An e-voting-based data gathering scheme for decision support systems

V.I. Galanis and E.K. Ikonomakis

Computational Intelligence Laboratory (CILab),
Department of Mathematics,
University of Patras,
GR-26110, Patras, Greece
and
University of Patras Artificial Intelligence Research Center (UPAIRC),
GR-26110, Patras, Greece
E-mail: basgal@master.math.upatras.gr
E-mail: eki@math.upatras.gr

G.C. Meletiou*

Epirus Institute of Technology,
P.O. 110, GR-47100 Arta, Greece
and
University of Patras Artificial Intelligence Research Center (UPAIRC),
GR-26110, Patras, Greece
E-mail: gmelet@teiep.gr
*Corresponding author

M.N. Vrahatis

Computational Intelligence Laboratory (CILab),
Department of Mathematics,
University of Patras,
GR-26110, Patras, Greece
and
University of Patras Artificial Intelligence Research Center (UPAIRC),
GR-26110, Patras, Greece
E-mail: vrahatis@math.upatras.gr

Abstract: In this contribution, a protocol for secure data gathering is introduced that utilises existing e-voting protocols to gather data records instead of votes. The protocol relies on two distinct authorities for the secure gathering of data and makes use of a token for each database record, making unencrypted data accessible only to the system participants. The protocol is suitable for decision support systems among non-mutually trusted parties as it protects the anonymity and privacy of the parties by dissociating the data from their origin and protects both the validity of the data sent by each party and the system by intruders and malicious participants.

Keywords: cryptography; electronic data gathering; e-voting; database record labelling; privacy preserving.

Reference to this paper should be made as follows: Galanis, V.I., Ikonomakis, E.K., Meletiou, G.C. and Vrahatis, M.N. (2010) 'An e-voting-based data gathering scheme for decision support systems', *Int. J. Decision Sciences, Risk and Management*, Vol. 2, Nos. 1/2, pp.36–45.

Biographical notes: Vassilios I. Galanis received his Diploma in Applied Mathematics from the School of Applied Mathematics and Physical Sciences of the National Technical University of Athens in 2006 and his MSc in Mathematics of Computers and Decision Making from the Departments of Mathematics and Computer Engineering and Informatics, University of Patras, Greece in 2009. He is currently a PhD candidate at the same department. His research interests are computational intelligence methods and computational mathematics in cryptography and cryptanalysis. He is the Co-author of four publications.

Emmanouil K. Ikonomakis received his Diploma in Mathematics from the Department of Mathematics, University of Patras in 2005 and his MSc in Computational Mathematics and Informatics: Computer Mathematics and Artificial Intelligence, from the Department of Mathematics, University of Patras, Greece, in 2009. He is currently a PhD candidate at the same department. His research interests are computational intelligence methods in text mining. He is the Co-author of four publications.

Gerasimos Meletiou is a Professor of Applied Mathematics at Epirus Institute of Technology. He is a Graduate of the Department of Mathematics of the University of Athens, where he also completed his Doctoral degree in 1984. He has taught both at postgraduate (University of Patras) and undergraduate level (Epirus Institute of Technology, Hellenic Open University). His research interests, as expressed in publications (he has published more than 30 papers in international journals), involve computational algebra, finite fields, data security, cryptography, data mining, discrete mathematics, median structures, etc.

Michael N. Vrahatis received his PhD in Computational Mathematics (1982) from the University of Patras, Greece. He is a Professor of Computational Mathematics in the University of Patras, since 2000 and serves as the Director of the Computational Intelligence Laboratory of the same department. He has participated in the organisation of over 70 conferences serving at several positions, and participated in over 200 conferences, congresses and advanced schools as a Speaker or Observer. He is a member of the Editorial Board of eight international journals. His work consists of over 340 publications that have been cited by researchers over 4,075 times. His work includes topological degree theory, systems of non-linear equations, numerical and intelligent optimisation, data mining and unsupervised clustering, cryptography and cryptanalysis as well as computational and swarm intelligence.

1 Introduction

In today's digital world, information is the most valuable commodity. Businesses and organisations keep records of every information necessary in databases in order to gain knowledge from their analysis. To meet this need, the development of methods that extract knowledge from a database, using data mining techniques, is an active area of research and significant effort has been devoted to it.

The reliability of a decision support system is based on the broadness of information available in the organisation's database. The anonymous and secure gathering of databases containing the relevant data would make possible the acquisition of the necessary information while maintaining the confidentiality of the data by making the tracing of their individual source infeasible. Fields like economics, marketing, supply chain management and others benefit greatly from the application of knowledge discovery techniques. Let us consider the scenario of comparative case studies in a corporate sector. Although comparative case studies are a valuable tool for decision-making, it requires the sharing of confidential information of each company's database. In addition, a single database of a company may not contain sufficient data for effective knowledge extraction due to the fact that several factors limit the global scope of these rules. The cognition derived from that company's database is restricted to the profile of that single company, such as the company's size, the scope of its consumer target group and its product range. So, it is clear that in order to have effective knowledge extraction, the need for unification of a number of relevant databases, so as to have the broadest range of information available, is obvious. This need contradicts the confidentiality of corporate data. Thus, we need a way to amass and unify these databases without revealing each database's origin, thus, protecting the confidentiality of the data.

A suitable real life example would be hospital databases holding individual patient records. As with corporate data, medical records of individual patients in a hospital database are not accessible by anyone other than the patients themselves and the corresponding hospital. An automated decision support system could utilise such medical records by checking the available data in the medical database and providing information to doctors regarding previous treatment of relevant cases in order to determine an optimal treatment for each case. Such a system would be as effective as the variety and volume of data that would be available to it while not violating the confidentiality of the medical records.

Thus, the need for protocols that will allow each organisation to exchange data from their database with other organisations for mutual interest in such a way that every organisation is not identified by its individual data arises. In the rest of this contribution, each organisation will be referred to as Alice due to a widespread convention used in cryptography.

These protocols need to satisfy the following prerequisites:

a *completeness:* all valid data are gathered correctly if all organisations follow the protocol

b *privacy:* it is infeasible to associate individual databases to the organisations

c *eligibility:* only legitimate organisations are allowed to send data

d *authentication:* each organisation sends valid data

e *verifiability:* each organisation is able to verify whether its data are correctly included in the aggregated data.

With the exception of the fourth, these requirements emulate the ones needed for secure voting systems on a public network.

Privacy preserving data gathering protocols use one or even no authorities to gather the data. In our approach, we propose a protocol that fulfils the aforementioned requirements.

2 The protocol of the double 'Bulletin Board-Tallier' authority scheme

Let us consider the scenario where there are two authorities participating in the data gathering procedure. These are named by convention the Bulletin Board and the Tallier by convention used in the description of e-voting systems. The Bulletin Board serves as the record keeper of the voters assuring their eligibility in participating in the elections and preventing non-eligible participants from being able to take part in the process, while the Tallier is the deposit of eligible votes. We use a modification of e-voting protocols to fulfil the requirements for such a system. Specifically, the present scheme is based on the works of Her and Abe (1998), Her et al. (2005), Cramer et al. (1997), Fujioka et al. (1992), Stadler (1996) and Laskari et al. (2004). The scheme relies on the labelling of database records, each record using a suitable structure that allows the verification of the eligibility of each record without exposing any additional data that would reveal the identity of the sender of each record.

Alices are organisations working on the same field. The Bulletin Board and Tallier entities could be either a national authority, an organisation or a machine. The Bulletin Board holds a list of the identities of all Alices that will participate in the electronic data gathering process.

Each record of a database is assumed to be a vector of values (numerical or categorical) of the form $\overline{w}_j=(w_{j1}, w_{j2}, \ldots, w_{jm})$. Each must send $\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_k$ to all other Alices via the Bulletin Board and the Tallier. The procedure begins with Alice having the $\overline{w}_j$ record signed by the Experts Union forming the pairs $A_j = [\overline{w}_j, Sign_{EU}(\overline{w}_j)]$ for all the Experts Union records. Alice verifies the Experts Union signature.

Each Alice proceeds to attach a label to each A_j in the form of y_j, where y_j is the building element of the identification scheme. Our protocol allows for a variety of identification schemes available to be used to establish the eligibility of each record sent by an Alice to the Tallier. Such schemes include the Feige-Fiat-Shamir scheme (Feige et al., 1988), the Guillou-Quisquater scheme (Guillou and Quisquater, 1990), the Schnorr identification scheme (Schnorr, 1991) as well as schemes proposed by Stadler (1996) and Cramer et al. (1997). Then, for every record, Alice has the label signed by the Experts Union, $Sign_{EU}(y_j)$. Thus, for each record, we have the formation of a set $[w_j, Sign_{EU}(w_j), Sign_{EU}(y_j)]$. When each Expert has completed the process, Alice verifies the Experts Union signature on every record $\overline{w}_j$ and label y_j and, then, signs for the Expert the number of total records, denoted by *nor*. Alice sends a message to the Bulletin Board containing a triple of the form $[nor, Sign_A(nor), Sign_{EU}(nor)]$.

The Bulletin Board verifies the Experts' Union signature on every triple, deduces from the triples the total number of expected records and sets a deadline for the y_j receipt of all the sets of labels.

For each record $\overline{w}_j$, Alice forms the set of labels $lab_j = [y_j, Sign_{EU}(y_j)]$.

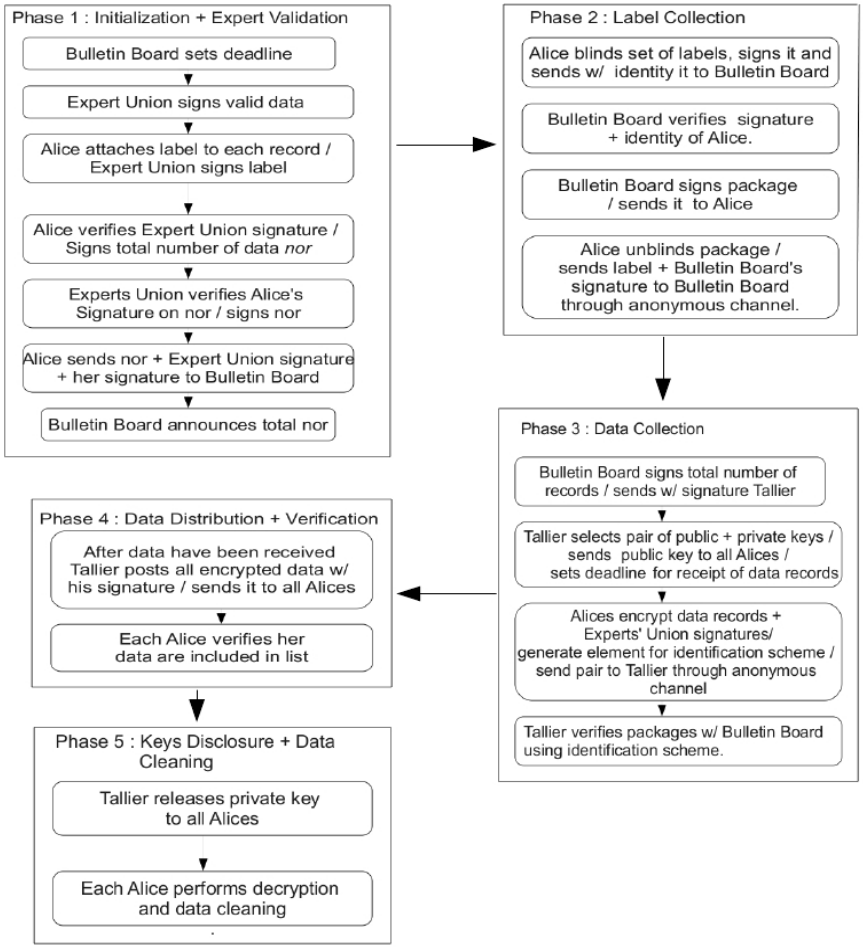
Alice, then, blind signs the lab_j as $e_j = blind[lab_j, r_j]$ where r_j is a randomly selected blinding factor (Chaum, 1982). Alice also signs e_j, $Sign_A(e_j)$ and sends the triple $[Id_A, e_j, Sign_A(e_j)]$ to the Bulletin Board to be verified. Id_A is an identity used by the corresponding Alice A.

The Bulletin Board verifies Alice's signature and checks if Alice is registered on the list of participants. If this is true, it signs e_j as $SignBB(e_j)$ and sends it back to Alice. Alice verifies the Bulletin Board's signature over her set of labels lab_j, unblinds it and sends the pair $[lab_j, SignBB(e_j)]$ to the Bulletin Board via an anonymous channel (Chaum, 1982; Abe, 1998; Ohkubo et al., 1990; Ogata et al., 1997). The Bulletin Board, then, proceeds

to merge the lab_j sets to construct the *lab* set of labels which contains the representations of the labels of every record w_j of every Alice. These messages are of very small size, and therefore, they do not add a significant size to the packages transmitted.

After that, the Bulletin Board signs the total number of expected records, denoted by nor_T, signs it with $SignBB(nor_T)$ and sends $[nor_T, SignBB(nor_T)]$ to the Tallier. The Tallier sets a deadline for the receipt of all the packages of $\overline{w}_j$ from every Alice and selects a pair (D_0, E_0) of private and public keys, announcing E_0 to all Alices for the encryption of all packages. The public key E_0 is used by all Alices for the encryption of all packages $A_j = [w_j, Sign_{EU}(\overline{w}_j)]$.

Figure 1 Flowchart of the protocol



Subsequently, Alice creates $(E_0(\overline{w}_j, Sign_{EU}(\overline{w}_j)), y'_j)$ and sends it to the Tallier via an anonymous channel. The y'_j denotes the corresponding element to y_j, with which the Tallier performs the eligibility authentication with the Bulletin Board using the identification scheme to generate y_j. The corresponding y'_j needs to fulfil certain

requirements to enable the Bulletin Board to prove to the Tallier that the w_j package sent to it by Alice is an eligible package without revealing any information regarding Alice's identity or the contents of the package.

The identification schemes are based on the concept zero-knowledge proofs to fulfil the above set of requirements.

Thus, the Tallier engages in the identification procedure via the identification scheme used until it is satisfied that the Bulletin Board indeed possess y_j sent to it by the corresponding Alice and so that the package received is a legal package. Once the packages number reach nor_T, the Tallier sends all the data with its signature on them to all Alices. Each Alice then verifies that her data are correctly included in the list.

Finally, the Tallier releases the private key D_0 to all Alices and each Alice, in turn, performs decryption of all the data and data cleaning by discarding any unverified records. In conclusion, all valid data are sent to all Alices.

The procedures involved in the proposed scheme can be summarised in Figure 1.

3 Security analysis

The security analysis of the proposed scheme is performed by considering various possible attack scenarios. After the expiration of the deadline, the Tallier checks the packages. Despite the increased time complexity of the scheme, the system benefits from the significant reduction of database record transmissions. This results in greatly reduced space complexity in terms of data volume and network overhead. We should note that due to the use of anonymous channels in the protocol neither the Bulletin Board nor the Tallier can be passive cheaters. Also, due to the nature of their role in the protocol, Alices cannot be passive cheaters. Thus, our security analysis shall confine itself to active cheating cases on behalf of the system's entities. The following cases are examined.

3.1 Alice acts as a cheater

1 Assume that one Alice wants to cheat the other Alices in order to see their data without submitting any of her data. In this case, the Bulletin Board can prove which Alice has not sent any data by checking the list of participants and the corresponding identifications Id_A it has received. So, it can ask Alice to provide it with its signature $SignBB(e_j)$ which Alice cannot forge.

2 Suppose Alice sends her labelset, receiving the Bulletin Board's signature on her labelset, but does not send the database record $\overline{w}_j$ to the Tallier. After the expiration of the deadline, the Tallier checks the package $[nor_T, SignBB(nor_T)]$ that it has received from the Bulletin Board and compares it to the total number of packages received through the anonymous channel.

   If the numbers do not agree, the Tallier does not release the private key and announces to all Alices that they have offended the regulations and also the exact number of data that have not been sent. The Tallier announces a final deadline for the release of missing data. If the total number of records is not sent until the new deadline expires, then the offender Alice can be disclosed by the Experts' record in the Bulletin Board. The same also holds if an Alice does not send all of her data.

3 Suppose that Alice sends invalid data in order to mislead all the other Alices. Those data will not be signed by the Expert and will not be verified, resulting them being discarded from the list in the last stage of the protocol leaving only valid data in. Also, the malicious Alice will be immediately exposed.

4 If an Alice or a malicious intruder tries to intercept the labels of one of Alice's database records then due to the identification schemes proposed for use in our protocol, this adversary will have to solve an instance of a computationally infeasible problem such as the factoring problem or the Diffie-Hellman problem to deduce any information regarding the corresponding package and its label.

3.2 The Bulletin Board acts as a cheater

1 Communications between each Alice and the Bulletin Board are registered submissions, since they are both signed by the sender. This means, the sender cannot cheat. Specifically, suppose that the Bulletin Board tries to abuse Alice by sending her a signed but faulty package $SignBB(e_j)$. Then, the Bulletin Board's verification over Alice's labelset fails and Alice can prove the fraud.

2 The Bulletin Board cannot compromise the protocol by adding a label to the labelset, as it signs only blinded data.

3.3 The Tallier acts as a cheater

Suppose the Tallier decides to use the private key D_0 prior to releasing it to all Alices so as to change the valid $\overline{w}_j$ with a record $\overline{w}'_j$ of its own to mislead the Alices. Then, Alice discovers her changed record $E(\overline{w}'_j)$ within the list and does the following.

She forces the Tallier to post $E(\overline{w}'_j)$ on the Bulletin Board, before she posts through an anonymous channel to the Bulletin Board the package $(E((\overline{w}_j), Sign_{EU}(w_j)), y'_j)$. After that, the Bulletin Board performs authentication of the package sent by Alice and verifies its eligibility. Since the encrypted records are different, Alice has proved that the Tallier has replaced the record and, thus, is a cheater without having revealed anything about her identity. Thus, verifiability is assured.

3.4 The Tallier and Bulletin Board cooperate as cheaters

Suppose the Tallier and the Bulletin Board reveal to each other the labels sent to each one of them by Alice. Since Alice has provided the different labels to both via an anonymous channels, it is not possible for them to connect Alice's identity to those labels.

4 Complexity issues

The number of signings, encryptions, decryptions, blindings, unblindings and sendings that take place within the protocol along with the identification scheme is called the complexity of a privacy preserving data gathering protocol. The total complexity of this proposed scheme is:

$$K_T = (12 + s) \cdot nor_T + 6e + 5n + \varepsilon + 2 \cdot \left(nor_T \cdot (k+1) \right)$$

where s is the running time for the identification scheme, n is the number of Alices, e is the number of Experts Alices employ, ε is the number of sendings from the Bulletin Board to The Tallier and $2 \cdot (nor_T \cdot (k+1))$ is the total cost for an anonymous channel with k-MIXes for nor_T sendings and one receiver, as described in Chaum (1982). Despite the increased time complexity of the scheme, the system benefits from the significant reduction of database record transmissions. This results in greatly reduced space complexity in terms of data volume and network overhead.

5 Comparison to other schemes

The focus of this contribution is the modification of electronic voting schemes for the purpose of data gathering in a privacy-preserving manner. Such a protocol was first introduced in Laskari et al. (2005), where the use of a Bulletin Board is introduced for the purpose of anonymously gathering database records instead of votes. Another protocol for this purpose was introduced in Laskari et al. (2004), where the data gathering process is carried out without the use of authorities but at great computational cost.

The novelty feature of our protocol is that the reconstruction of the key for the decryption of the unified database is accomplished by the Alices themselves, that is the entities that contribute the database records, at a small additional complexity cost. In this way, the Bulletin Board only gathers encrypted data, thus, having no access to actual database records. Additionally, due to the use of labels, there is reduced traffic of database records in the network, making this protocol suitable for large size databases.

6 Conclusions and future research

In this contribution, we propose a privacy preserving data gathering protocol that facilitates the creation of concentrated databases into which the use of data mining techniques allows the extraction of knowledge to support decision-making systems. Such decision support systems are a valuable tool to businesses and organisations, as by using data mining techniques they achieve the grouping of data and the extraction of rules and, thus, of new knowledge, which is immediately usable.

This work focuses on data gathering instead of processing distributed data to different sources, since, to support a viable real-time decision support system, we need continuous access to the data as well as unsupervised data mining techniques.

Privacy-preserving data gathering is a research topic with a great variety of applications. Future directions of research include the construction of new protocols utilising other types of e-voting schemes as well as novel techniques and the expansion of the protocol in order to exploit data that are produced in a ubiquitous computing framework. An important research topic, however, is the pairing of protocols of this type with privacy-preserving data mining protocols in order to produce automated, privacy-preserving decision support and electronic evaluation systems.

References

Abe, M. (1998) 'Universally verifiable mix-net with verification work independent of the number of mix servers', in *Lecture Notes in Computer Science*, Vol. 1403, pp.437–447.

Chaum, D.L. (1982) 'Blind signatures for untraceable payments', in *Advances in Cryptology, Proceedings of Crypto 82*, pp.199–203.

Cramer, R., Gennaro, R. and Schoenmakers, B. (1997) 'A secure and optimally efficient multi-authority election scheme', in *Lectures Notes in Computer Science*, Vol. 1233, pp.103–118.

Feige, U., Fiat, A. and Shamir, A. (1988) 'Zero-knowledge proofs of identity', *Journal of Cryptology*, Vol. 1, pp.77–94.

Fujioka, A., Okamoto, T. and Ohta, K. (1992) 'A practical secret voting scheme for large scale elections', in *Proceedings of ASIACRYPT 92 Lecture Notes in Computer Science*, Vol. 718, pp.244–251.

Guillou, L.S. and Quisquater, J.J. (1990) 'A paradoxical identity-based signature scheme resulting from zero-knowledge', in CRYPTO 88 ed. S. Goldwasser, *LNCS Lecture Notes in Computer Science*, Vol. 403, pp.216–231, Springer Verlag.

Her, Y.S. and Abe, M. (1998) 'Universally verifiable mix-net with verification work independent of the number of mix servers', in *Lecture Notes in Computer Science*, Vol. 1403, pp.437–447.

Her, Y.S., Imamoto, K. and Sakurai, K. (2005) 'E-voting system with ballot cancellation based on double encryption', in *Preproc. of International Workshop on Information Security Applications 2005*, Vol. 1, pp.525–532.

Laskari, E.C., Meletiou, G.C. and Vrahatis, M.N. (2004) 'Recent approaches to electronic data gathering with privacy', in *Proceedings of the 1st International Conference From Scientific Computing to Computational Engineering*, Athens, Greece.

Laskari, E.C., Meletiou, G.C., Tasoulis, D.K. and Vrahatis, M.N. (2005) 'Privacy preserving electronic data gathering', *Mathematical and Computer Modelling*, Vol. 42, pp.739–746.

Ogata, W., Kurosawa, K., Sako, K. and Takatani, K. (1997) 'Fault tolerant anonymous channel', in Information *and Communications Security – First International Conference*, Vol. 1334, pp.440–444, Springer Verlag.

Ohkubo, M., Miura, F., Abe, M., Fujioka, A. and Okamoto, T. (1990) 'An improvement on a practical secret voting scheme', in *Proceedings of the Second International Workshop on Information Security*, in M. Mambo and Y. Zheng (Eds.): *Lecture Notes in Computer Science*, Vol. 1729, pp.225–234, London.

Schnorr, C. (1991) 'Efficient signature generation by smart cards', *Journal of Cryptology*, Vol. 4, pp.161–174.

Stadler, M. (1996) 'Publicly verifiable secret sharing', in *Advances in Cryptology – EUROCRYPT'96*, *Lecture Notes in Computer Science*, Vol. 1070, pp.190–199.

Appendix

A short explanatory Appendix for cryptographic terms in this paper:

- public key: publicly available cryptographic key for encrypting a message (i.e., information)
- private key: key for decrypting an encrypted message from the corresponding public key, available only to its owner
- signature: cryptographic scheme for associating a message to the identity of an entity, demonstrating the authenticity of that information
- blind (signature): a form of signature in which the content of the message is encrypted before being signed
- anonymous network: a network in which it is infeasible to trace the identity of an entity from the message it sent through that network.