Private and Resilient Data Aggregation

Mathieu Cunche, Cédric Lauradoux, Marine Minier

Private and Resilient Data Aggregation

Mathieu Cunche, Cédric Lauradoux, Marine Minier

RESEARCH REPORT

N° 8330

July 2013

Project-Team Privatics

ISSN 0249-6399   ISRN INRIA/RR--8330--FR+ENG

Private and Resilient Data Aggregation

Mathieu Cunche, Cédric Lauradoux, Marine Minier

Project-Team Privatics

Research Report n° 8330 — July 2013 — 17 pages

Abstract: Sensors are commonly deployed in hostile environment, and consequently a number of research works have focused on data aggregation schemes designed to be tolerant to attacks on sensor nodes. In parallel, schemes ensuring the confidentiality of sensor data have been proposed to address the emerging privacy concerns. We note that resilience against tampering attacks requires access to the sensor node's data, while in privacy-preserving systems this data must remain confidential. In this work, we aim to reconcile these two seemingly conflicting objectives. We present a novel private and resilient aggregation system, in which an aggregator combines the data collected from sensor nodes and forwards the resulting sum to an analyst. Our scheme protects the privacy of the users from both honest-but-curious aggregator and analyst, while enabling the filtering of fake data values using a Private Range Test protocol.

Key-words: resilience, privacy, data aggregation, sensor networks

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Agrégation de données résiliente et privée

Résumé : Les réseaux de capteurs peuvent être déployés dans un environnement hostile. Ainsi un nombre de travaux de recherche se sont intéressés à des systèmes d'agrégation de données tolérant aux attaques sur les noeuds. Parallèlement des techniques garantissant la confidentialité des données collectées par les réseaux de capteurs ont été proposées afin de faire face à la problématique de vie privée. La tolérance aux attaques sur les capteurs nécessite un accès aux données retournées par ceux-ci, alors que la protection de la vie privée nécessite justement que ces données restent confidentielles. Le but de ce travail est de réconcilier ces deux objectifs qui apparaissent comme conflictuels. Nous présentons un nouveau système d'agrégation capable de tolérer les attaques sur les noeuds tout en préservant la confidentialité des données des capteurs. Ce système inclut un agrégateur qui collecte et combine les données provenant des capteurs et renvoi le résultat à l'utilisateur final appelé analyste. Ce système protège la vie privée des utilisateurs face à un couple agrégateur/analyste "curieux mais honnête". Il permet également de données les données contrôlées par un attaquant en utilisant un protocole de "Private Range Test" basé sur la théorie du calcul sécurisé multipartie.

Mots-clés : résilience, vie privée, agrégation de données, réseaux de capteurs

Contents

1 Introduction

Sensor networks and smart metering systems are becoming increasingly popular. The payload of these networks is often highly sensitive for the users and the absence of security or privacy features has enabled researchers to demonstrate significant privacy leaks in such networks [2, 25]. However, any security and privacy solution has to consider the potentially conflicting goals of the *analyst* and the *end user.* First, the *analyst* expects to receive *useful data* from a group of deployed *sensors.* Therefore, it should be possible to link the sensors to the data they are providing to the system or, as a minimum requirement, a set of sensors controlled by the attacker should not be able to alter the quality of the information delivered to the analyst. The *end-users* wish to protect their privacy and avoid *profiling* and *de-anonimization*, therefore they expect the analyst to have access only to coarse grained information. Our work aims to reconcile the requirements of both parties, in the context of sensor networks performing *data aggregation.*

Data aggregation enables a trade-off between communication and computation costs, and is often used in sensor networks to save energy or improve capacity. The values of the sensors are combined by an *aggregator*, and the result is then sent to the analyst for further processing. Numerous works [10, 26, 28, 24] have targeted independently the resilience or the privacy aspects of data aggregation schemes, but very few have attempted to solve both problems at the same time. In our work, resilient aggregation ensures the robustness of the aggregation result in the presence of fake sensor readings. Indeed, nodes are deployed in an hostile environment and, in addition to potentially being faulty, are vulnerable to attacks. An adversary that controls a node can send fake readings to the aggregator and influence the result to their advantage. A solution supported by many research works consists of verifying that the reported values belong to a predetermined valid range. This is potentially highly intrusive, if it is done on plaintext data.

In parallel, solutions to preserve privacy using homomorphic encryption have been proposed. Here, the aggregator does not have access to plaintext sensor data and aggregation is performed by combining encrypted sensor data values. However, this step in the (right) direction of privacy prevents the aggregator from inspecting the data to determine whether it is in the valid range. *This paper addresses the challenge of enabling the verification of the range of encrypted data values prior to using them in the aggregation.*

We consider a flat sensor network architecture, in which all sensors are directly connected to an aggregator that collects and combines the received values, before sending the aggregate to an analyst. The sensors can be controlled by an attacker and send fake values to influence the final result. Our contribution focuses on aggregation functions based on the sum operation.

Our contributions are as follows. We propose a *system that provides private and resilient data aggregation*, based on additive homomorphic encryption and Private Range Test. The former enables data aggregation while ensuring data confidentiality. Private Range Test allows the aggregator to test if an encrypted data lies in a given interval, without having to access the plaintext or gaining any additional information other than the result of the test. In the proposed system, the values reported by the sensors are individually verified using the Private Range Test. We then propose an optimization of the system to reduce the verification cost by the Aggregated Private Range Test (effective for scenarios with a small number of compromized nodes, i.e. up to 17), however with some reduction of accuracy. We have implemented and tested the Private Range Test protocol, demonstrating the practicality of our scheme with a Range Test taking only 108$msec$ to execute.

The paper is organized as follows. Section 2 defines our goal for private and resilient aggregation and the model of adversary. The core idea of our protocol is given in Section 3. To overcome the potentially high complexity of the first protocol, a second scheme implementing compromise between accuracy and complexity is presented in Section 4. Practical considerations are discussed in Section 5. Other possible solutions are summed up in Section 6 whereas Section 7 gives the position of this work in the literature. Finally, Section 8 concludes this paper.

2 System goals and the security model

We are considering a system including: a *set of n sensors*, an *aggregator* in charge of aggregating the data received from the sensors and an *analyst* that must obtain the sum $y = \sum_{i=1}^{n} x_i$ where the x_i denote the values measured by the sensors. At each round, every sensor transmits a unit of data to the aggregator which in turn transmits the aggregate to the analyst. The system architecture is shown in Fig. 1.

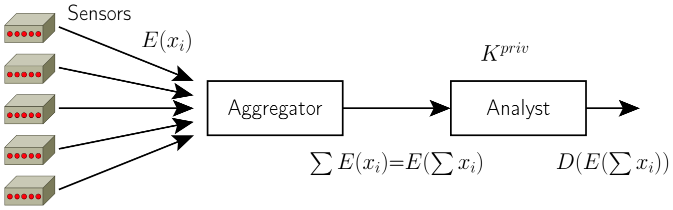

Figure 1: System architecture comprising an aggregator and an analyst.

Most systems providing privacy preserving aggregation include an aggregator located between the data sources and the analyst. Its role is to aggregate the data received from the data sources and forward the result, y, to the analyst. This architecture has been adopted in several privacy-preserving aggregation solution [1, 13, 11] as it allows the distribution of aggregation and decryption tasks to two distinct entities, thus limiting the risk of a privacy breach.

A report is considered as *bogus* (resp. *genuine*) when its value is different from (resp. equal to) the real value.

The proposed system aims to satisfy the following objectives:

- *Privacy* - The internal and external attackers should not be able to access the data of individual sensors.

- *Resilience* - The system should be tolerant to bogus data sent by the compromised or faulty sensors. Even in the presence of bogus data, the system must still be able to compute the aggregate with a small error margin.

- *Complexity* - Computational cost for the elements of the system should be as low as possible. In particular the operations performed by the sensors should be in line with the limited computation resources of embedded systems.

We make the following hypothesis. First, there is an interval $[0,..,q]$, call the *valid interval*, in which should fall all *genuine* data measured by the sensors. This is for instance the case of a number of physical measurements like temperature inside a building, or energy consumption of a household. Thanks to the homomorphic property of the encryption scheme any interval can be shifted to an interval of the form $[0,..,q]$. Reports falling in the valid interval are said to be *valid*, and reports that do not fall in this interval are said to be *invalid*. We assume that a *genuine* report is always valid, and that a *bogus* report can be either *invalid* or *valid* (a value different from the real value but that remains in the valid range) .

Second, the values x_i can be controlled by an attacker, called the *correctness attacker*, whose goal is to control the aggregated value y. This assumption was the motivation of the seminal of Wagner [28] and it was later used in subsequent works [6]. As nodes are very cheap devices, they are not assumed tamper-resistant: they are an obvious target for the adversary to mount an attack. Let y be the result of the aggregation without the action of an attacker and $\hat{y}$ the aggregation result after the action of the attacker. The error induced by the attack is denoted: $\Delta = |y - \hat{y}|$. The goal of the attacker is to maximize the error Δ. To counter this attacker, we aim at designing a system that limits the impact of the attack on the result, *i.e.* minimize Δ.

Third, there is an attacker, called the *privacy attacker* whose goal is to obtain information on the sensors' data. This attacker correspond to the aggregator and the analyst that are assumed to be *honest-but-curious* adversaries, i.e. they are willing to execute correctly the communication protocol but they are interested in getting information on the individual values x_i. Therefore, the aggregator is not allowed to manipulate the values x_i in the clear. We also assume that the aggregator and the analyst are not colluding.

To the best of our knowledge, this is the first time that these two attackers (the *correctness attacker* and the *privacy attacker*) are merged to in the analysis of a data aggregation system.

Finally, we assume that all the cryptographic materials (keys, encryption algorithms, etc.) are already loaded into the sensors. We assume that this step is done through secure channel or other methods. We also assume that there exist secure channels between the aggregator and the analyst.

3 Invalid data filtering using Range Test

In this section, we propose a new private and resilient aggregation scheme. The privacy core of our design comes from the use of a partially homomorphic public key cryptosystem as in [23]. We are using an additive homomorphic encryption scheme; we are therefore limited to aggregation functions based on the addition such as the sum or the average.

The resilience of our scheme relies on the private range test protocol proposed in [21, 22]. It filters out invalid values while preserving their confidentiality. The private range test RT is a two party protocol in which two non-colluding parties A_1 and A_2 can verify if a ciphertext c lies in a given range $[0, .., q]$ without revealing the plaintext. A short description of the range test protocol proposed in [22] is made in Appendix A. This protocol requires two homomorphic public key cryptosystems, $E_1(.), D_1(.)$ and $E_2(.), D_2(.)$, which private keys are held by the analyst. Only the first cryptosystem is used by the sensors to encrypt their data. The second cryptosystem is employed by the aggregator and the analyst for the range test protocol. Let us denote $RT(A_1, A_2, c, [0, .., q])$ the invocation of a range test protocol between A_1 and A_2. The function returns TRUE if c belongs to $[0, .., q]$ and FALSE otherwise. In the rest of the paper, A_1 refers to the aggregator and A_2 to the analyst.

Our system is presented in Fig. 2 and uses the following protocol composed of four steps:

1. **Data encryption:** each sensor takes the sensed value x_i, encrypts it with the analyst public key and an homomorphic cryptosystem into $c_i = E_1(x_i)$. Then each sensor sends the c_i value to A_1.

2. **Data-filtering:** The aggregator A_1 uses the range test $RT(A_1, A_2, c_i, [0, .., q])$ to filter out the invalid reading received during the previous step. During this phase the aggregator A_1 collaborates with the data analyst A_2, in order to perform the range test and to know if each value c_i belongs to the interval $[0, .., q]$ (see Algorithm 1).

3. **Aggregation:** The aggregator A_1 sums the valid values thanks to the homomorphic properties of the cryptosystem and sends the aggregated result to the analyst A_2.

4. **Result decryption:** The analyst A_2 decrypts the aggregated result received from the aggregator A_1. This result is the sum of the valid values transmitted by the sensors.

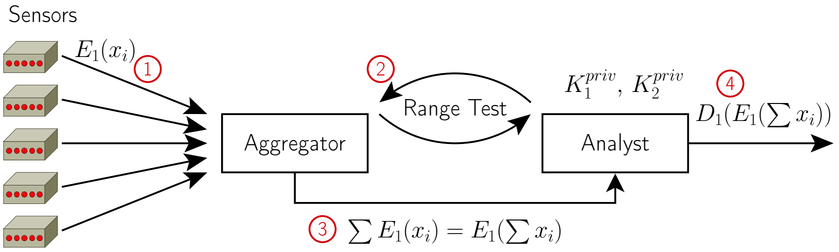

Figure 2: Private and resilient aggregation with range test.

During the protocol, the individual values are tested but never revealed to the analyst neither to the aggregator thanks to the range test (see details in Appendix A). They only know if individual values lie on a given interval or not.

Algorithm 1 Data Filtering

```
Require: $[0..q], \{c_1, \dots, c_n\}$
Ensure: $\forall c \in V,\ c \in [0, .., q]$
  $V \leftarrow \varnothing$
  for each $c \in \{c_1, \dots, c_n\}$ do
    if $\mathrm{RT}(A_1, A_2, c, [0, .., q]) = \mathsf{TRUE}$ then
      $V \leftarrow V \cup \{c\}$
    end if
  end for
```

4 Aggregated Range Test

4.1 Protocol description

As noted before, the verification step puts the aggregator and the analyst under a heavy computational load as they are performing a range test on each ciphertext. We propose to reduce the number of range tests by performing a partial aggregation before the verification step. The ciphertexts are divided and aggregated by subgroups before being submitted to the range test on a larger interval. Each subgroup is submitted to a *q-homogeneous range test* (see Definition 1); i.e. a subgroup composed of r values $\{x_i\}_{1 \leq i \leq r}$ is tested on the interval $[0, .., rq]$ (rather than on $[0, .., q]$ for the basic approach).

Definition 1 *Let $\{x_i\}_{1 \leq i \leq r}$ be r values, $\{\lambda_i\}_{1 \leq i \leq r}$ be r integers, $q \in \mathbb{Z}$ and let $Agg = \sum_{1 \leq i \leq r} \lambda_i x_i$ be the corresponding sub-aggregate. We say that the double inequality $\alpha \leq Agg \leq \beta$ is q-homogeneous iff, $\alpha = a.q$, $\beta = b.q$ and $\beta - \alpha = \Lambda q$, where $\Lambda = \sum_{1 \leq i \leq r} \lambda_i$, and a and b are two integers. The associated range test $RT(A_1, A_2, E_1(Agg), [\alpha, .., \beta])$ is said q-homogeneous.*

The complete protocol works as follows:

1. **Data encryption:** each sensor takes the sensed value x_i, encrypts it with the first homomorphic cryptosystem into $c_i = E_1(x_i)$ and transmits it to the aggregator A_1.

2. **Partial aggregation:** The aggregator A_1 aggregates together r received ciphertexts: $Agg = E_1(x_1) \times \cdots \times E_1(x_r) = E_1(x_1 + \cdots + x_r)$.

3. **Data-filtering:** The aggregator A_1 collaborates with the analyst A_2 in order to perform the range test $RT(A_1, A_2, Agg, [0, .., rq])$ on the partial sum testing if the Agg value belongs to the interval $[0, .., rq]$ or not.

   - If the Agg result belongs to $[0, .., rq]$, then the current aggregate is marked as valid.
   - If not, two strategies are possible: the aggregator A_1 completely discards the partial sum Agg or it narrows down the investigation in order to identify and discard invalid reports until a valid partial sum is obtained (see Subsection 4.3 for more details).

4. **Aggregation:** The aggregator A_1 combines the valid partial sums thanks to the homomorphic properties of the first cryptosystem and sends the aggregated result to the analyst A_2.

5. **Result decryption:** The analyst A_2 decrypts the aggregated and encrypted result received from the aggregator A_1. This result is the sum of the values comprised in the valid partial sum.

One advantage of the aggregated range test is that the cost of the aggregation operation is much smaller than the cost of a range test and it could be re-used for the final aggregation.

4.2 Bounds on the aggregate error

This technique reduces the global number of range tests (from n calls to the range test to n/r calls to the range test), but it also reduces the efficiency of the filtering, as an invalid value can pass the aggregated range test. However we can derive bounds on the error that an attacker can cause as a function of the number of nodes controlled by the attacker, the number of sensors, n, and the aggregation subgroup size, r.

Lets consider a subgroup of size r, in which the *correctness attacker* controls one element. Note that the reasoning would be the same if the attacker wanted to minimize the aggregate value. The worst case is when the $r-1$ genuine values are equal to 0. The maximum value that the compromised sensor can send without being detected is qr. The attacker can therefore add an error of $\delta = q.(r-1)$ in each aggregation subgroup. Since the reports of a subgroup are aggregated before the test, it does not matter whether if the error comes from a single sensor or from multiple ones. In other words, controlling multiple values per subgroup would not increase the maximum error that the *correctness attacker* could induce without being detected. In a system of size n, an attacker controlling at least one value in b subgroups can induce a global error on the sum of amplitude $\Delta = b.q.(r-1)$. The corresponding error on the average is $\Delta' = \Delta/n = b.q.(r-1)/n$.

If the attacker has managed to compromise at least one sensor in each subgroup, then $b = n/r$, the error on the sum is $\Delta = n/r.q.(r-1) = n.q.(r-1)/r$ and the error on the average is $\Delta' = q.(r-1)/r$. The amplitude of this error could be considered as too large (it is of the order of the expected aggregate). However, controlling at least one element per subgroup can be a difficult task for the attacker, especially when those subgroups are small. If the attacker is able to choose the compromised nodes, it requires n/r compromised sensors to achieve the previous errors. The size, r, of the subgroup can be adapted according to the expected power of the attacker. In addition the subgroup can be randomly created at each round, making impossible for the attacker to adapt the values sent to the aggregator according to the identity of the nodes composing each subgroup.

4.3 Subgroup testing strategy

The efficiency of our system is highly dependent on the strategy adopted for the aggregated test. Our objective is to minimize the verification cost while maximizing the resilience of our system. We propose a strategy that ensures that the error is smaller than q with $2\log_2(r)$ range tests in average. Our proposed strategy can be divided into two parts:

- The set of ciphertexts is randomly partitioned into subgroups of equal size r.
- A dichotomic search is performed on each subgroup.

The dichotomic search is described in Algorithm 2. This algorithm implements a recursive search of the invalid value(s). Each step, a set of ciphertexts is submitted to a range test. If the result is positive, i.e. the output of the range test is TRUE, the function simply returns the set of ciphertexts. If the answer is negative and the set is not reduced to one element, this set is recursively divided into two subsets of equal sizes and each subset is in turn tested with a q-homogeneous range test. If the answer is negative, and the subset is reduced to one element, then the ciphertext is discarded by returning the empty set.

Algorithm 2 Dichotomic search

```
Require: $q$,$Agg$
  $r \leftarrow \#Agg$
  $t \leftarrow RT(A_1, A_2, Agg, [0, .., rq])$
  if $t =$TRUE then
    return $Agg$
  else
    if $s > 1$ then
      $Agg_1, Agg_2 \leftarrow$ Split($Agg$)
      $Agg'_1 \leftarrow$ Search($Agg_1$)
      $Agg'_2 \leftarrow$ Search($Agg_2$)
      return $Agg'_1 \cup Agg'_2$
    else
      return $\{\varnothing\}$
    end if
  end if
```

Note that this algorithm does not discard all the invalid reports. It is possible that invalid data remain in the subgroup, if their error's amplitude is low enough to not be detected by the aggregated range tests. However is this case, the induced error will be small as we will show in the following paragraph.

4.3.1 Bounds on the error

This algorithm ensure that for each subgroup, the set of r' ciphertexts returned by the search procedure does not contain an error with an amplitude larger than $q.r'$. Indeed, according to Theorem 1, the search procedure only returns valid sets, i.e. sets Agg' of r' elements such as $RT(A_1, A_2, Agg', [0, .., r'q]) =$ TRUE.

Theorem 1 *Let Agg' be an aggregated set of ciphertexts returned by the search function. Then $RT(A_1, A_2, Agg', [0, .., r'q])$ is always TRUE.*

Proof The proof is straightforward by recursion.

Since the error in each subgroup is bounded by qr', then the global error is bounded by $(qr')(n/r) \leq qn$.

4.3.2 Complexity

The complexity of this strategy varies depending on the number, the amplitude and the positions of the invalid values. For any subgroup of size r, at least one range test needs to be performed, but up to a maximum of $2r - 1$ range test may be required, depending on the execution of the dichotomic search algorithm on a subgroup of size r.

A subaggregate of r values can be represented by a binary tree with r leaves (one for each value), and a depth of $\log_2(r)$. Each non-leaf node corresponds with an aggregated value. The dichotomic search can be seen as an exploration of this tree. If a node fails to pass the range test, the algorithm pursues by exploring the corresponding subtree; otherwise, if the node passes the range test, the corresponding subtree is not explored. This tree is composed of $2r - 1$ nodes; therefore, the maximum number of range tests to perform is $2r - 1$. To attain this maximum, the number of invalid data must be at least $r/2$ (in order to require a full depth and width exploration). Another interesting case is when there is a unique invalid data and the first range test return FALSE. In this case a search following one branch is performed until the invalid leaf is identified. During this in-depth search, 2 nodes are explored for each of the $\log_2(r)$ level of the tree. This results in the execution of $2\log_2(r)$ range tests.

If we summarize for the full set, the minimal cost is n/r, the maximum cost is $(n/r)(2r-1) = 2n - n/r$, and if there is at least one corrupted data per subgroup the cost is $(n/r)\log_2(r)$. In the worst case scenario, the aggregated range test approach is more costly than the simple approach, but this requires that at least half of the data are corrupted. If a random subgroup selection strategy is adopted this would require that the attacker control half sensors, which would correspond to a rather powerful attacker.

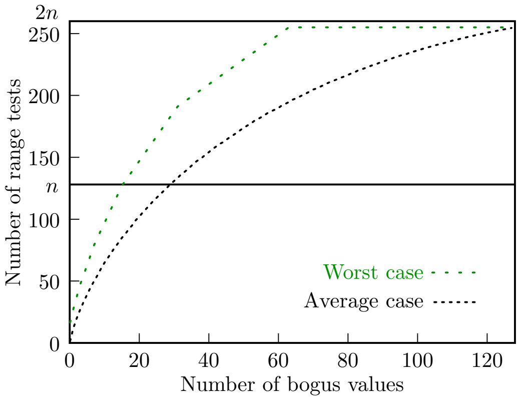

Figure 3: Comparison of the direct application of range test versus the aggregated range test for $n = 128$. The average results are given by simulation of 1000 verifications.

Fig. 3 shows the cost in terms of range tests as a function of the fraction of invalid values for a system comprising 128 sensors. For the worst case scenario, if the fraction of invalid values is lower than 17, then the aggregated approach is more efficient, otherwise the standard approach with a constant cost is more efficient. In the average case, the aggregated approach is better for a number of invalid values lower than 27.

4.4 Data privacy

In the aggregate range test protocol, the aggregator obtains information on linear combination of the data. Thus the aggregator could combine the outcomes of those tests to gain information on individual values. We show that if the aggregator follows a specific rule, i.e. the *homogeneous range tests rule*, then it gains no more information than any range test provided on individual values.

Lemma 2 *Any linear combination of q-homogeneous double inequalities is also q-homogeneous.*

Proof The proof is straightforward considering first multiplication by a constant and second combination of two inequalities.

Theorem 3 *Let* $\{x_i\}_{1\leq i\leq n}$ *be* n *values,* $\{a_j\}_{1\leq j\leq m}$ *be* m *sub-aggregate and* ${T_j}_{1\leq j\leq m}$ *the corresponding range test. If all the* T_j *are q-homogeneous range test, then* $\forall\, i$*, the output of the* m *range test* T_j *does not provide more information on* x_i *than the output of any range test* $RT(x_i, [mq, .., (m+1)q])$ *where* $m \in \mathbb{Z}$.

Proof Let $\mathcal{L}$ be the linear system representing the output of the $T_{j_{1\leq j\leq m}}$ range tests. If those inequalities are all q-homogeneous, then any linear combination is also q-homogeneous. In particular for any inequality $a \leq x_i \leq b$, therefore we have $a = c.q$, $b = d.q$ and $b - a = q$ and this result correspond to a q-homogeneous range test on the value x_i.

According to Theorem 3, if all the range tests performed by the aggregator and the analyst are following the *homogeneous range tests rule* then any combination of those range tests provide the same information that a single range test on this value.

5 Complexity

Up to our knowledge, this is the first time that private range tests are applied in practice, and there is currently no official implementation of these tests. In Table 1, we give a first evaluation of the computational cost of a range test in term of number of operations. As encryption/decryption operations are much more expensive than addition and multiplication (due to the use of underlying modified El-Gamal cryptosystem), we sum up the computational complexity in terms of encryption/decryption operations. A test has a significant cost: our approach to reduce the number of test execution is therefore justified.

Table 1: Computation cost (encryption and decryption) for one value.

	Enc.	Dec.
Data producer	1	-
Aggregator (A_1)	$2+18$	-
Analyst (A_2)	8	24

To know how practical is our scheme, we have implemented a range test. Our implementation used the El-Gamal cryptosystem with 1024-bit keys. The implementation was done using GMP 5.0.2 and GCC version 4.6.3 on an Intel(R) Core(TM) i7 processor. For $t = 1$, a range test is executed in 108ms.

For the data producer (*e.g.* sensors), the main complexity is the implementation of El-Gamal cryptosystem. The elliptic curves El-Gamal cryptosystem has been already investigated for sensor in [27] and is fully feasible. It also benefits from the support of TinyECC [18].

6 Other solutions

Other solutions to perform private and resilient data aggregation are possible. A first solution consists to limit the range of the plaintext by choosing an appropriate encryption scheme. The ciphertexts domain can match the desired range. Encrypting data on a field $\mathbb{F}_q$ ensures that the aggregated values belong to the range $[0, .., q-1]$. However, this range may not include the result of the aggregation (sum) leading to an overflow or an invalid result. If n values lying on $[0, .., q-1]$ are to be aggregated, the aggregation must be performed in the field $\mathbb{F}_{qn}$ to ensure the result's correctness. Before the aggregation, each encrypted value in $\mathbb{F}_q$ must be plunged in the field $\mathbb{F}_{qn}$. Performing this operation without breaching the confidentiality of the data requires the use of secure multiparty computation (SMC) [30]. The corresponding SMC protocols are associated with high complexity operations that involve the sensors themselves. They are not suitable for devices with limited resources.

Another possible solution is based on cryptographic accumulators [3, 12]. Accumulators allow to test whether or not an element belongs to a group. Bloom filters [19] can be used to design accumulators. These accumulators rely on the one-wayness property of the underlying hash functions of a Bloom filter. More precisely, we could imagine a system where the nodes individually send their encrypted values $c_i = E(x_i)$ to the aggregator using a stream cipher which is an homomorphic symmetric key primitive as shown in [8]. Then, the aggregator asks the analyst to provide the corresponding Bloom filter containing the ciphering values of the interval $E(0), .., E(q)$ for each c_i. The aggregator then tests whether or not the received c_i value belongs to the Bloom filter, i.e. testing if the c_i value is in the valid interval. The aggregator computes the sum of the c_i that have correctly passed the previous test and sends the aggregated result to the analyst that could deciphers correctly the sum thanks to the homomorphic property of the underlying primitive and because the analyst shares with each sensor the corresponding secret key. As previously presented, this solution could also be adapted using subgroup testing for more efficiency. However, this solution remains costly in terms of number of exchanged bits between the aggregator and the analyst: the size of the Bloom filter increases linearly according to the size of the considered interval and one filter must be sent per value received by the aggregator.

Finally, one could use Zero Knowledge Proofs (ZKP) protocols, such as [4], to prove that sensor values belong to a given interval. In those ZKP protocols, a prover proves to a verifier that it holds a value lying in a given range. A such ZKP protocol could be applied to our problem by assigning to each sensor the role of prover, while the aggregator would play the role of verifier. However ZKP are generally costly in term of computations and communications. For instance, in the protocol proposed in [4] would require that each sensor compute 20 exponentiations per value (against only one in our scheme). ZKP-based solution therefore appear much more costly for the sensors that our solution based on private range test.

7 Related work

The issue of corrupted or faulty sensors sending bogus reports has been considered in multiple works. Wagner was the first to introduce the problem of resilience against faulty sensors in aggregation systems [28]. In addition to a nice formalization of the problem, he studied the intrinsic resilience of the common aggregation functions (average, min/max, median) and proposed several algorithms to further improve their robustness. Similarly, Buttyan *et al.* introduced a solutions [6] [7] capable of filtering outliers in sensor networks. Buttyan's schemes assume that the measurements follow a statistical model and use this information to detect and discard outliers before aggregation, thus creating a resilient aggregation system. As opposed to our solution, Buttyan's schemes require the hypothesis that the data follow a particular distribution, whereas our solution does not require any assumption on the data other than the fact that valid measurements lie in a limited range or set. Moreover, privacy is not considered in all these works.

A number of works have considered the problem of misbehaving or faulty nodes in tree-based aggregation systems deployed on wireless sensor networks. In [15], faulty or malicious behavior in the aggregation process are detected with the aid of cryptographic primitives. However, this solution only applies to aggregators inside the tree and cannot detect corrupted leaves (sensors). In SIA [24] the authors consider the problem of data aggregation in a wireless sensor network where nodes and aggregators can be compromised and may try to corrupt the aggregated data. They provide an efficient solution for the case of a corrupted aggregator. However, in case of corrupted sensors, their work relies on the intrinsic resilience of the aggregation function as in [28]. Contrary to our work, those resilient aggregation systems do not take into account the privacy of the sensed data. In fact one of their main requirements is that the data are available in clear in order to be inspected by various parties of the network. This contradicts our goal to preserve the confidentiality of users data.

Systems ensuring privacy have been built using homomorphic encryption [29, 13] or secure multiparty computation [5]. In those systems, each node encrypts or encodes its value before sending it to the analyst, either directly or through a set of intermediate forwarders and aggregators. In those works, the problem of resilience against malicious or faulty sensors is simply ignored because the encryption prevents any verification on the data.

A scheme to privately compute statistical queries over distributed databases has been proposed by Chen et. al. in [11]. This work focus on queries that can receive binary answer (TRUE or FALSE). Answers are encrypted by the data sources with the Goldwasser-Michali (GM) bit cryptosystem [14]. This cryptosystem provides semantically secure encryption of binary values. An aggregator is then in charge of forwarding and anonymizing the encrypted answers to the analyst. Similarly to our solution, the employed cryptosystem permits to check the validity of each ciphertext without learning the plaintext. Thanks to this feature, the aggregator is able to filter out the bogus answers while preserving data's privacy. We note that the solution we are proposing in this paper also use a ciphertext verification. As opposed to our problem, this work is limited to binary data and cannot deal with integers, since the GM cryptosystem is reduced to binary values. In addition, even if the GM cryptosystem is homomorphic, this feature is not exploited to perform data aggregation in the encrypted domain. In fact, the GM cryptosystem working on binary values, the homomorphic operations are reduced to the exclusive-or which is of limited use in our problem.

We note that resilience and privacy in aggregation systems have both been considered by [16] and [17]. However the robust aspect of this work only considers sensor or link failure. These solutions can tolerate sensors that *omit* or *fail* to send their data, but cannot tolerate sensors that send bogus data.

In [9], the authors introduce ABBA, a solution for secure aggregation in wireless sensor networks. The presented system supports a privacy preserving aggregation mechanism as well as an integrity verification that enable the sink to detect with a high probability the presence of bogus data in the aggregate. When the integrity check fails, the aggregated value must be discarded, and it is not possible to neither identify nor isolate the source(s) of the bogus data. To overcome this issue, one could adopt a dichotomic approach by requesting aggregate values of smaller sets of sensors in order to narrow down the faulty or compromised sensors. This would significantly reduce the privacy guarantees of the system, as the aggregated values would be computed from a smaller number of sensors, increasing the risk of information leak. In addition, such a dichotomic method would only be effective if the faulty/compromised sensors have a constant behavior, i.e. if they send bogus data at each round. Intermittent failures of some sensors or sporadic data corruption would be hard to detect with a dichotomic approach. This weakness could be exploited by an attacker to mount a Denial of Service (DoS) attack: ensuring that at least one bogus data is included in the aggregate value at each round would imply that all the aggregated values are discarded. As opposed to the system proposed in this work, ABBA relies solely on low-cost symmetric key cryptography. One could envision an hybrid system taking advantages of the strength of each scheme: the low complexity ABBA system could be sufficient when no or few bogus values are detected, and above a certain threshold of bogus values, a switch to our Range-Test based system could be employed.

8 Conclusion

We have presented a private and resilient data aggregation system based on a private range test. This system allows to preserve privacy of the data while verifying that an attacker controlling a subset of the sensors has a limited impact on the aggregated data. This approach could be compared with the notion of accountability of the end-users as done for example in [23], even if the verification is not done on individual values but on sub-aggregates.

We have considered in the paper bogus readings from the sensors which attempt to influence the sum obtained by the analyst. It would be interesting to consider a scenario in which some sensors collude with the analyst in order to threaten the privacy of a particular user. Application of our scheme to other aggregation functions such as the median, minimum and maximum would be worth investigating.

References

[1] Gergely Ács and Claude Castelluccia. I have a DREAM!: differentially private smart metering. In *International conference on Information hiding - IH'11*, Lecture Notes in Computer Science 6958, pages 118–132, Prague, Czech Republic, 2011. Springer-Verlag.

[2] Ross Anderson and Shailendra Fuloria. On the security economics of electricity metering. In *9th Annual Workshop on the Economics of Information Security, WEIS 2010, Harvard University, Cambridge, MA, USA, June 7-8, 2010*, 2010.

[3] Josh Cohen Benaloh and Michael de Mare. One-way accumulators: A decentralized alternative to digital sinatures (extended abstract). In *Advances in Cryptology - EUROCRYPT '93*, volume 765 of *Lecture Notes in Computer Science*, pages 274–285. Springer, 1993.

[4] Fabrice Boudot. Efficient proofs that a committed number lies in an interval. In *Proceedings of the 19th international conference on Theory and application of cryptographic techniques*, EUROCRYPT'00, pages 431–444, Berlin, Heidelberg, 2000. Springer-Verlag.

[5] Martin Burkhart, Mario Strasser, Dilip Many, and Xenofontas Dimitropoulos. Sepia: privacy-preserving aggregation of multi-domain network events and statistics. In *Proceedings of the 19th USENIX conference on Security*, USENIX Security'10, pages 15–15, Berkeley, CA, USA, 2010. USENIX Association.

[6] Levente Buttyán, Péter Schaffer, and István Vajda. RANBAR: RANSAC-based resilient aggregation in sensor networks. In *ACM Workshop on Security of ad hoc and Sensor Networks - SASN 2006*, pages 83–90, Alexandria, VA, USA, October 2006. ACM.

[7] Levente Buttyán, Péter Schaffer, and István Vajda. Resilient aggregation with attack detection in sensor networks. In *4th IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2006 Workshops), 13-17 March 2006, Pisa, Italy*, pages 332–336. IEEE Computer Society, 2006.

[8] Claude Castelluccia, Aldar C-F. Chan, Einar Mykletun, and Gene Tsudik. Efficient and provably secure aggregation of encrypted data in wireless sensor networks. *ACM Trans. Sen. Netw.*, 5(3):20:1–20:36, June 2009.

[9] Claude Castelluccia and Claudio Soriente. Abba: A balls and bins approach to secure aggregation in wsns. In *6th International Symposium on Modeling and Optimization in*

Mobile, Ad Hoc, and Wireless Networks and Workshops, WIOPT 2008, pages 185–191, Berlin, Germany, March 31 - April 4 2008.

[10] Aldar C.-F. Chan and Claude Castelluccia. A security framework for privacy-preserving data aggregation in wireless sensor networks. *TOSN*, 7(4):29, 2011.

[11] Ruichuan Chen, Alexey Reznichenko, Paul Francis, and Johannes Gehrke. Towards statistical queries over distributed private user data. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, NSDI'12, pages 13–13, Berkeley, CA, USA, 2012. USENIX Association.

[12] Nelly Fazio and Antonio Nicolisi. Cryptographic Accumulators: Definitions, Constructions and Applications. Technical report, New York University, 2002. Available at `http://www-cs.ccny.cuny.edu/~fazio/research.html`.

[13] Flavio D. Garcia and Bart Jacobs. Privacy-friendly energy-metering via homomorphic encryption. In *Proceedings of the 6th international conference on Security and trust management*, STM'10, pages 226–238, Berlin, Heidelberg, 2011. Springer-Verlag.

[14] Shafi Goldwasser and Silvio Micali. Probabilistic Encryption and How to Play Mental Poker Keeping Secret All Partial Information. In *ACM Symposium on Theory of Computing - STOC*, pages 365–377, San Francisco, CA, USA, May 1982. ACM.

[15] Lingxuan Hu and David Evans. Secure aggregation for wireless networks. In *Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT'03 Workshops)*, SAINT-W '03, pages 384–, Washington, DC, USA, 2003. IEEE Computer Society.

[16] Marian Kamal Iskander, Adam J. Lee, and Daniel Moss é. Privacy and robustness for data aggregation in wireless sensor networks. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 699–701, New York, NY, USA, 2010. ACM.

[17] Marek Jawurek and Florian Kerschbaum. Fault-tolerant privacy-preserving statistics. In Simone Fischer-Hübner and Matthew Wright, editors, *Privacy Enhancing Technologies*, volume 7384 of *Lecture Notes in Computer Science*, pages 221–238. Springer, 2012.

[18] An Liu and Peng Ning. Tinyecc: A configurable library for elliptic curve cryptography in wireless sensor networks. In *Proceedings of the 7th international conference on Information processing in sensor networks*, IPSN '08, pages 245–256, Washington, DC, USA, 2008. IEEE Computer Society.

[19] Michael Mitzenmacher. Compressed bloom filters. In *PODC '01: Proceedings of the twentieth annual ACM symposium on Principles of distributed computing*, pages 144–150, New York, NY, USA, 2001. ACM.

[20] Kun Peng, Colin Boyd, Ed Dawson, and Byoungcheon Lee. An Efficient and Verifiable Solution to the Millionaire Problem. In *Information Security and Cryptology - ICISC 2004*, Lecture Notes in Computer Science 3506, pages 51–66. Springer Berlin Heidelberg, Seoul, Korea, December 2005.

[21] Kun Peng, Colin Boyd, Ed Dawson, and Eiji Okamoto. A Novel Test. In Lynn Margaret Batten and Reihaneh Safavi-Naini, editors, *Australasian Conference on Information Security and Privacy ACISP 2006*, volume 4058 of *Lecture Notes in Computer Science*, pages 247–258. Springer, 2006.

[22] Kun Peng and Ed Dawson. A range test secure in the active adversary model. In *Australasian symposium on ACSW frontiers - ACSW '07*, pages 159–162, Ballarat, Australia, 2007. Australian Computer Society, Inc.

[23] Raluca A. Popa, Andrew J. Blumberg, Hari Balakrishnan, and Frank Li. Privacy and accountability for location-based aggregate statistics. In Yan Chen, George Danezis, and Vitaly Shmatikov, editors, *ACM Conference on Computer and Communications Security - CCS 2011, Chicago, Illinois, USA*, pages 653–666. ACM, 2011.

[24] Bartosz Przydatek, Dawn Song, and Adrian Perrig. Sia: secure information aggregation in sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, SenSys '03, pages 255–265, New York, NY, USA, 2003. ACM.

[25] Ishtiaq Rouf, Hossen Mustafa, Miao Xu, Wenyuan Xu, Rob Miller, and Marco Gruteser. Neighborhood watch: security and privacy analysis of automatic meter reading systems. In *Proceedings of the 2012 ACM conference on Computer and communications security*, CCS '12, pages 462–473, New York, NY, USA, 2012. ACM.

[26] Jing Shi, Rui Zhang, Yunzhong Liu, and Yanchao Zhang. Prisense: Privacy-preserving data aggregation in people-centric urban sensing systems. In *INFOCOM 2010. 29th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 15-19 March 2010, San Diego, CA, USA*, pages 758–766. IEEE, 2010.

[27] Osman Ugus, Dirk Westhoff, Ralf Laue, Abdulhadi Shoufan, and Sorin A. Huss. Optimized Implementation of Elliptic Curve Based Additive Homomorphic Encryption for Wireless Sensor Networks. In *2nd Workshop on Embedded Systems Security, WESS'2007*, Salzburg, Austria, October 4 2007.

[28] David Wagner. Resilient aggregation in sensor networks. In *ACM workshop on Security of ad hoc and sensor networks - SASN '04*, pages 78–87, Washington DC, USA, October 2004. ACM.

[29] Dirk Westhoff, Joao Girao, and Mithun Acharya. Concealed data aggregation for reverse multicast traffic in sensor networks: Encryption, key distribution, and routing adaptation. *IEEE Transactions on Mobile Computing*, 5(10):1417–1431, October 2006.

[30] Andrew C. Yao. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, SFCS '82, pages 160–164, Washington, DC, USA, 1982. IEEE Computer Society.

A Range Test Description

We describe here the range test proposed in [21]. Note that an extended version of this protocol secure in the active adversary model is presented in [22]. This range test scheme requires two additive homomorphic[1] semantically-secure encryption schemes (E_1, D_1) and (E_2, D_2) and involves two parties: a tester and an authority. The tester holds the encrypted values while the authority hold the private keys of the two cryptosystems K^1_{priv} and K^2_{priv}. The public keys of both encryption systems are known by the tester and the authority. In addition, the size of the encryption systems' message space $\mathbb{Z}_{p_1}$ and $\mathbb{Z}_{p_2}$ must satisfy: $p_2 \geq 3p_1$ and p_2 must be

[1] i.e. $E(m_1)E(m_2) = E(m_1 + m_2)$.

prime. The range involved in the test can be any $\mathbb{Z}_q$ with the condition $5q \leq p_1$. The range test described in [21] relies on a particular Specialized Zero Test [20].

A.1 Specialized Zero Test

Like the range test, this protocol involves two parties, a tester A_1 and an authority A_2, as well as an homomorphic asymmetric cryptosystem. The public key is known by both parties but only the second party (the authority) knows the private key. Given a set of n ciphertexts, this protocol allows to verify if one of the ciphertext is null, without revealing any information about the plaintexts. A specialized zero test involving two participants A_1 and A_2 on the ciphertexts $\{c_1, \ldots, c_n\}$ output TRUE if at least one ciphertext is null and false if no null ciphertext is found. It is denoted $ZM(A_1, A_2 | c_1, \cdots, c_n)$ and it works as follows:

- Q_1 chooses a permutation $\pi()$ on $\{1, \cdots, n\}$ and random integers r_i from $\mathbb{Z}_{p_2} \backslash \{0\}$ for $i = 1, \cdots, n$. Then he calculates $c'_i = c_{\pi(i)}^{r_i}$ for $i = 1, \cdots, n$. He sends $(c'_1, c'_2, \cdots, c'_n)$ to A_2.
- A_2 calculates $d_i = D_2(c'_i)$ for $i = 1, \cdots, n$ one by one until one d_i is found to be zero or all the n ciphertexts are decrypted. A_2 publishes the output of the zero test as follows:

$$ZM(A_1, A_2 | c_1, \cdots, c_n) = \begin{cases} \mathsf{TRUE} \text{ if zero found in } d_i \\ \mathsf{FALSE} \text{ if no zero in } d_i \end{cases}$$

A.2 Range Test

In [21], the authors first introduce two range tests efficient in passive adversary model. The first test called basic range test is implemented as follows:

- The tester (A_1) divide the ciphertext c into 2 ciphertext c_1 and c_2 such that $c_1 = E_1(m_1)$ and $c_2 = c/c_1$, where m_1 is randomly chosen in $\mathbb{Z}_{p_1}$. He sends c_2 to A_2.
- A_2 calculates $m_2 = D_1(c_2), c'_2 = E_2(m_2)$ and $e_2 = E_2(m_2 \bmod q)$. He then sends c'_2 and e_2 to A_1.
- A_1 calculates $c'_1 = E_2(m_1)$ and $e_1 = E_2(m_1 \bmod q)$. He then performs with the help of A_2 a specialized zero test:

$$\begin{aligned} ZM(A_1, A_2 | e_1 e_2/(c'_1 c'_2), e_1 e_2/(c'_1 c'_2 E_2(q)), \\ e_1 e_2/(c'_1 c'_2 E_2(p_1 \bmod q)), e_1 e_2/(c'_1 c'_2 E_2(p_1 \bmod q - q)), \\ e_1 e_2/(c'_1 c'_2 E_2(p_1 \bmod q + q)), \end{aligned}$$

This test is denoted as $BR(A_1, A_2 | c)$. The precise range test that finally allows to test if an element belongs to an interval makes two calls to the BR tests: $BR(A_1, A_2 | c)$ and $BR(A_1, A_2 | E_1(q - 1)/c)$ in a random order. The output of the precise test (denoted $PR(A_1, A_2 | c)$) is TRUE if the two basic range tests output TRUE and FALSE otherwise. In this case, the precise test proves that $0 \leq D_1(c) < q$.

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399