

RECENT APPROACHES TO ELECTRONIC DATA GATHERING WITH PRIVACY

Elena C. Laskari^{**‡}, Gerasimos C. Meletiou^{†‡}, and Michael N. Vrahatis^{**‡}

^{*} Dept. of Mathematics, University of Patras, GR-26110 Patras, Greece

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

[‡] University of Patras Artificial Intelligence Research Center (UPAIRC),
University of Patras, GR-26110 Patras, Greece

e-mail: elena@math.upatras.gr, web page: <http://www.math.upatras.gr/~elena>
gmelet@teiep.gr

vrahatis@math.upatras.gr, web page: <http://www.math.upatras.gr/~vrahatis>

Keywords: electronic data gathering, privacy, anonymity, cryptographic protocols

Abstract. *The spread of database technologies has allowed corporations and scientific organizations to collect huge amounts of data. Analysis of the contents of these databases can contribute to knowledge discovery. Discovery of broader and more accurate knowledge in areas such as medicine, education, biology, economics, and engineering among others, is very important. Extraction of this knowledge requires the collection and analysis of data originating from several, ideally all, organizations on the field. In this paper, we propose a protocol for electronic data gathering that ensures precise data gathering and provides privacy for each contributing organization. The proposed protocol does not require a Central Tabulating Facility and proceeds solely with the participants. For completeness purposes a review on an alternative protocol of ours for electronic data gathering that makes use of a Central Tabulating Facility is also given.*

1 INTRODUCTION

The spread of database technologies has allowed corporations and scientific organizations to collect data in previously inconceivable magnitudes. It has been widely recognized that the appropriate analysis of the contents of these databases can contribute to knowledge discovery. Data mining is a collection of mathematical tools, designed to extract knowledge directly from data. The conclusions derived from the application of data mining techniques on a single database, reflect cognition that is embedded in the specific data. These rules may be sufficient for the needs of a single organization, but they cannot be considered broad knowledge, due to various limitations. Consider for example, the cognition derived from the database of a hospital. The validity of the extracted conclusions may be restricted to people living in a narrow region, with specific environmental conditions, nutrition habits, etc. The discovery of broader and more accurate knowledge in areas such as medicine, education, biology, economics, and engineering, among others, is very important, both for each organization individually, as well as for the development and progress of the field as a whole. Extraction of this knowledge requires the collection and analysis of data originating from several, ideally all, organizations on the field. However, organizations are not always willing to expose their data, due to issues related to privacy, anonymity and competition.

We consider a scenario where several organizations want to electronically gather their data in a common database and share it among them. The data have to be collected in a way that ensures the anonymity of the corresponding organizations. Furthermore, each organization must be assured that no participating organization can cheat by not providing any, or by providing corrupt data, and still get an opportunity to acquire the gathered data. In principle, the security requirements that must be met for data gathering with privacy protocol are the following:

- a. **Completeness:** All valid data are gathered correctly if all organizations follow the protocol.
- b. **Privacy:** Infeasibility of associating individual databases to organizations.
- c. **Eligibility:** Only legitimate organizations are allowed to send data.
- d. **Authentication:** Each organization sends valid data.
- e. **Verifiability:** Each organization is able to verify whether its data are correctly included in the

aggregated data.

In a previous work, we proposed a protocol for privacy preserving electronic data gathering, based on modern e-voting protocols, which make use of a Central Tabulating Facility^[1]. For completeness purposes, a review on this protocol is given in Section 2. As an extension of this work, in the present paper, induced by poll free e-voting protocols, a protocol for data gathering with privacy without the use of a Central Tabulating Facility is proposed. The description of the e-voting protocol without a poll is given in Section 3.1, and in Section 3.2 the details of the proposed data gathering scheme are discussed. Security analysis of the proposed scheme is given in Section 3.3 and complexity issues are described in Section 3.4. Conclusions and directions for future research are given in Section 4.

2 RELEVANT WORK

In a relevant work about electronic data gathering protocols^[1], we have considered the existence of a third party, namely the *Collector*, who is responsible for gathering the data of all legitimate organizations and for the distribution of the aggregation of the data back to all organizations, as a Central Tabulating Facility. 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 organizations that will participate in the 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 organization must send its records $\bar{w}_1, \bar{w}_2, \dots, \bar{w}_k$ to all other organizations 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*. The difference is that each 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 $D_{EU}(H(\bar{w}_j)) = \text{Sign}_{EU}(\bar{w}_j)$, where D_{EU} is a common private key of the Experts' Union and Sign_{EU} denotes the common signature of all Experts. Thus, each Expert forms the pairs $[\bar{w}_j, \text{Sign}_{EU}(\bar{w}_j)] = A_j$ for all his records and passes them to the organization, keeping a copy for his own archive. The organization signs for the Expert the number of the total of records, denoted by *nor*, he hands over to her. When each Expert completes this process, he sends, through an *anonymous channel*^[2], to the Collector a message containing: (a) the number *nor* of records; (b) the organization's signature on *nor*; and (c) the Experts' Union signature on *nor*. The Collector sets a deadline for the receipt of all the data. The Collector also selects a pair (D_0, E_0) of private and public keys and announces E_0 to all organizations. The public key E_0 is used by all organizations for the encryption of all packages A_j to $z_j = E_0(A_j)$. Subsequently, each organization blinds z_j as $e_j = \text{Blind}(z_j, r_j)$, where r_j is a randomly selected blinding factor^[3]. Each organization also signs e_j as $\text{Sign}_{org}(e_j)$ using its signature. The triple $[Id_{org}, e_j, \text{Sign}_{org}(e_j)]$, containing the organization's identity, its data, and its signature, is sent to the Collector by the organization. The Collector verifies the organization's identity and its signature and checks if it 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 the organization. The organization verifies the Collector's signature over its data, it un-blinds $\text{Sign}_C(e_j)$, and derives the Collector's signature on z_j , that is $\text{Sign}_C(z_j)$. The pair $[z_j, \text{Sign}_C(z_j)]$ is sent to the Collector through an anonymous channel. The Collector verifies his signature and considers z_j as a valid data record.

When the deadline is reached and all organizations 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 organizations. Each organization verifies that its data are correctly included in the list by the z_j 's that originated from it. Finally, the Collector announces his private key D_0 to all organizations. The organizations decrypt all the z_j 's deriving all A_j 's. Then they verify each A_j by the Experts' Union signature, and perform data cleaning by discarding from the list any non-verified data. From the verified A_j 's the organizations retrieve

the \bar{w}_j 's and start data processing.

Security analysis of the above scheme has shown that the organizations need not be trusted and the Collector of the data recognizes the Experts' Union signature and the organizations' identities, without being able to associate the databases to the corresponding organization. Other relevant studies to the direction of data collecting do not usually consider the privacy of the senders of the data or the authentication of the data. Significant work has also been done on processing data that are distributed to many sources^[4,5,6] but these approaches do not consider data gathering.

3 ELECTRONIC DATA GATHERING PROTOCOL WITHOUT A COLLECTOR

3.1 Poll free e-voting protocol

Nowadays, electronic voting is widely studied, as traditional elections are characterized by several deficiencies; they are time and work consuming, voters need to move and swarm to voting centers, etc. To this aim several e-voting protocols for secure elections have been proposed^[7,8,9,10]. Most of these protocols make use of at least one Central Tabulating Facility (CTF). An e-voting scheme without a CTF, where the voters watch each other, has been designed by Merritt *et al.*^[11,12] and it is briefly described below.

Assume that each voter, i , where $i = 1, \dots, N$, has a public and a private key, and every voter knows the public keys of all other voters.

- 1) Each voter chooses his vote and follows the procedure below^[13]:
 - a) He attaches a random string, R_0 , to his vote, V .
 - b) He successively encrypts his vote along with the attached string, using the public keys of voters 1 through N , in that order, producing $E_N(E_{N-1}(\dots(E_1(V, R_0))\dots))$, where E_j is the encryption function that uses the j -th voter's public key.
 - c) He repeats Step b, including this time a random string within each layer of encryption. Thus, at this point the vote becomes $E_N(R_N, E_{N-1}(\dots(R_2, E_1(R_1, E_N(E_{N-1}(\dots(E_1(V, R_0))\dots))))\dots))$, where R_i , $i = 1, \dots, N$, are random strings. The voter saves the intermediate results after each successive encryption. These results will be used later in the protocol to confirm that his vote is among the counted votes.
- 2) Each voter sends his encrypted message to the N -th voter. He decrypts all the votes with his private key and removes all the random strings at that level. Then, he scrambles the order of all the votes and sends the result to the $(N-1)$ -th voter. This procedure is repeated for all the voters from N to 1 . Thus, only the inner encryptions remain, and each vote looks like: $E_N(E_{N-1} \dots (E_1(V, R_0)) \dots)$.
- 3) Subsequently, the voter N decrypts all the votes with his private key, verifies that his vote is among the set of votes, signs all the votes, and sends the results to all other voters. The $(N-1)$ -th voter verifies and deletes the N -th voter's signature. He verifies that his vote is among the set of votes, signs the votes and sends them to all other voters. This procedure is repeated for all voters from N to 1 . Partway through this step the votes look like: $Sign_{i+1}(E_i \dots (E_1(V, R_0)) \dots)$, where $Sign_i$ is the signature of voter i .
- 4) All voters verify the signature of voter 1 , and they ensure that their vote is among the set of votes by looking for their random string among the votes. Then, everyone removes the random strings for each vote and tallies the votes.

This protocol works and is also self-adjudicating. The number of voters remains fixed throughout the process, thus, ballot stuffing and dropping is easily detected. Votes cannot be replaced by a malicious party, since any attempt of replacement can be discovered, as reported in^[13]. The scrambling of votes provides anonymity, and the inner random string, R_0 , allows participants to ensure that their vote is in the final tally. However, three problems arise with this scheme. The first and most important problem is that the protocol requires an enormous amount of computations. Its complexity limits its practical implementation to a small number of people and it would never work in a real election^[13]. Second, voter 1 learns the result of the election before anyone else does, although he cannot affect the outcome. The third problem is that voter N can copy anyone else's vote, even though he does not know its content beforehand. This could be a problem if we consider a three-person election,

since the result of the election will be identical to the copied vote. Therefore, it must hold that $N \geq 4$, with N relatively small.

3.2 The proposed protocol

We modify the prescribed e-voting protocol for the purposes of electronic data gathering, as follows: first, we divide the participating organizations in groups of proper size, such that each group comprises of at least four organizations, and the total number of groups is divided by a number different than 2. This partitioning scheme ensures both security and simplicity, as it will be analyzed in detail later on. This partitioning can also meet specific requirements, such as regional partitioning etc.

As in the protocol described in Section 2, each record \bar{w}_j has to be accompanied by a digital signature, which verifies its validity. This digital signature comes from the Expert in the form of D_{EU} , which is a common private key of the Experts' Union. The common signature of all Experts is denoted again as $Sign_{EU}$. Thus, each Expert forms the pair $[\bar{w}_j, Sign_{EU}(\bar{w}_j)] = A_j$ for each of his records, and passes them to the corresponding organization, keeping a copy for his own archive. The organization signs for the Expert the number of the total of records, nor , that he has passed to it. When each Expert completes this process, he sends to all other organizations, a message containing: (a) the number, nor , of records that he has passed to the organization; (b) the organization's signature on nor ; and (c) the Experts' Union signature on nor . Thus, all organizations are aware of the total number of records that each organization must send, as well as the total number of records.

At the first level, each organization follows the e-voting protocol as described in Section 3.1, applying all the four steps. Step 1 is applied separately on each pair A_j of the organization. Thus, at Step 2, all voters get the total number of their group's encrypted records. If the number of received records does not agree with the number of expected records, then the protocol is interrupted. Steps 3 and 4 are applied subsequently. At the end of the first level, the total of all records of the groups are privately gathered.

At the second level, one agent from each group (or the whole group as a unit) is responsible for further gathering of the data among the different groups. Each group announces its agent to all other organizations by sending in public a message with the agent's identity, signed by all the members of the corresponding group. The determination of the agent of each group can be done also prior to the data gathering processes. Super-groups comprised of the agents are defined, and the prescribed e-voting protocol is applied again. From this level up, Step 1 is not applied to each record individually, but to packages of records of predefined size, in order to reduce the complexity of the scheme. Definition of the proper packages' size is described later on. If remaining records exist in the sub-groups, these are encrypted and sent individually (one by one). Steps 2 to 4 are applied subsequently. The second level procedure is applied iteratively, until all records of all organizations are gathered together. The assembled database is announced in public for all participating organizations. Each organization can be assured of the authenticity of the database by verifying the signatures of the last agents (agents of the upper level), which are attached to the records of the database. The aforementioned procedures are summarized in the following steps:

Phase 1: 1st Level procedures

- (a) All participating organizations are divided into groups of appropriate size.
- (b) Each Expert signs with the Experts' Union signature each valid data record; he passes each signed record to the corresponding organization and receives its signature, verifying the number of records he provided.
- (c) The Expert sends to all organizations the exact number, nor , of records that he has signed for the specific organization, along with its signature, as well as the Experts' Union signature on nor .
- (d) Each organization follows Steps 1 to 4 of the poll free e-voting protocol, distributing its records to his group. At Step 1, the organization encrypts separately its pairs A_j . At the end of this phase, each group has gathered all its data records.

Phase 2: 2nd Level procedures

- (e) An agent for each group is defined.
- (f) Agents are divided into super-groups of proper size.
- (g) The proper size of record packages to be distributed subsequently among the agents of each super-group is computed and announced to each super-group.
- (h) Each agent follows Steps 1 to 4 of the poll free e-voting protocol, distributing the records of his sub-group to his super-group. At Step 1, the agent encrypts the records of his sub-group in packages of size defined at Step (g). If remaining records exist, the agents encrypt them individually and send them after the packages.
- (i) Phase 2 is repeated until all records of all organizations are gathered together.

As we have already mentioned, the rules for partitioning the participating organizations into groups are imposed by security and simplicity considerations. Initial groups require at least four participants in order to overcome the problem of the e-voting scheme of copying each other's files. Though, the number of participants in each group should not be too large, since it increases the complexity of Phase 1. From level 2 of data gathering and onward, a number of groups that is different than 2 is needed, to avoid revealing the origin of records.

At the 1st Level procedures, the scheme demands that data records are encrypted separately by each organization. This is necessary to ensure that no knowledge concerning the origin of any record can be gained. Thus, privacy of the organizations is preserved. From the 2nd Level and onward, this demand is not so pressing, as all group's data records are gathered together and no one can link a specific data record with its initial origin. Thus, the agents can encrypt data records in packages of predefined size, which cannot be identified as they are shuffled. This is done in order to reduce the complexity of the proposed scheme. The size is defined for each group by a different integer, m (modulo), such that a reasonable number of record packages is created for each member of the group, and either all members of the group have remaining records, or none of the members of the group has remaining records. Remaining records, if they exist, are encrypted separately one by one. For example, let x_1, x_2, x_3, x_4 , be the number of records of four members of a group, respectively. Then, an integer m is selected, such that $x_i = \lambda_i m + \mu_i$, $i = 1, \dots, 4$, where λ_i is the number of record packages of member i , and either all μ_i 's are equal to zero or all μ_i 's are nonzero. The proposed scheme does not use anonymous channels. All submissions come from a specific sender to a specific receiver, and they are authorized. Furthermore, protocols for "secret selling secrets" and identification tags (or alias), which are usually used in e-voting schemes, are not needed in our case.

3.3 Security analysis

Considering various possible attacks, the security analysis of the proposed electronic data gathering scheme is investigated below. In contrast to the e-voting schemes, in the case of data gathering, the organizations have limited cheating motivation, since any interruption of the protocol may imply heavy penalties on the responsible organization, as well as termination of its collaboration with the other organizations. In all cases of interruption of the protocol, the Experts can be called to reveal the responsible entity.

Organization is an active cheater:

- (1) Assume that one organization intends to cheat the other organizations in order to see their data without submitting any of its data, or by submitting just a part of it. In this case, at step (d) of the 1st Level, and specifically at the inner Step 2 of the voting protocol, the total number of the corresponding group's received records, and the total number of expected records, will not agree. Thus, the procedure will be interrupted before ending; the cheating organization will not obtain the decrypted data of the group.
- (2) Assume that an organization copies one of its valid records and sends it several times, instead of its other records, in order to cheat the other organizations. In this case, at the end of the 1st Level, there will exist several identical records in the group. Since each valid record is signed by the Experts' Union, it can be ensured that no identical signed records will exist. Thus, the procedure will be interrupted. Notice that, the uniqueness of each record can also be assured by the attachment of a random number to it by the Expert.
- (3) Suppose that an organization copies another organization's records. Then at the inner Step 2 of the voting protocol (i.e. at step (d) of the 1st Level), there will exist identical encrypted messages passed to the organizations. This is not acceptable and the procedure will be interrupted. Furthermore, the cheating organization will not obtain any decrypted data.
- (4) Suppose that an organization sends invalid data in order to mislead the other organizations. Then, at the end of the 1st Level, the missing signatures of the Experts' Union will identify the invalid data.

Agent is an active cheater:

- (1) Assume that one agent intends to cheat the other organizations in order to see their data without submitting any of its data or by submitting just a part of it. In this case, at step (h) of the 2nd Level, the total number of the corresponding super-group's received records, and the total number of expected records, will not agree. Thus, the procedure will be interrupted and his group can expose the cheater.
- (2) Assume that an agent copies some of his group's valid records or packages, and sends it instead of the real records, in order to cheat the other agents. Then, at the end of the 2nd Level, there will exist identical records and his group can expose the cheater agent.
- (3) If an agent copies another agent's records, then there will exist identical encrypted messages at the inner step 2) of the voting protocol (step (h) of the 2nd Level), and the protocol will be interrupted by the other agents before the decryption of the records. Thus, the agent will not obtain any of their records and his group can expose him.

- (4) Suppose that an agent sends invalid data in order to mislead the other agents. Then, at the end of step (h), the missing signatures of the Experts' Union will identify the invalid data records, the protocol will be interrupted, and his group can expose the cheater agent.

Oscar is an active cheater:

- (1) Oscar, a malicious party, attempts to pass invalid records to the organizations. Oscar's signature at step (d) will not be verified as a signature of a legitimate organization and the records will be discarded. In general, all submissions come from a specific sender to a specific receiver and are authorized. Thus, any malicious party can be discovered immediately.

3.4 Complexity Issues

Let us compute the number of operations in the gathering process. As a complexity measure we consider the number of: encryptions, decryptions, signatures, verifications and sendings. We assume that all the above operations have the same computational cost, equal to 1.

The Merritt election protocol for N voters, as it is described in^[13], requires:

1. $2N^2$ encryptions,
2. N sendings,
3. N^2 decryptions + N^2 sendings,
4. N^2 decryptions + N^2 signatures + $N^2(N-1)$ sendings + $N^2(N-1)$ verifications.

Therefore, the total number of operations is

$$\#(operations) = 2N^3 + 4N^2 + N = f(N). \quad (1)$$

Assume that each voter is an organization and we deal with N organizations. Let each organization have x_i records. Then, the total number of records is $x = x_1 + \dots + x_N$, and the total number of operations is,

$$\frac{x}{N} f(N). \quad (2)$$

Consider the election protocol and let the voters be grouped in groups of size n , and the groups further grouped in groups of n elements of higher order, etc. Thus, we derive a "tree" with s levels, with number of voters equal to $N = n^s$. We apply the protocol described in Section 3.2, and we obtain the following number of operations:

$$sn^{s-1} f(n) = \log_n(N) \times N \times \frac{f(n)}{n}. \quad (3)$$

Finally we combine the previous results; assuming that we have $N = n^s$ organizations, and each organization has x_i records, with $x = x_1 + \dots + x_N$. The number of operations is:

$$s \times x \times \frac{f(n)}{n}. \quad (4)$$

Finally let us assume a "tree" with s levels, where n is not constant. Let also n_{\max} denote the maximum number of branches per node. According to Section 3.2, $n_{\max} \approx 5$. Then, $N \leq n_{\max}^s$ and,

$$\#(operations) \leq sx \frac{f(n_{\max})}{n_{\max}}, \quad (5)$$

where, in general, $\frac{f(k)}{k} = 2k^2 + 4k + 1$ is a polynomial of second degree.

4 CONCLUSIONS AND FUTURE RESEARCH

A modification of a poll free e-voting protocol for secure privacy preserving electronic data gathering is proposed. The proposed scheme meets all five requirements of completeness of scheme, privacy of organizations, eligibility of participating organizations, authentication and verifiability of data that are necessary for electronic data gathering with privacy. An alternative privacy preserving electronic data gathering protocol that makes use of a Collector and also satisfies the requirements is reviewed.

Future work will consider the application of data mining algorithms in the proposed setting. Finally we would like to point out that the proposed scheme could be further extended to applications such as e-Census with privacy. We intend to propose such an application in a future work.

ACKNOWLEDGMENT

We acknowledge the partial support by the “Archimedes” research programme awarded by the Greek Ministry of Education and Religious Affairs and the European Union.

REFERENCES

- [1] Laskari, E.C., Meletiou, G.C., Tasoulis, D.K., Vrahatis, M.N. (2004), “Privacy preserving electronic data gathering”, *Mathematical and Computer Modelling*, accepted for publication.
- [2] Abe, M. (1998), “Universally verifiable mix-net with verification work independent of the number of mix-servers”, *Lecture Notes in Computer Science*, Springer-Verlag, Vol. 1403, pp. 437—447.
- [3] Chaum, D.L. (1982), “Blind signature for untraceable payments”, *Proceedings of CRYPTO'82, Advances in Cryptology*, D. Chaum, R. L. Rivest, and A. T. Sherman Eds., Plenum, NY, pp. 199—203.
- [4] Agrawal, R., Srikant, R. (2000), “Privacy-preserving data mining”, *Proceedings of the ACM SIGMOD Conference on Management of Data, Texas*, ACM Press, pp. 439—450.
- [5] Lindell, Y., Pinkas, B. (2002), “Privacy preserving data mining”, *Journal of Cryptology*, Vol. 15, pp. 177—206.
- [6] Vaidya, J., Clifton, C. (2002), “Privacy preserving association rule mining in vertically partitioned data”, *Proceedings of the Eighth ACM SIGKDD International Conference, Canada*.
- [7] Chang, C., Wu, W. (1997), “A secure voting system on a public network”, *Networks*, Vol. 29, pp. 81—87.
- [8] Cramer, R., Gennaro, R., Schoenmakers, B. (1997), “A secure and optimally efficient multi-authority election scheme”, *Lecture Notes in Computer Science*, Vol. 1233, pp. 103—118.
- [9] Karagiannopoulos, M.G., Meletiou, G.C., Vrahatis, M.N. (2004), “A note on a secure voting system on a public network”, *Networks*, Vol. 4, pp. 224—225.
- [10] Schoenmakers, B. (1999), “A simple publicly verifiable secret sharing scheme and its application to electronic voting”, *Lecture Notes in Computer Science*, Vol. 1666, pp. 148—164.
- [11] DeMillo, R., Lynch, N., Merritt, M. (1982), “Cryptographic protocols”, *Proceedings of the 4th ACM Symposium on the Theory of Computing*, pp. 383—400.
- [12] DeMillo, R., Merritt, M. (1983), “Protocols for data security”, *IEEE Computer*, Vol. 16, pp. 39—50.
- [13] Schneier, B. (1996), *Applied Cryptography*, John Wiley and Sons, Inc., New York.