EFFECT OF A DATA MINING EXERCISE

[1]Pradeep Agrawal & [2]Yudhisthir Sharma

Research Scholar, Monad University, Hapur (India)

ABSTRACT

To papers entitled "privacy Preserving Data mining" appeared in 2000. Although both addressed a similar problem, constructing decision threes from private training data, the concepts of privacy were quite different. One was based on data obscuration, i.e., modifying the data values so real values are to disclosed (Agrawal and Srikant 2000) [5]. The other used secure multiparty computation (SMC) to "encrypt" data values (Lindell and Pinkas 2000) [6], ensuring that no party learns anything about another's data values. We first describe SMC, and then give additional background on data obscuration. We also discuss a problem that has received little attention: How do we constrain data mining if it is possible that the result along violate privacy?

I. INTRODUCTION

1.1 APPROACHES TO PRIVACY IN DATA MINING

1.1.1. Secure Multiparty Computation

The ideal of secure multiparty computation (SMC) (Yao 1986: Gldrech, Micali, and Wigderson 1987) [7] is that the parties involved learn nothing but the results, informations; we have a trusted third party to which all parties give their input. The trusted party computes the output. SMC enable this with the trusted third party. There may be considerable communication between the parties to get the final results, but the parties don't learn anything from this communication. The computation is secure if given just one party's input and output from those runs; we can simulate what would be seen by the party. In this case, to simulate means that the distribution of the simulated view over many runs are computationally indistinguishable. We may not be able to exactly simulate every run, but over time we cannot tell the simulation from the real runs.

1.1.2. Obscuring Data

Another approach to privacy is to obscure data: making private data available, but with enough noise add the exact value cannot be determined. One approach, typically used in census data, is to aggregate items. Knowing the average income for a neighborhood is not enough to determine the actual income of a resident of that neighborhood. An alternative is to add random noise to data values, the mine the distorted data. While this lowers the accuracy of data mining results, research has shown that the loss of accuracy can be small relative to the loss of ability to estimate an individual item.

1.1.3. Perfect Privacy

One problem with the above is the tradeoff between privacy and accuracy of the data mining results. SMC does better, but at a high computational and communication cost. In the "web survey example, the respondents could engage in a secure multiparty computation to obtain the results, and reveal no information that is not contained in the results. However getting thousand of respondents to participate synchronously in a complex protocol is impractical. While useful in the corporate model, it is not appropriate for the web model. Here we present a solution based on moderately trusted third parties – the parties are not trusted with exact data, but trusted only not to collude with the data receiver."

1.2 Individual Privacy

Most legal efforts have been directed to protecting date of the individual. For example; the European Community regulates personal date (Official Journal of the European Communities 1995):"Personal data" shall mean any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specifies to his physical, physiological, mental, economic, cultural or social identity and specified that data can be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member states shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

1.3 Network Privacy

Design and operations for privacy mechanisms for the network. Which is the fourth fallacy of distributed computing? Network privacy mechanisms, such as network firewalls and network intrusion detection devices, are generally a convenient and scalable point to apply privacy controls and are an important locale for defining chokepoints and zones. Zones define logical and/or physical boundaries around a group of systems, for example the DMZ pattern in web applications. Chokepoints define places to cross boundaries into and out of zones, where special privacy considerations apply.

1.4 Application Privacy

Deals with two main concerns: 1) protecting the code and services running on the system, who is connecting to them, and what is output from the programs through a combination of secure coding practices, static analysis, threat modeling, participation in the SDL, application scanning, and fuzzing. 2) Delivering reusable application privacy services such as reusable authentication, authorization, and auditing services enabling developers to build privacy into their system. Privacy frequently collaborates with software architects and developers in this area to build privacy into the system.

II THE SUCCESS OF A DM EXERCISE IS DRIVEN TO A VERY LARGE EXTENT BY THE FOLLOWING FACTORS

2.1.1. Availability of data with rich descriptions

This means that unless the relations captured in the database are of high degree, extracting hidden patterns and relationships among the various attributes will not make any practical sense. Availability of a large volume of data: This is mostly mandated for statistical significance of the rules to hold. Absence of say, at least a hundred thousand transactions will most likely reduce the usefulness of the rules generated from the transactional database.

2.1.2. Reliability of the data available

Although a given terabyte database may have hundreds of attributes per relation, the DM algorithms run on this dataset may be rendered defunct if the data itself was generated by manual and error prone means and wrong default values were set. Also, the lesser the integration with legacy applications, the better the accuracy of the data set.

2.1.3. Ease of quantification of the return on investment (ROI) in DM

Although the earlier three factors may be favorable, unless a strong business case can be easily made, investments in the next level DM efforts may not be possible. In other words, the utility of the DM exercise needs to be quantified vis-a-vis the domain of application.

2.1.4. Ease of interfacing with legacy systems

It is commonplace to find large organizations run on several legacy systems that generate huge volumes of data. A DM exercise which is usually preceded by other exercises like extract, transformation and loading (ETL), data filtering etc, should not add more overheads to system integration.

It must now be noted that e-commerce data, being the result of on-line transactions, do satisfy all the above proper criteria for data mining. We observe that once the back-end databases are properly designed to capture customer buying behaviour, and provided that default data take care of missing and nonexistent data, the first issue of availability of data with rich descriptions is taken care of. Similarly, the reliability of data collected is also ensured because it is possible to increase the so called no-touch-throughput in e-commerce transactions. Technologies like ebXML, BizTalk and Rosetta Net enhance the quality of data that is generated.

The ROI in DM exercises related to e-commerce can be easily quantified. For instance, mining the web logs certainly enhances web server architecture-related decisions. Improved web server availability results in faster transactions, thus increasing the revenue. Observe that increasing the number of transactions directly results in improved profits. Lastly, e-commerce systems usually follow the MVC (Model- View-Controller) pattern with the business execution systems conforming to the model tier, the browser being the view tier and interfacing mechanisms like Java Servlets or Microsoft ASP forming the controller tier. Data mining in e-commerce mostly relies on the controller for generating the data to mine on. Thus integration issues also do not surface in this case. In summary, it is little surprise that e-commerce is the killer application for data mining.

III CONCLUSION

Recently, the SANS Institute has made web application privacy the number one threat in their Top Twenty Privacy Attack Targets (2006 Annual Update). The analyst community agrees, noting over 75% of applications is vulnerable and 70% of attacks are now focused on these custom applications. Custom applications and services are the hackers' favorite target. The technology is evolving and connecting so quickly that it has been very difficult for the privacy community to keep up. The attackers know this and they're taking full advantage. Application privacy is challenging, and there are many tempting approaches out there. We're here to tell you that if you want to get value out of your application privacy efforts, put a plan in place that will drive deep visibility into application privacy. Then you can manage with metrics. In our experience, organizations that establish an application privacy team are the most likely to succeed. The team should be responsible for both facilitating visibility and leading efforts to improve privacy. Typically, those teams do training, verification, process, tools, architecture, etc.

REFERENCE

[1]. David Geer. Taking Steps to Secure Web Services.Computer.

[2]. Secure, Reliable, Transacted Web Services: Architecture and Composition. IBM Corporation, Microsoft Corporation, 2003 September.

[3]. Olwyn Dowling, Sarah Evans. Is SSL enough privacy for first-generation Web Services? Web Services.Org, 2003.

[4]. UcheOgbuji. The Past, Present and Future of Web Services. WebServices.Org,2003.

[5]. David Chappell. New Technologies Help You Make Your Web Services More Secure. MSDN Magazine, 2003 April.

[6]. Ray Djajadinata. Yes, You Can Secure Your Web Services Documents. Java World, 2002 August.

[7]. Adam Bosworth. Developing Web Services.Crossgain Corporation, 2001.

[8]. Yuichi Nakamur, Satoshi Hada, Ryo neyama. Towards the Integration of Web Services Privacy on Enterprise Environments.IEEE, 2002.

[9]. Eric Newcomer. Understanding Web Services: XML, WSDL, SOAP, and UDDI. Addison Wesley, 2002 May.

[10]. Huysmans J., Baesens B., Vanthienen J.: Using Rule Extraction to Improve the Comprehensibility of Predictive Models. FETEW Research Report (2006) 1-55.

[11]. Doughetry, J. Kohvi, M. Sahami, M.: Supervised and unsupervised discretiazation of continuous features. I: Int. Conf. Machine Learning (1995) 194-202.