deleting personally identifying information.

Second, many techniques restrict their analysis only to data within a dataset in order to sanitize it. This is referred to as a closed world assumption. The problem lies in the fact that there exists information outside this dataset, which can help an adversary to infer data that was supposedly hidden within the dataset. It may be impossible to determine who an attacker may be, let alone estimate how much of this outside information is available to him/her. Therefore, while sanitizing data, we must assume an "open world" scenario, wherein any information can be used by any attacker to de-sanitize a dataset with some hidden data.

And finally, almost all techniques consider data sanitization as a "yes" or a "no" problem. But if this was the case, then sanitized datasets like the Netflix Prize dataset and the AOL Query dataset would never have been breached for privacy violations. In fact, it is rare for analysts to estimate risks and vulnerabilities that may arise in sanitized datasets. One of the reasons behind this is the lack of a comprehensive process to sanitize data. For example, if a privacy policy fails to capture the privacy requirements of stakeholders, then a vulnerability in the sanitized dataset may arise. But such comprehensive assessments are seldom performed. Moreover, one cannot make assumptions as to what information may or may not be present external to the dataset, as this may lead to vulnerabilities that can be exploited by an adversary.

II. PRIVACY AND ITS CHALLENGES?

We define privacy as the ability of an entity to control information about itself. Most commonly, this entity is a person, business organization or government. Each entity may have a different set of requirements regarding the disclosure of their information. For example, people may choose to give their information to a social network, if it can guarantee control over the disclosure of this information in a way that is acceptable by its users. Such requirements guide privacy policies that need to be precise, comprehensive, and universal. However, cultural and legal differences can make this a challenging problem, as policies can be interpreted differently in different regions. For example in Europe, the EU Data Protection Directive has given the citizens a "right to be forgotten". Under this right, the citizens can request removal of any information from their past that is "no longer needed for any

legitimate purpose". However, implementing such a directive on a search engine basedin United States might result in a clash of policies. Also, there may be insufficient technical tools to implement the policies. For example, in the above example, it can be very challenging to implement the "right to be forgotten" directive on search engines, whose underlying idea is to remember and search through all past history on the internet.

III. DATA SANITIZATION

The process of data sanitization involves removing or modifying parts of a dataset which could reveal sensitive information or if disclosed together, a subset of parts which could reveal sensitive information. It should be mentioned that the term privacy can be given different meanings within a single policy. For example, consider the following datasets: D_1 which consists of usernames with their corresponding salaries, and D_2 which consists of usernames with their corresponding diseases. The privacy policy for D_1 could say that it is sufficient sanitization to change salaries to broad ranges rather than exact numbers. Alternatively, the requirement for sanitizing D_2 may be to replace usernames with pseudorandom, unique numbers while keeping the disease names listed. Therefore "removing sensitive data to protect privacy" does not always mean the same thing and can vary highly depending on context.

We can formalize the above discussion in the following way. Consider a dataset D and

let D' be a subset of D, such that D' contains sensitive information. The policy: P , is a function of P_p and P_u, where P_p is the privacy policy and P_u is the utility policy, such that:

$$D \times P \rightarrow D \setminus D'$$

IV. THE COMPLEXITY IN DATA SANITIZATION

There are many aspects to the problem of data sanitization; the data and how it can be interpreted, the various policies and how they can be interpreted, and the information that can be derived from the data with or without using the external information. The way data and policies are interpreted is important, because different interpretations can lead to different solutions to the problem. But these interpretations depend upon the assumptions that are made while analyzing the problem. The fundamental problem here lies in how to determine whether each of those assumptions is correct or not.

When considering assumptions relating to the data itself, adding data fields will cause an increase in the number of values, which adds more complexity when analyzing the relationships among them. This again presents the attacker with more information that can be correlated with externally available information, thereby, making inferencing easier. If the goal is to hide some sensitive information, it will generally involve some loss of the usefulness of data, as sensitive values must be hidden. But consequently, if there is a goal of having data to analyze it, then some values (which may or may not be sensitive) may need to be revealed. This requires a tradeoff, and resolving the conflict can be highly complex.

V. CONCLUSION

In this article we have proposed an iterative model for effectively sanitizing data, which uses relationship analysis and helps predict what data and relationships, if present external to the dataset, may help an adversary in

de-sanitizing the sanitized dataset. We have also shown how data and relationships can be formally and graphically represented to allow analysis using different properties and algorithmic methods.

Data sanitization is the problem of removing sensitive information while retaining its statistical integrity to comply With an analysis requirement. Fundamentally, some information must be removed from a dataset to ensure non-disclosure of sensitive information. This is governed by a privacy policy. Correspondingly, some information within the same dataset must be retained to ensure analytical requirements that are governed by an analysis policy.

Overall, there has to be a balance between privacy and utility. If we do not anonymize data appropriately, crucial data and relationships may be revealed in the sanitized dataset that have the potential to be correlated with information present in the external world. This inference is what leads to revealing sensitive information. On the contrary, if we over-sanitize the data, analytically useful statistical relationships and correlations within the dataset are destroyed, which may provide little or no utility value to the dataset. This leads to a question is data sanitization that enables both privacy and utility possible?

REFERENCES

1. ArgoUML. http://argouml.tigris.org/.
2. California Civil Code Section 1747.08(b). http://www.leginfo.ca.gov/cgi-bin/ displaycode?section=civ&group=01001-02000&file=1747-1748.95.
3. Chapter 93 Section 105 MA General Laws. https://malegislature.gov/Laws/
4. GeneralLaws/PartI/TitleXV/Chapter93/Section105.
5. Internet Movie Database (IMDb). http://www.imdb.com/.
6. Massachusetts' Weld Collapses at Commencement. http://articles.latimes.com/
7. 1996-05-19/news/mn-5935_1_undergraduate-commencement.
8. Netflix Prize. http://www.netflixprize.com.
9. OWL - web ontology language overview. http://www.w3.org/TR/owl-features/. Privacy: Internet: minors. Senate bill no. 568. chapter 336. http://leginfo.
10. legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201320140SB568. RDF Schema. http://www.w3.org/TR/REC-rdf-syntax/.
11. Yolo County Public Employee Salaries and Benefits Information.
12. http://www.yolocounty.org/index.aspx?page=364.
13. A Firm Foundation for Private Data Analysis, volume 54, New York, NY, USA, January 2011. ACM.