TOWARDS EFFICIENT CRYPTOGRAPHY FOR PRIVACY PRESERVING DATA MINING IN DISTRIBUTED SYSTEMS

Emmanouil Magkos and Vassilis Chrissikopoulos
Department of Informatics, Ionian University, Palaia Anaktora, 49100, Corfu, Greece

Keywords: Data mining, Privacy, Security, Cryptography, Distributed systems.

Abstract: A common fact for both businesses and physical entities is that sensitive, accurate information would be more easily diffused if adequate measures for protection were in place. This could also lead to higher quality data mining results, in a privacy preserving manner. Recent research has proved that it is possible to provide both privacy and accuracy assurances in a distributed computing scenario, where all participants may be mutually untrusted, without the presence of an unconditionally trusted third party. We believe that valuable knowledge can be borrowed from the vast body of literature on e-auction and e-voting systems, in order to be adapted to privacy preserving data mining systems in a distributed environment. These systems tend to balance well the efficiency and security criteria, because they need to be implementable in medium to large scale environments.

1 INTRODUCTION

Following the explosion of Information and Communications Technologies in the last decade, we are witnessing the advent of the digital era in an emphatic way: Most businesses and organizations have transformed their processes and relationships with their partners and customers into fully electronic ones, while the convergence of telecommunication networks have boosted the use of the Internet for everyday activities. Furthermore, advances in both software and hardware, combined with emerging technologies such as Web 2.0, Grid computing, and Semantic Web, have facilitated the collection, storage and processing of large volumes of digital data, at a relatively low cost.

Data Mining (DM) techniques are well known for extracting valuable, and usually not obvious, information from large quantities of data (Chen et al, 1996). Data mining has broad applications in areas related to market research, as well as to financial and scientific research. Most techniques can be categorized into two generic classes (Dunham, 2002): In the *predictive class*, original (*i.e.*, training) data are processed for handling, describing, or predicting future data or phenomena. In the *descriptive class*, DM techniques are trying to develop a better description for objects in databases. In this paper we focus on descriptive techniques, which can also be seen as a necessary layer that facilitates higher level (*e.g.*, predictive) DM techniques. For example, *classification* techniques aim at classifying objects of a database into a discrete category, based on the values of some attributes in a transaction. *Association rules* DM statistically process a set of transactions in order to extract high frequencies of occurring patterns such as “customers who buy milk also tend to bye cookies”. In *summarization* techniques, the general idea is to use statistics, in order to better describe and present a large set of data at an abstraction level, for further processing (Chen et al, 1996).

Privacy surveys have shown that Web users may be willing to divulge some personal information, in exchange for getting something valuable in return (*e.g.*, personalization and customization services, or better search results) (Yang et al, 2005). Additionally, in the corporate environment, while most organizations normally recognize the importance of privacy protection (Deloitte, 2007) (although their theoresis may be influenced by factors such as laws and privacy regulations, comformity with security standards, reputation and the fear of competition), there are cases where different organizations may have strong incentives in sharing private information for extracting valuable knowledge. For example, two medical institutions would find it useful to selectively pool

Magkos E. and Chrissikopoulos V. (2008).
TOWARDS EFFICIENT CRYPTOGRAPHY FOR PRIVACY PRESERVING DATA MINING IN DISTRIBUTED SYSTEMS.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 301-304
DOI: 10.5220/0001531403010304

their medical records for doing research in epidemiology or improving medical diagnosis. Or, several competing businesses would like to pool their customer transaction data, and get in return a research analysis on the market trends. Unfortunately, this is usually either not possible or not secure, mainly due to confidentiality concerns (Lindell and Pinkas, 2000). A common fact for both businesses and physical entities is that sensitive, accurate information would be more easily diffused if adequate measures for protection and security were in place. This could also lead to higher quality data mining results, in a privacy preserving manner.

2 SECURITY IN DISTRIBUTED DATA MINING SYSTEMS

In large intra-organizational environments, data are usually shared among a number of distributed databases, for security or practicality reasons, or due to the organizational structure of the business. Data can be partitioned either *horizontally*, where each database contains a subset of complete transactions (Lindell and Pinkas, 2002; Kantarcioglu and Clifton, 2004), or *vertically*, where each database contains shares of each transaction (Vaidya and Clifton, 2002). The role of a *data warehouse* is to collect and transform the dispersed data to an acceptable format, before they will be forwarded to the DM subsystem. Such central repository raises privacy concerns, especially if it used in an *inter-organizational* setting where several entities, mutually untrusted, may desire to mine their private inputs, both securely and accurately. Alternatively, data mining can be performed locally, at each database (or intranet), and then the subresults be combined to extract knowledge, although this will most likely affect the quality of the output (Vaidya and Clifton, 2002).

If a general discussion was to be made about protecting privacy in distributed databases, we would point to the literature for *access control* and audit policies, authorization and information flow control (*e.g.*, multilevel and multilateral security strategies (Anderson, 2001)), security in the application layer (*e.g.*, database views), and Operating Systems security among others. However in this paper we assume that appropriate security and access control exist in the intra-organizational setting, and we mainly focus on the inter-organizational setting where a set of mutually untrusted entities wish to execute a miner on their private databases. As an alternative layer of protection, original data can be suitably altered (*e.g. randomized*) (Agrawal and Srikant, 2000) or *anonymized* before given as an input to a miner, or queries in statistical databases may be restricted (Anderson, 2001). The problem with data perturbation is that in highly distributed environments, preventing the *inference* of unauthorized information by combining authorized information is not an easy problem (Ferrer, 2002). Furthermore, in most perturbation techniques lies a *tradeoff* between protecting privacy of the individual records and at the same time establishing accuracy of the DM results (Wang and Zhang, 2007). In the next session we discuss the important role of cryptography in privacy preserving data mining.

3 CRYPTOGRAPHY IN PRIVACY PRESERVING DATA MINING

At a high abstraction level, the problem of privacy preserving data mining between mutually untrusted parties can be reduced to the following problem for a two-party protocol: Each party owns some private data and both parties wish to execute a *function F* on the union of their data without sacrificing the privacy of their inputs (Pinkas, 2002). In a DM environment, for example, the function *F* could be a classification function that outputs the class of a set of transactions with specific attributes, a function that identifies association rules in partitioned databases, or a function that outputs aggregate results over the union of two statistical databases.

In the above distributed computing scenario, an "ideal" protocol would require a trusted third party who would accept both inputs and announce the output. However, the goal of cryptography is to relax or even destroy the need for trusted parties. Contrary to other strategies, crypto mechanisms usually do not pose dilemmas between the privacy of the inputs and the accuracy of the output.

In the academic literature for privacy preserving data mining, following the line of work that begun with Yao (Yao, 1986), most theoretical results are based on the *Secure Multiparty Computation* (SMC) approach (*e.g.* Lindell and Pinkas, 2002; Vaidya and Clifton, 2002; Kantarcioglu and Clifton, 2004). SMC protocols are *interactive protocols*, run in a distributed network by a set of entities with private inputs, who wish to compute a function of their inputs in a privacy preserving manner. The goal is that no more information is revealed to an entity in the computation than can be inferred from that participant's input and output (Goldwasser, 1997).

For example, in (Kantarcioglu and Clifton, 2004), there are three sites with sales transactions, where each transaction contains an itemset, and all three sites wish to combine their itemsets in order to find the most popular items (*e.g.*, thus being able to mine association rules). Each site uses a symmetric cipher that is also *commutative*, *i.e.*, $E_A(E_B(X)) = E_B(E_A(X))$, encrypts its itemset and passes it to the next site, until all itemsets have 3 layers of encryptions. Then, all encrypted itemsets are again passed around, with each site decrypting, until the complete set is revealed.

Unfortunately, as has been noted in the literature, SMC protocols require multiple communication rounds among the participants, and privacy usually comes at a high performance and communication cost (Pinkas, 2002).

We believe that research for privacy preserving DM could borrow knowledge from the vast body of literature on secure *e-auction* (Naor et al, 1999) and *e-voting* systems (Gritzalis, 2002). These systems are not strictly related to data mining but, they exemplify some of the difficulties of the multiparty case (this has been pointed out first by (Pinkas, 2002) but it only concerned e-auctions, while we extend it to include e-voting systems as well). Such systems also tend to balance well the efficiency and security criteria, in order to be implementable in medium to large scale environments. Furthermore, such systems fall within our distributed computing scenario and have similar architecture and security requirements, at least at our abstraction level.

In a sealed bid e-auction for example, the function *F*, represented by an auctioneer, receives several encrypted bids and declares the winning bid. In a secure auction, there is a need to protect the privacy of the loosing bidders, while establishing accuracy of the auction outcome and verifiability for all participants. Or, in an Internet election, the function *F*, represented by an election authority, receives several encrypted votes and declares the winning candidate. Here the goal is to protect the privacy of the voters (*i.e.*, unlinkability between the identity of the voter and the vote that has been cast), while also establishing eligibility of the voters and verifiability for the election result.

During the last decade, a few cryptographic schemes for conducting online e-auctions and e-elections have been proposed in the literature. Research has shown that it is possible to provide both privacy and accuracy assurances in a distributed computing scenario, where all participants may be mutually untrusted, without the presence of an unconditionally trusted third party. Most efficient schemes of both worlds (*e.g.*, Parkes et al, 2006, Damgard et al, 2003), follow the *homomorphic model* (Cramer et al, 1997), originally proposed in the e-voting setting, where the privacy of the inputs and the accuracy of the results can be universally verified, thanks to the algebraic properties of several randomized encryption algorithms. With homomorphic encryption there is an operation $\oplus$ defined on the message space and an operation $\otimes$ defined on the cipher space, such that the "product" of the encryptions of any two private inputs is the encryption of the "sum" of the inputs:

$$E(M_1) \otimes E(M_2) = E(M_1 \oplus M_2) \qquad (1)$$

This property allows, for example, either to tally votes as aggregates or to combine shares of votes (*e.g.*, Cramer, 1997; Schoenmakers, 1999), without decrypting single votes.

We have to note that in the most practical schemes of both worlds of e-auctions and e-voting, a third party is still needed to preclude communication between participating entities, however no third party is blindly trusted on either the privacy of the inputs or the accuracy of the results. This is achieved by using robust cryptographic primitives such as *threshold decryption* (Desmedt, 1994). For example, a set of *M* voting authorities in a (*t*, *M*) threshold public-key encryption system share a private key, and there is one public key corresponding to the shared private key. After all encrypted inputs have been submitted, any subset of *t* honest and functioning authorities are able to combine their key shares and decrypt the encrypted aggregate.

Selecting the most efficient mechanisms from both worlds could also be facilitated by the fact that when considering distributed DM systems, the *threat model* seems to be relaxed, thus better balancing the trade-off in favour of efficiency: for example, verifiability does not have to be universal (but atomic), and there seems not to be an obvious need for receipt-freeness (as in elections).

The first implementation of the homomorphic model of (Cramer et al, 1997) in the DM setting, to the best of our knowledge, has recently appeared in (Yang et al, 2005). In a simplified version of their scheme, which involves horizontally partitioned databases, each client holds a record of personal data (*e.g.*, age, income, marital status, history of accidents) and a data miner can pose questions such as "How many people have income > 25 and are married?". Each client sends to the data miner the (homomorphic) encryption of "1" ("yes") or "0" ("no") and the system computes the aggregates.

While the scheme in (Yang et al, 2005) is based on the homomorphic model of (Cramer et al, 1997) that supports *1-out-of-2* ("yes"/ "no") selections, we believe that future research could also look at some very efficient extensions of the homomorphic model, where *1-out-of-L* or *K-out-of-L* selections are allowed (e.g., Baudron et al, 2001; Damgard et al, 2003). In this way, the overall bits of information that a database sends to the miner could be increased, leading to new possibilities.

4 CONCLUSIONS

We believe that valuable knowledge can be borrowed from the vast cryptographic literature on e-auction and e-voting systems, in order to be adapted to the specific requirements for privacy preserving data mining systems in a distributed environment. These systems tend to balance well the efficiency and security criteria, because they need to be implementable in medium to large scale environments.

Of course, further research is needed to choose and then adapt the specific cryptographic techniques to the DM environment, taking into account the kind of databases to work with, the kind of knowledge to be mined, as well as the kind of specific DM technique to be used.

REFERENCES

Agrawal, R., Srikant, R., 2000. Privacy-preserving data μining. In *ACM SIGMOD Conference on Management of Data*. ACM Press, pp. 439-450.

Anderson, R., 2001. *Security engineering – A guide to building dependable distributed systems*. Wiley Computer Publishing.

Baudron, O., Fouque, P., Pointcheval, D., Poupard, G., Stern, J., 2001. Practical Multi-Candidate Election System. In *20th ACM Symposium on Principles of Distributed Computing*. ACM Press, pp. 274–283.

Chen, M., Han, J., Yu, P., 1996. Data mining: An overview from a database perspective. In *IEEE Transactions on Knowledge and Data Engineering*. IEEE Press, Vol. 8 (6), pp. 866-883.

Cramer, R., Gennaro, R., Schoenmakers, B., 1997. A secure and optimally efficient multi-authority election scheme. In *European Transactions on Telecommunications*. Vol. 8 (5), pp. 481-490.

Damgard, I., Jurik, M., Nielsen, J., 2003. *A generalization of Paillier's public-key system with applications to electronic voting*. Manuscript. Available at: www.daimi.au.dk/~ivan/GenPaillier_finaljour.ps

Deloitte, 2007. *Global security survey 2007*. Deloitte Touche Tohmatsu. Available at: http://www.deloitte.com/dtt/cda/doc/content/arg_cons_encuesta-global-Seguridad-2007_20071031(2).pdf

Desmedt, Y., 1994. Threshold Cryptography. In *European Transactions on Telecommunications*. Vol. 5(4), pp. 449–457.

Dunham, M., 2002. *Data mining, introductory and advanced topics*. Prentice Hall.

Ferrer, J. (Ed.), 2002. *Inference control in statistical databases, from theory to practice*. Springer, LNCS Vol. 2316.

Goldwasser, S., 1997. Multi-party computations: Past and present. In *16th Annual ACM Symposium on principles of Distributed Computing*. ACM, pp. 1-6.

Gritzalis, D. (Ed.), 2002. *Secure electronic voting: trends and perspectives, capabilities and limitations*. Kluwer Academic Publishers.

Kantarcioglu, M., Clifton, C., 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *IEEE Transactions on Knowledge and Data Engineering*. IEEE Press, Vol. 16 (9), pp. 1026-1037.

Lindell, Y., Pinkas, B., 2000. Privacy preserving data mining. In *Advances in Cryptology - CRYPTO '00*. Springer, LNCS Vol. 1880, pp. 36–53.

Naor, M., Pinkas, B., Sumner, R., 1999. Privacy preserving auctions and mechanism design. In *1st ACM conference on Electronic commerce*. ACM Press, pp. 129 – 139.

Parkes, D., Rabin, M., Shieber, S., Thorpe, C., 2006. Practical secrecy-preserving, verifiably correct and trustworthy auctions. In *8th ACM International Conference on Electronic Commerce*. ACM Press, pp. 70 – 81.

Pinkas, B., 2002. Cryptographic techniques for privacy-preserving data mining. In *SIGKDD Explorations*. ACM Press, Vol. 4(2), pp. 12-19.

Schoenmakers, B., 1999. A Simple Publicly Verifiable Secret Sharing Scheme and Its Application to Electronic Voting. In *Advances in Cryptology–CRYPTO'99*. Springer LNCS Vol. 1666. pp. 148-164.

Vaidya, J., Clifton, C., 2002. Privacy preserving association rule mining in vertically partitioned data. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, pp. 639-644.

Wang, J., Zhang, J., 2007. Addressing accuracy issues in privacy preserving data mining through matrix Factorization. In ISI'07, IEEE International Conference on Intelligence and Security Informatics. IEEE Press, pp. 217-220.

Yang, Z., Zhong, S., Wright, R., 2005. Privacy-preserving classification of customer data without loss of accuracy. In SDM'05 SIAM Data Mining Conference.

Yao, A., 1986. How to generate and exchange secrets. In 27th Symposium on Foundations of Computer Science. IEEE Press, pp. 162–167.