CPSC 4660 – Database Management Systems Fall 2015 Course Project

Research Project

0011164024

Research Project

Security in Data Mining

By: Vusi Ka-Caleni

Data mining as a field looks at large data sets and attempts to extract information from the data sets, and use them in an interpretable way for future use. Some problems that arise when using data mining is the compromising of security and sensitive information contained in the data. Over the span of this research paper we will take an in-depth analysis of the many different aspects of security and data mining. Beginning with a look at PPDM, privacy preserving data mining which is supposed to reduce the risk brought by using data mining operations[1]. Through this protocol we will be able to look more closely at the attributes of data mining which are affected. With data mining we have many different levels in which data can be accumulated and analyzed for example the data provider, collector miner and decision maker all play vital roles in protecting, changing and analyzing information. With these different layers and approaches to protecting privacy in data mining we will be looking at, at what point does data mining violate privacy. For example when we obtain valid results from private input data, then only the results of the data is displayed, but is the results themselves an invasion of privacy and what risks are there for the data to be breached and accessed by persons who should not have access to the said information. Also we will be briefly looking at Data mining and its uses for cyber security and national security, as many of the articles only give vague information of the approaches used in regards to data mining for governmental purposes. This is to be expected as the algorithms and methodologies used would be a national security hazard if released to the general population. None the less we will briefly look at ways in which data mining plays a role in detecting and auditing, instances of viruses, Trojan horses or any other methodologies in which a person could attempt to break into secure networks and access classified or private information. Another way in which we will be looking at how data mining effects security, is using the techniques to mask data, so using data mining to hide information rather than discover trends and patterns. This aspect of data mining could also be valuable in the matters of security, for example if one wanted information not to be located by data mining techniques then they would need to understand how these techniques worked and how to avoid them, thus masking the data. In relation to masking to data in order to protect our privacy, randomization is an efficient approach to do so. Thus we will look at how effective this approach is to protecting our data and basic framework used to randomize data.

To begin, we must first comprehend that over time there has been a growth in interesting privacy-preserving data [1] as recorded information has exponential grown over the past decades due to memory becoming less expensive and society becoming more technologically inclined. Privacy-preserving entails adding noise onto an attribute, this can affect the data in many ways, but there are many more effective ways of protecting a user’s data in order to preserve security. The first method and perhaps the simplest is by allowing the user to remain anonymous, as introduced Murats article [1], if we introduced a medical diagnostic test that was available to the general public to determine the likelihood of a person contracting a disease, then for obvious the user of the publicly available program may want to keep their own identity hidden from the mass population. Due to the fact that if they are more likely to contract a certain disease employers may be less likely to hire the person, or insurance costs for health maintenance may increase due to the user being a greater risk from the general population. If however the program only asked the user questions such as ancestor death causes, diet, daily exercise and other generic non-person specific questions then we could see how the collecting of this data could indeed keep refrain from identifying any one person. However an insure could improve on this data by implementing ‘guesses’ [1] for certain unknown variables and grouping trends of known variables to certain geographical locations, for example if McDonalds was only located on the Westside of Lethbridge and the user inputted that their diet consisted of McDonalds four times a week ad through this program we saw that people who ate McDonald’s four times a week contracted diabetes then the insurance company could use such information to see which users where located on the west side of Lethbridge, and eating McDonalds four times a week, and thus charge them more for health insurance. The way in which the program would be improvised as discussed in the paper would be by introducing public data and sensitive data, in which the sensitive data would be hidden, thus any independent company or persons would gain no new information implementing any new techniques to analyze the data. Then we arrive at the next problem of protecting the user’s information, which would be the case of publishing the collected data of users in general. So the insurance company could attempt to look at the total sum of the data and distinguish which information is the sensitive data, in order make their predictions. This however would very difficult due to the fact that although the company is able to access the final results it would not be able to see the relationships between the data. Through this example we have demonstrated a basic way in which we can protect user information from certain security risks from the data mining model. By identifying what information is sensitive to whom is it sensitive too and what else can be determined from the information.

­­­

We can now begin to see how access to data models proves to be an important factor in much of a user’s security being at risk, however the prevention of basic privacy loss in many cases is not absolute. Even by implementing the methods described above of making certain attributes hidden we cannot always guarantee that information will remain private by relying on the above method alone. We can thus classify privacy loss using the Bayesian classification error, for a given set of data. Within this data we can look at the summation of the data in relation to the classified data, this data is within the bounds of the protected information, but a problem arises when certain attributes that may not be deemed as private at the time are outside of these protected bounds, as the data collector determines what is believed to be sensitive information, one my overlook a certain attribute of a data sets as personal data meant to be kept private thus allow for the data to be exposed and become a security risk for the user. The user may then have the data protected but due to the mismanagement of data classification the privacy is compromised. There are four basic ways in which the data may be compromised, when a classifier produces sensitive data when given access to public information, the classifier taking public and unknown data, the classifier taking public and sensitive data, and lastly when an adversary has access to the sensitive data of the users. With these we can begin to see the potential for the complications that arise when looking at the security risks in data mining.

With these security risks we can look at more in depth models and algorithms that are used to address these issues in relation to protecting the user’s data, from the many different ways in which data of a user may become compromised through the use of data mining. The first approach we will look at will be the randomization of data in order to aid in the preservation of user security. Randomization itself is a very economical approach to preserving privacy in data mining [2]. With the use of randomization we must understand the basic framework in which it is implemented. Assuming we are using the server-client model, then random variables distributed over the data are within our set, thus we have sets of data including the original data and those containing random variables, or the randomized data. Now the strength of the this “noise” depends on the method we apply to data which in this case is the mixture model, using this model we can have different types for the data sets, for example one may be discrete while the other continuous[2]. How much data we have to examine effects the density of the sets, and how we distribute over set with the noise. Now when looking at this method of protecting data ways in which we could analyze our ability to protect the data would be to examine our ability to translate the data with noise back into the original set, using the nonparametric MLE is one way in which we could do this, or even applying kernel Estimators. Though we must acknowledge that more randomized the data becomes less we can understand the information and the more degraded it becomes. So by implementing randomization into or data we can indeed maintain privacy preserving data mining but we lose information accuracy. From the first approach we demonstrated how implementing more simplistic methods have limitations on the ability to maintain the data security, now with the use of a more complex methods we see other problems start to arise, in reference to the degrading of the information. At this point we can perhaps begin to realize that each approach for preserving security in data mining has its own drawbacks.

Now when examining the security implications from data mining it is truly important to not to forgot the process that goes into the basic collection of data. Firstly let us examine the different types of users that could access the information [5]. The data provider, which would be the user who owns the data that is to be collected. The data collector, this entails the user who collects the data which is obviously provided from the previous user. Then we have the data miner who performs the task of mining through the data provided by the collector and lastly the decision maker. This user perhaps has the most influence in affecting privacy issues of the data as this user makes decisions on the data based on the data mining results. Now looking at each user on an individual basis in regards to the data being analyzed we can further understand the security risks that each one contributes towards and is effected by. In regards to the data provider, the major concern would be attempted to control the sensitive information that might be given, we discussed this earlier with the approach to make certain information private or inaccessible. For the data collector major a concern would be that the personal information of the user would be attached to the data which would violate the providers privacy, so data modification could be implemented here in order better to protect information, which would again touch on what we have discussed in reference to the randomization of data sets in order to better protect the privacy of the provider. This would have to be implemented in a way which does not cause harmful degrading of the information. Next the data miner, whose task pertains to applying data mining algorithms to extract useful information from the information provided in the data. This user must still take into account PPDM( privacy preserving data mining) to ensure that user data is protect from third parties who may also access the data. Lastly we can look at the decision maker, who takes the mining results directly from the data miner or from some information transmitter. This user’s task pertains to interpreting the information to determine if the data that has been mined is credible.

Looking closer at the data provider we have seen how limiting the access can indeed aid in the protecting of sensitive information, but we must also consider the trade off, of privacy for the users benefit. Certain websites that force users to register before being able to access any content on the webpage. In this case the data provider could simply provide false information in order to maintain their privacy such methods that could be used in regards to use on the internet would be. Using ‘sock puppets’ to hide a user’s true activity, as sock puppets create a false online identity which participants of an internet community speak while pretending to be someone else, so by implementing this a data provider would be interpreted as a different user, so their own unique activities online would not necessarily be attributed to them. Another basic method in which the data provider could maintain their security of data would be to use a fake identity to create false information, an example of this would be Apple Inc. implementing Techniques to pollute electronic profiling “which can help to protect user’s privacy. This patent discloses a method for polluting the information gathered by ‘‘network eavesdroppers’’ by making a false online identity of a principal agent, e.g. a service subscriber. The clone identity automatically carries out numerous online actions which are quite different from a user’s true activities. When a network eavesdropper collects the data of a user who is utilizing this method, the eavesdropper will be interfered by the massive data created by the clone identity. Real information about of the user is buried under the manufactured phony information” ([5]). Lastly we can look at the method of using security to mask ones identity, so when the user signs up for a web service or purchases an item online more often than not the user most provide an email address, credit card information, postal code, etc. In order to protect this data a user could download a browser extension take MaskeMe for example, this extension gives the user manageable aliases (masks) which provided the false information needed, to protect ones privacy.

Granted these methods exist and are readily available for use by the basic user, we must acknowledge that in most cases users will not implement these methods to protect their data. Due to a lack of knowledge or simple lack of understanding many of the security risks that data mining can pose on an individual’s privacy. In today’s world many people use the internet to access social networks, so we have to ability to examine how many social networks being the data collectors, implement methods to inure user security and the maintenance of user privacy. As social networks better their ability to discover social patterns amongst users, in order to allow users to locate their peers on their websites/applications, network analysis becomes more important. In many cases the analysis of the network must be sent to a third party. So once again we have the issue of maintaining security for the user, even if the social network removes the identities from the published data, the published data could still lead to the exposing of sensitive information as we have discussed above. As social networks are typically molded as graphs, preserving privacy would entail anonymizing graph data, which can become very difficult as anonymizing graph data requires a much different approach then table data [5]. We can break this security problem into three categories, firstly modeling the adversary’s background knowledge about the network. Second modeling the information loss in the anonymizing social network data, hence the degrading of the information. Lastly we consider that anonymizing data for a social network is different than that of relational data as, applying anonymity to a group of tuples in a relational table doesn’t necessarily have an effect on other tuples whereas anonymizing graph data, or changing the information on one edge could affect the entire network. To deal with the issues outlined above in relation to the privacy implications of graphs we will need to understand how to approach edges in the graphs, namely edge randomization, modification, and clustering-based generalization.

Namely when an adversary wishes to de-anonymize the edges of a graph they use their background knowledge to this by learning about the relationship between the de-anonymized data and the user. This would an attack model approach, were the background attributes like the attributes of vertices, vertex degrees, linked relationships, neighbors, subgraphs and graph metrics can all be of a security issue. Take for example the widely known algorithm of seed-and-grow, which implements these attributes to identify users of a graph which have been made anonymous. By taking a seed sub-graph which could be caused by the collision of a small group of users and then based on this irregularity which identifies users, then grows the identifying property out to the whole graph. This attack on user security implements the cumulative degrees of n-hop neighbors of a vertexes the local feature and then combines this with a simulated annealing-based graph matching to makes anonymous edges no longer anonymous. Another attack to identify users in a social network, would be the mutual friend attack which works by identifying an anonymous user based on already identifying their friends. Also we have the degree attack, where data of the user is examined as belonging to a community and not just existing as an individual edge, so the attack examines the edges information determines what attributes can associated with the community and from there identifies the induvial user. From these examples we can clearly see how a third party could use certain algorithms to identify the users of a social network, even after having their data anonymized. This is a very important factor that the data collector has to account for if they wish to provide the proper safety tool for protecting the in security and privacy of the users who have provided their information.

We will now look at and examine the privacy model for protecting privacy of data in graphs. As we now know that many complications arise in the social networks, such as making information anonymous without degrading the data, which proves to be very difficult as each edge in the graph is dependent on the other and also third parties being able to de-anonymize the edges based on their vertices and binding attributes. Such models include the implementation of k-NMF, which looks to prevent aspects of the attack model for identifying the user based on their data [5]. Basically for this approach we stop the algorithms from identifying the user by insuring that the graph data with ‘e’ edges have at least k – 1 other edges of the same connections then it’s can be guaranteed by implementing the algorithm that the probability of the edge being identified by those implementing the attack models as described above would be no greater than 1/k, where we have k-degrees, k-neighbors, k-automorphism, k-isomorphism, and k-symmetry. Thus the greater the graph is in size the more the probability of using the basic attack model identifying the user dwindles.

Now in regards to graphs we examine the utility implications by maintaining the privacy of our data user. Many different methods can be implemented such as looking at the topological properties which is a method of examining the graph at a deep level, spectral properties which aims to examine the eigenvalues of the graphs adjacency matrix and other derived matrix’s, and the aggregate network queries, which calculates the aggregate on some paths or subgraphs that satisfy the query conditions. These methods help us to understand functionality that may have been lost by implementing the k-anonymization algorithms we discussed above in regards to K—NMF. So these Algorithms for network data publishing perform edge insertion and/or deletion operations, and they try to reduce the utility loss by minimizing the changes on the graph degree sequence. Which we can then examine using the topological, aggregate and spectral properties to gage if indeed the graph’s information has been degenerated.

Throughout this paper we have demonstrated the many different security risks that are posed by data mining, up until this point we have spoken about those who wish to collect such data as commercial companies perhaps simply wanting to gain monetary use out of identifying interesting traits from the data sets, but this is a rather shallow approach to the implications of data mining. When wanting to understand more of the gravity that lies within privacy in data mining, we can look towards the cyber security applications [6]. In which governments use data mining to unusual behaviors and patterns online in order to determine possible cyber-attacks. By implementing link analysis governments could trace back many viruses to those individuals responsible of creating or unleashing the viruses on to the interweb. This would of course by many be considered a good example of the applications of compromising security of the user, as the user in this case attempts to harm those around her. Also we may use classification to group various cyborg attacks into communities and then use these communities to help determine where the cyber-attacks originate and also to predicted where future cyber-attacks could possibly occur from. With this outlook on the security implications of data mining we see the positive in being able to identify the user by analyzing their data provided, whereas in contrast to before we saw the importance in maintaining their security. This is a spot where we must understand the tradeoffs that occur in data mining, as many of the techniques that are used to identify these online attacks, could then be applied for the commercial purposes as we have discussed above, in reference to insurance companies attempting to identify users at high risk of certain diseases and social networks releasing sensitive data of their users. As cyber-attacks cost corporations billions of dollars [7]. We must not ignore the fact that preventing these measures by implementing the algorithms developed by governments would be a cost effective method in stopping cyber-attacks. For example when governments analyze e-mails and phone calls of users to determine variances in the conversations of those people deemed to be threats to the state. Could be applied in a verity of ways to interpret variances user e-mails and phone calls for commercial purposes.

We must now understand that in data mining no one algorithm has the ability to fix all the potential security risks or compromise all of the security a user possesses, or even to better manage data as a whole due to the difficulty of fitting data types to certain algorithms [3]. But rather when approaching the issues in data mining we attempt to use the methods that would best protect the data that your user interacts with. Take for example Snap Chat and Google, with both deal with sensitive information in regards to their users. How they approach maintaining user privacy would be very different. As when using data mining techniques on user data in snapchat one would be access photos while with google it would information relating to search queries. Implementing anonymity in searches as discussed previously could add a level of protection for users who search on google, however implementing anonymity with Snap Chat would definitely not have the same security benefits as users generally transmit photos of themselves, and of places where they live/travel. In that example we clearly demonstrate the necessity of interpreting data uniquely from each other. Also as google implements graph searches and we have seen the complexity related to attempting to anonymize edges on a graph, they would again have a variance in the methodology of maintaining privacy of the data. This is not to say that we cannot have generalized approaches to data. As generalized approaches to data often require us to do some type of “transformation” [3], this transformation can be seen as the pre-processing of the data mining. Such as Knowledge discovery in databases(KDD), with this we have a stark difference between other pre-processing applications, such as the missing value handling, because for the missing value handling applications the data is considered to be independent, which means they can be done without knowing what data mining algorithm technique is being used. To use the generalized method we must consider the data types in the databases and their relations to the many different data mining techniques. In regards to the different data mining techniques we have four typical forms of data mining, textual data forms which reads the texts from documents, temporal data forms which looks at time-series data stored in temporal data forms, transaction data forms which looks at transactions of items exchanged, and relational data forms which are the most come to be used and looks at many different data that are stored in data tables. Also for generalized approaches we must look at the six kinds of widely used data mining techniques, such as multilevel data generalization, which is to observe the data stored in databases. Second we have mining association rules which looks at associations between items. Third we have data classification, which looks at classifying data sets, fourth we have clustering analysis, which amaryllises the similarities of patterns. Fifth we have pattern based which is similar to the clustering analysis as it aims to search for patterns in the database and lastly we have the discovery and analysis of time series trends which aims to find tends through periods of time. All of these approaches come together to allow us to make a generalized approach to given data sets.

Thus, throughout this paper we have examined the many different aspects of security and privacy of users, which refers to user anonymity in the mining of data, and also insuring this data does not have the ability to be de-anonymized by third parties gaining access to the given data sets. The tradeoffs of degrading the information that is caused by the very action of attempting to prevent third parties from being to de-anonymize the user’s information, which imposes its own complications to the management of data. We have also looked at the complications that arise from different types of data, like graph data in which attempting to change the information of one node can affect the entire graph and so preserving the information becomes more difficult while trying to maintain the user’s security. We have looked at more general approaches to attempting to protect the user’s data by implementing methods to translate the data set into more tangible sets that can be examined by the generalized approaches. Overall when attempting to protect a user’s security, instantly we can see many problems that arises after seeing the complexity of data mining and security. As we progress into the future these issues we only become more complex as the field of computer science and mathematics continues to improve and technology continues become cheaper and more efficient. We can clearly see how there is no absolute method to protect a user’s security and there is no absolute method to corrupt the security of the users information, but rather a constant battle between the two opposing sides of attempting to access and attempting to deny access user information

References

1.

“When do data mining results violate privacy?”, Murat Kantrarcioglu, Jiashun Jin, Chris Clifton. Conference:  Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Published August 22 2004.

2.

“Optimal randomization for privacy preserving data mining”, Yu Zhu. Lei Liu. Conference: Proceedings of the tenth ACM SIGKDD international conference on Knowledge sdiscovery and data mining. Published August 22 2004.

3.

“Data types generalization for data algorithms”. Mon-Fong Jiang. Conference: Systems, Man Cybernetics, 1999, IEEE SMC ’99 Conference Proceedings. (vol 3). Published October 12 1999

4.

“Bottom-up generalization: a data mining solution to privacy protection”. Ke Wang. Conference: Data Ming, 2004. ICDM’04 FOURTH IEEE International Conference on. Published Nov 4 2004

5.

“Information Security in Big Data: Privacy and Data Mining”. Lei Xu. Conference: Access, IEEE (Volume:2). Published 09 October 201

6.

“Data Mining for Malicious Code Detection and Security Applications”. Thuraisingham, Bhavani. Conference: Intelligence and Security Informatics Conference (EISIC), 2011 European. Published sept 14 2011

7.

“Data mining for security applications”. Thuraisingham, Bhavani. Conference: Machine Learning and Applications, 2004. Proceedings. 2004 International Conference on. Published December 18 2004.