



Privacy Preserving Electronic Data Gathering

E. C. LASKARI

Computational Intelligence Laboratory, Department of Mathematics, University of Patras
GR-26110 Patras, Greece
and

University of Patras Artificial Intelligence Research Center (UPAIRC)
University of Patras, GR-26110 Patras, Greece
elena@math.upatras.gr

G. C. MELETIOU

A.T.E.I. of Epirus, P.O. Box 110
GR-47100 Arta, Greece
and

University of Patras Artificial Intelligence Research Center (UPAIRC)
University of Patras, GR-26110 Patras, Greece
gmelet@teiep.gr

D. K. TASOULIS

Computational Intelligence Laboratory, Department of Mathematics, University of Patras
GR-26110 Patras, Greece
and

University of Patras Artificial Intelligence Research Center (UPAIRC)
University of Patras, GR-26110 Patras, Greece
dtas@math.upatras.gr

M. N. VRAHATIS

Computational Intelligence Laboratory, Department of Mathematics, University of Patras
GR-26110 Patras, Greece
and

University of Patras Artificial Intelligence Research Center (UPAIRC)
University of Patras, GR-26110 Patras, Greece
vrahatis@math.upatras.gr

Abstract—Most organizations, from private business to scientific institutes, own large information databases. Analysis of these databases can be very beneficial to the owner organizations, as it contributes to knowledge discovery. To extract broader and more accurate conclusions, knowledge discovery techniques need to be applied on a collection of databases of different organizations involved in the same field. This paper addresses two issues associated with electronic data gathering: confidentiality of the organization that supplies a particular database, and authentication of the provided data. © 2005 Elsevier Ltd. All rights reserved.

Keywords—Privacy, Electronic data gathering, E-voting protocols, Discrete logarithm.

We would like to thank the editor and the reviewers for their concern. Moreover, we acknowledge the partial support by the “Archimedes” research programme awarded by the Greek Ministry of Education and Religious Affairs and the European Union.

1. INTRODUCTION

Nowadays, business and scientific organizations collect data in great magnitude and organize them in databases. Analysis of these databases can be very beneficial, as it leads to knowledge discovery. To this end, considerable effort has been devoted to the development of methods that contribute to knowledge discovery, using data mining techniques [1]. Rules derived from a single database capture cognition that is embedded in the specific data. These rules may be sufficient for the needs of a single organization but several factors may limit the global scope of these rules. For example, cognition derived from the database of a hospital may be restricted to people living in a narrow region, with specific environmental impacts, specific nutrition habits, etc. Applying knowledge discovery techniques on the collection of databases of all the organizations that are active in a particular area is likely to produce much broader knowledge.

Discovery of broader and more accurate knowledge on fields like medicine, education, economics, and others, is very important for each organization individually, but it is also important for the field as a whole. Ideally, to extract this knowledge one would require the collection and analysis of data originating from all organizations on the field. However, this is not always possible and the organizations are in general unwilling to expose their data, due to issues related to privacy and confidentiality.

In this paper, we consider the scenario where several organizations want to electronically gather all their data in a common database, which will be announced to all of them. Furthermore, the data have to be collected in such a way that the corresponding source organizations will not be identified from the data. In the rest of the paper, organizations will be referred to as *Alices*, following a convention widely used in cryptography. The security requirements that must be satisfied by the proposed scheme for electronic data gathering with privacy, are the following.

1. *Completeness*: All valid data are gathered correctly if all Alices follow the protocol.
2. *Privacy*: It is infeasible to associate individual databases to the Alices.
3. *Eligibility*: Only legitimate Alices are allowed to send data.
4. *Authentication*: Each Alice sends valid data.
5. *Verifiability*: Each Alice is able to verify whether her data are correctly included in the aggregated data.

All aforementioned requirements with the exception of the forth, are identical to the ones needed for secure voting system on a public network (for example, see [2–6]). The requirement imposed by some e-voting systems of just one vote per voter is not necessary in our case. Thus, to meet the aforementioned objectives we need to modify e-voting protocols to meet the additional requirement of authentication of transferred data without the restriction of just one package submission. The proposed scheme considers the existence of a third party, namely, the *Collector*, who is responsible for gathering the data of all legitimate Alices and for the distribution of the aggregation of the data back to all Alices. It also requires a verifier for each data record. The details of this approach are discussed in Section 2. Security analysis of the proposed scheme is given in Section 3 and complexity issues are described in Section 4. Conclusions and outlines for further research are given in Section 5.

2. E-GATHERING THROUGH MODIFICATION OF E-VOTING PROTOCOLS

The proposed approach is a modification of recent e-voting protocols to ensure privacy preserving in electronic data gathering. Alices are organizations working on the same field, such as different hospitals, pharmaceutical companies, banks, schools, etc. The Collector could be either a national authority or an organization, like the World Health Organization, a manager or a director of an organization, or even a machine. The Collector holds a list of the identities of all Alices that will participate in the electronic data gathering.

Each record of a database is assumed to be a vector of values (numerical or categorical) of the form $\bar{w}_j = (w_{j1}, w_{j2}, \dots, w_{jm})$. Each Alice must send all $\bar{w}_1, \bar{w}_2, \dots, \bar{w}_k$ to all other Alices via the Collector.

Each record \bar{w}_j has to be accompanied by a digital signature verifying its validity. This digital signature must come from an expert on the field, e.g., a doctor for medical data or a bank manager for financial data, called *Expert*. Each Alice, as an organization, may employ more than one Expert. The Experts' digital signature is implemented using a hash function H . The record \bar{w}_j is signed as

$$\text{Sign}_{\text{EU}}(\bar{w}_j) = D_{\text{EU}}(H(\bar{w}_j)),$$

where D_{EU} is the private key of the Experts Union (common for all Experts) and Sign_{EU} denotes the common signature of all Experts. Thus, each Expert forms the pairs,

$$[\bar{w}_j, \text{Sign}_{\text{EU}}(\bar{w}_j)] = A_j,$$

for all his records and passes them to Alice, keeping a copy for his own archive. Alice verifies the Expert's Union signature on every record and then she signs for the Expert the number of total records, denoted by nor , he hands over to her. The Expert verifies Alices signature on nor .

When each Expert completes this process, he sends to the Collector a message containing,

- (a) the number nor of records;
- (b) Alice's signature on nor ; and
- (c) the Experts' Union signature on nor , i.e., a triple of the form,

$$[\text{nor}, \text{Sign}_A(\text{nor}), \text{Sign}_{\text{EU}}(\text{nor})].$$

The Collector verifies the Experts' Union signature on every triple and deduces from the triples the total number of expected records. The Collector sets a deadline for the receipt of all the data and also selects a pair (D_0, E_0) of private and public keys, announcing E_0 to all Alices. The public key E_0 is used by all Alices for the encryption of all packages A_j to $z_j = E_0(A_j)$. Subsequently, each Alice blinds z_j as

$$e_j = \text{Blind}(z_j, r_j),$$

where r_j is a randomly selected blinding factor [7]. Alice also signs e_j as $\text{Sign}_A(e_j)$ using her signature. The triple $[Id_A, e_j, \text{Sign}_A(e_j)]$, containing Alice's identity, her blinded data, and her signature, is sent to the Collector by Alice. The Collector verifies Alice's identity and her signature and checks if Alice is registered on the list of participants. If this is true, he signs e_j as $\text{Sign}_C(e_j)$ and sends it back to Alice.

Alice verifies the Collector's signature over her data, she unblinds $\text{Sign}_C(e_j)$, and derives the Collector's signature on z_j , that is $\text{Sign}_C(z_j)$. All the pairs $[z_j, \text{Sign}_C(z_j)]$ of signed records are separately transmitted to the Collector through an *anonymous channel* [3,8–10]. The Collector verifies his signature and considers z_j as a valid data record.

When the deadline is reached and all Alices have sent their data, the Collector puts the aggregation of the encrypted data, along with his signature on them, to a list as pairs,

$$[z_j, \text{Sign}_C(z_j)],$$

and announces them to all Alices.

Each Alice verifies that her data are correctly included in the list by the z_j that originated from her. Finally, the Collector announces his private key D_0 to all Alices. Alices decrypt all the z_j deriving all A_j . Then, they verify each A_j by the Experts' Union signature, and perform data cleaning by discarding from the list the nonverified data (if such exist). From the verified A_j , Alices retrieve the \bar{w}_j and start data processing.

The procedures involved in the proposed scheme can be summarized in the following steps.

- Phase 1: Initialization and Expert Validation
 - (a) The Collector announces the deadline for the receipt of all data and a public key E_0 .
 - (b) Each Expert signs with the Experts' Union signature each valid data record and passes it to the corresponding Alice. Alice verifies the Experts' Union signature on every record and signs the nor for the Expert. The Experts verify Alices' signature on nor .
 - (c) The Expert then sends to the Collector the exact number of records nor he has signed for Alice, along with her signature and the Experts' Union signature on nor . The Collector verifies the Experts' Union signature on the nor 's and accepts them.
 - (d) Each Alice encrypts the pairs of records and the corresponding signatures using the public key E_0 .
- Phase 2: Blind Signing and Data Collection (for each Alice)
 - (e) Alice blinds the encrypted data, she signs them using her private key, and sends them along with her identity to the Collector.
 - (f) The Collector verifies Alice's signature and identity, and accepts the package.
 - (g) The Collector signs the data and sends it back to Alice.
 - (h) Alice verifies the Collector's signature, she unblinds the data and sends them back to the Collector in form of separate records through an anonymous channel.
 - (i) The Collector identifies his signature on the encrypted data and accepts them.
- Phase 3: Data Distribution and Verification
 - (j) After the deadline and the receipt of all data, the Collector posts all the encrypted data with his signature on them, in a list and sends them to all the Alices.
 - (k) Each Alice verifies that her data are correctly included in the list.
- Phase 4: Keys Disclosure and Data Cleaning
 - (l) The Collector releases the private key D_0 to all Alices.
 - (m) Each Alice performs decryption of all the data and data cleaning by discarding any unverified records. In conclusion, all valid data are exposed to all Alices.

3. SECURITY ANALYSIS

Next, we analyze the security of the proposed scheme by considering various possible attacks. The following cases are examined.

ALICE IS AN ACTIVE CHEATER.

- (i) 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 Collector can prove which Alice hasn't send data by checking the list of participants and the corresponding identifications Id_A he has received at Step (f). So, he can ask Alice to provide him with his signature $Sign_C(e_j)$. Notice that Alice cannot forge his signature.
- (ii) Suppose that one Alice sends her data at Step (e), gets the Collector's signature but she does not return the unblinded data back to the Collector at Step (h). After the expiration of the deadline the Collector checks the packages of the form,

$$[nor, Sign_A(nor), Sign_{EU}(nor)],$$

that he has received from the Experts, and computes the sum,

$$s = \sum_A nor.$$

Then, s , which is the number of expected data, is compared with the number of received data through the anonymous channel. If the numbers do not agree, he does not release

his private key and announces to all Alices that they have offended the regulations and also the exact number of data that have not been send. The Collector announces a final deadline for the receipt of missing data. If the total number of records is not sent up to the new deadline, then the offender Alice can be disclosed by the Experts. The same case holds if Alice doesn't send all her data.

- (iii) Suppose that Alice sends invalid data at Step (h) in order to mislead all the other Alices. These data will not be signed by the Expert (only original data records are signed at Step (b) with the Experts' Union signature Sign_{EU}), thus, the data will not be verified at the last step. Invalid data will be discarded from the list and only valid data will remain. Also, the malicious Alice can be immediately exposed by the Experts. Thus, the data authentication criterion is fulfilled.
- (iv) On Step (d) Alice encrypts her packages using the Collector's public key. The usual cryptosystems for e-voting schemes are based on the El Gamal cryptosystem. For the El Gamal cryptosystem the encryption of \bar{w}_j is

$$(g^k \bmod p, E_0^k (\bar{w}_j \bmod p)),$$

where k is a random key, p is a prime number, and g is the generator of \mathbb{Z}_p^* . Thus, if we suppose that Alice or Oscar (an intruder) wants to derive the Collector's private key D_0 , he has to solve the equation,

$$E_0 = g^{D_0} \bmod p,$$

which is identical to solving the discrete logarithm problem. If he wants to decrypt a specific message encrypted by an Alice, he has to deduce

$$E_0^k = g^{D_0 k} \bmod p,$$

from the quantities,

$$E_0 = g^{D_0} \bmod p, \quad \text{and} \quad g^k \bmod p.$$

Solving this problem is as difficult as solving the Diffie-Hellman problem [11].

THE COLLECTOR IS A PASSIVE CHEATER.

It is infeasible for the Collector or an intruder to associate individual records to Alices at Step (f), since each encrypted record is blinded by Alice. It is also infeasible for the Collector to detect the source of unblinded data since he receives them in separated records through an anonymous channel. Thus, privacy is ensured.

THE COLLECTOR IS AN ACTIVE CHEATER.

- (i) Communications at Steps (e) and (g) are registered submissions, since they are both signed by the sender. This means that the sender cannot cheat. Specifically, in Step (g), we suppose that the Collector tries to abuse Alice by sending her a signed but fault package $\text{Sign}_C(e'_j)$. Then, the verification at Step (h) of the Collector's signature over Alice's data fails and Alice can prove the fraud.
- (ii) At the verification phase the Collector's private key is not revealed until each Alice verifies that her data are correctly included in the list. If Alice's data z_i are not included in the list then Alice can force the Collector to sign again all data in the list. By publishing through an anonymous channel the signature of the Collector $\text{Sign}_C(z_i)$ on her encrypted data at Step (h), she can prove that for all signed encrypted data $\text{Sign}_C(z_j)$ in the list, it holds that $\text{Sign}_C(z_j) \neq \text{Sign}_C(z_i)$. In the same way, if Alice's data in the list are distorted, she can announce through an anonymous channel that some (or all) of her records are

distorted, and verify this fact by publishing her encrypted data signed by the Collector ($[z_i, \text{Sign}_C(z_i)]$) that she gets at Step (g). For all other Alices it is trivial to check if

$$[z_i, \text{Sign}_C(z_i)] \neq [z_j, \text{Sign}_C(z_j)],$$

for all the Alice's i records and all j records in the list. Thus, verifiability is secured.

- (iii) The Collector can try to mislead Alices by adding invalid data in the list. However, Alices will identify these invalid data since it will not have the Experts signature and it will be discarded at the verification phase.

4. COMPLEXITY ISSUES

For the computation of the complexity of the proposed electronic data gathering protocol, we consider as a measure the number of

1. encryptions and decryptions,
2. signatures and verifications,
3. sendings,
4. blindings and unblindings.

We assume that each one of these operations bears a computational cost equal to 1.

Let e be the number of all Experts, n the number of all Alices and $\sum_A \text{nor}$ the sum of all records that will be gathered through the process. Then, the computational cost of each step of the e-gathering protocol, described in Section 2, is

- 1: (a) none;
- 1: (b) at most $2 \sum_A \text{nor}$ signatures and $2 \sum_A \text{nor}$ verifications;
- 1: (c) e signatures + e sendings + e verifications;
- 1: (d) $\sum_A \text{nor}$ encryptions;
- 2: (e) $\sum_A \text{nor}$ blindings + $\sum_A \text{nor}$ signatures + $\sum_A \text{nor}$ sendings;
- 2: (f) $\sum_A \text{nor}$ verifications;
- 2: (g) $\sum_A \text{nor}$ signatures + $\sum_A \text{nor}$ sendings;
- 2: (h) $\sum_A \text{nor}$ verifications + $\sum_A \text{nor}$ unblindings + $f(\sum_A \text{nor}, k)$;
- 2: (i) $\sum_A \text{nor}$ verifications;
- 3: (j) n sendings;
- 3: (k) none;
- 4: (l) n sendings;
- 4: (m) $\sum_A \text{nor}$ decryptions.

Thus, the computational cost of the scheme is

$$\mathcal{K} = 15 \sum_A \text{nor} + 3e + 2n + f\left(\sum_A \text{nor}, k\right),$$

where the first three terms measure the computational cost of the gathering process, while the term $f(\sum_A \text{nor}, k)$ is the computational cost due to the use of an anonymous channel of communication with k MIXes.

The anonymous channel with k MIXes was initially proposed by Chaum in [3]. Its computational cost for m senders and only one receiver, is computed to be

$$z = 3m \times (k + 1).$$

Improvements of Chaum's protocol for anonymous channels with k MIXes have been proposed in [9,10]. The computational complexity associated with the use of these protocols is of the same order.

In the proposed scheme, we use an anonymous channel for n Alices (senders) with $\sum_A \text{nor}$ records and one receiver, the Collector. In this case, m corresponds to $\sum_A \text{nor}$, since each record is transmitted separately. Thus, the previous result for z becomes

$$z^* = f\left(\sum_A \text{nor}, k\right) = 3 \sum_A \text{nor} \times (k + 1).$$

Combining these results, the total computational cost of the proposed scheme is, at most,

$$\mathcal{K}_{\text{total}} = 15 \sum_A \text{nor} + 3e + 2n + \left(3 \sum_A \text{nor} \times (k + 1)\right).$$

This cost can be significantly reduced if a proper grouping of the records, is used in the communications, and will be considered in a future work. (Note that the term proper refers to a grouping that preserves the privacy of the Alices involved.) Proper grouping of the records can also be used in the simple sendings at Steps (e) and (g) of the proposed scheme, thereby reducing further the overall computational cost of the scheme. Finally, note that the number e of the Experts cannot exceed $\sum_A \text{nor}$.

5. CONCLUSIONS AND FUTURE RESEARCH

A modification of recently proposed e-voting systems for privacy preserving electronic data gathering is proposed. This modification includes an additional type of entities, namely Alices, which need not be trusted. Alices are the organizations that employ the Experts, which in turn correspond to the voters of the e-voting system. The data Collector recognizes the Experts' Union signature and Alices' identities, without being able to associate the databases to the corresponding Alices.

Most relevant studies on data collection do not consider the privacy of the senders of the data, or the authentication of the data. Significant effort has been devoted to address the issue of processing data that are distributed to many sources [12–14] but these approaches do not consider electronic data gathering.

The proposed scheme is a first approach that satisfies all five requirements of completeness of scheme, privacy and eligibility of Alices, authentication of data and verifiability of data, that are necessary for electronic data gathering with privacy.

In a future work we intend to study alternative approaches for the implementation of electronic data gathering with privacy, including distributed e-gathering without a collector. Moreover, combinations of the proposed scheme with other anonymous channel and e-voting protocols will be considered. The application of data mining in this setting need also be investigated. Finally, we would like to point out that the proposed scheme can be extended to other related and interesting applications, like e-census with privacy, which will also be considered in a future work.

REFERENCES

1. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, (1996).
2. C. Boyd, A new multiple key cipher and an improved voting scheme, *Lecture Notes in Computer Science* **434**, 617–625, (1989).
3. D.L. Chaum, Untraceable electronic mail, return addresses, and digital pseudonyms, *Communications of the ACM* **24** (2), 84–90, (1981).
4. R. Cramer, R. Gennaro and B. Schoenmakers, A secure and optimally efficient multi-authority election scheme, *Lecture Notes in Computer Science* **1233**, 103–118, (1997).
5. A. Fujioka, T. Okamoto and K. Ohta, A practical secret voting scheme for large scale elections, *Lecture Notes in Computer Science* **718**, 244–251, (1992).
6. B. Schoenmakers, A simple publicly verifiable secret sharing scheme and its application to electronic voting, *Lecture Notes in Computer Science* **1666**, 148–164, (1999).

7. D.L. Chaum, Blind signature for untraceable payments, In *Advances in Cryptology, Proceedings of CRYPTO '82*, (Edited by D. Chaum, R.L. Rivest, and A.T. Sherman), pp. 199–203, Plenum, New York, (1982).
8. M. Abe, Universally verifiable mix-net with verification work independent of the number of mix-servers, *Lecture Notes in Computer Science* **1403**, 437–447, (1998).
9. W. Ogata, K. Kurosawa, K. Sako and K. Takatani, Fault tolerant anonymous channel, *Lecture Notes in Computer Science* **1334**, 440–444, (1997).
10. C. Park, K. Itoh and K. Kurosawa, Efficient anonymous channel and all/nothing election scheme, *Lecture Notes in Computer Science* **765**, 248–259, (1994).
11. W. Diffie and M.E. Hellman, New directions in cryptography, *IEEE Transactions on Information Theory* **IT-22** (6), 644–654, (1976).
12. R. Agrawal and R. Srikant, Privacy-preserving data mining, In *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 439–450, ACM Press, (2000).
13. Y. Lindell and B. Pinkas, Privacy preserving data mining, *Journal of Cryptology* **15** (3), 177–206, (2002).
14. J. Vaidya and C. Clifton, Privacy preserving association rule mining in vertically partitioned data, In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Canada, (2002).