Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 827959, 9 pages
http://dx.doi.org/10.1155/2014/827959

Research Article

Towards Self-Awareness Privacy Protection for Internet of Things Data Collection

Kok-Seng Wong and Myung Ho Kim

School of Computer Science and Engineering, Soongsil University, Information Science Building, Sangdo-dong, Dongjak-gu, Seoul 156-743, Republic of Korea

Correspondence should be addressed to Myung Ho Kim; kmh@ssu.ac.kr

Received 10 February 2014; Accepted 6 May 2014; Published 19 June 2014

Academic Editor: Young-Sik Jeong

The Internet of Things (IoT) is now an emerging global Internet-based information architecture used to facilitate the exchange of goods and services. IoT-related applications are aiming to bring technology to people anytime and anywhere, with any device. However, the use of IoT raises a privacy concern because data will be collected automatically from the network devices and objects which are embedded with IoT technologies. In the current applications, data collector is a dominant player who enforces the secure protocol that cannot be verified by the data owners. In view of this, some of the respondents might refuse to contribute their personal data or submit inaccurate data. In this paper, we study a self-awareness data collection protocol to raise the confidence of the respondents when submitting their personal data to the data collector. Our self-awareness protocol requires each respondent to help others in preserving his privacy. The communication (respondents and data collector) and collaboration (among respondents) in our solution will be performed automatically.

1. Introduction

The Internet of Things (IoT) is now an emerging global Internet-based information architecture used to facilitate the exchange of goods and services. The concept of IoT is to allow living objects (humans or animals), devices (sensor), or object with embedded technologies to automatically transfer data over communication networks (wired or wireless networks) without human-to-human or human-to-computer interaction. IoT aims to utilize and extend the benefits of Internet such as always-on, data sharing, and remote access capabilities [1].

IoT enables data collection in every aspect of our life. Data collected from smart metering application allows the utility provider to analyze and improve its services. Also, these data can help the user to be aware of their energy consumptions and possible energy saving strategies. In an underwater environment, smart meter is particularly important because information can be detected, gathered, and sent to the sensor [2].

Let us consider the following scenario. A practitioner (data collector) would like to collect medical data from his patients (respondents) with implanted medical devices. Since medical data are highly sensitive information, respondents must be aware of the data to be collected. There are two main paradigms to protect the patient's privacy in this scenario. The first paradigm relies on the respondent's trust in the data collector while the second paradigm depends on the respondent's anonymity. If the respondents do not have confidence in the data collector, they may refuse to submit data or provide inaccurate data to the agency. If the submitted data from the respondents are not genuine, we can predict that the data collector will face the data utility problem because the analyzed results based on the collected data will not be accurate. In the second paradigm, we should prevent the reidentification problem. For instance, if the collected data are used for research purposes, the data collector should not be able to link any of the collected data to the real identity of any patient.

1.1. Challenges of IoT. Wireless sensor networks have been revolutionized by creating significant impact throughout the society [3]. Advances in wireless communication technology (e.g., efficient resource management [4] and performance improvement [5] in wireless network) enable the development and implementation of IoT applications. IoT-related applications include traffic congestion detection and waste management in smart cities, remote diagnostics in patients' surveillance system (e.g., Ubiquitous healthcare [6, 7]), and storage condition monitoring in supply chain control.

Along with potential benefits offered, the usage of IoT also raises some privacy concerns to the data owners. In particular, real-time data collection and data analysis in IoT applications may compromise the privacy of data owner. In practical, new data arrive continuously and up-to-date data should be used for analysis. The data collected at different times allows malicious providers to learn extra knowledge by cross-examining the data within a targeted timeframe. Therefore, a secure and privacy aware protocol should be implemented in IoT when data are collected automatically. Some new security and privacy challenges can be found in [8].

The development of radio frequency identification (RFID) technologies and the advances of network communication technologies motivate the forming of IoT [9]. Physical objects called u-things which are embedded or connected to communication networks, sensors, and computers are commonly found in our daily life [10]. In the context of IoT, u-things should be able to act automatically (e.g., autodetection and data transfer) and adaptively. The construction of smart u-things involves the following 7 challenges [11, 12]:

(i) surrounding situations (context),

(ii) users' needs,

(iii) things' relations,

(iv) common knowledge,

(v) self-awareness,

(vi) looped decisions,

(vii) ubiquitous safety (UbiSafe).

The ultimate goal of any ubiquitous intelligence is to make the u-things behave trustworthily in both other-aware and self-aware manners to some degrees and circumstances [13]. Therefore, it is important to design a self-awareness protocol to help data owners to protect their privacy.

In this paper, we will focus on the self-awareness challenge. In particular, we design a self-awareness protocol to increase the confidence of the data owner when the smart u-things automatically submit their data to the data collector.

1.2. Problem Statement. There are two challenges we aim to address in this work. Firstly, we want to protect the identity of each data owner from the data collector before and after the data collection process. Secondly, and more importantly, we want to guarantee the usefulness of the collected data by increasing the confidence of data owner.

The first challenge can be solved by using anonymity technology such as the onion routing (Tor) [14], anonymous proxy server [15], and mix network [16, 17]. These technologies are still under active investigation and their focuses are mainly on network traffic analysis, anonymous communication channel, and private information retrieval. Since our aim in this paper is not to design any of the specific anonymity technology, we refer readers to [15, 18] for the usage of these technologies.

The second challenge requires each respondent to help others in order to preserve his own privacy. This idea is motivated by the coprivacy concept in [19, 20]. Coprivacy (or cooperative privacy) considers the best option for a party to achieve his privacy protection is to help another party in achieving her privacy. The formal definition of coprivacy and its generalizations can be found in [19].

1.3. Our Contributions. In this paper, we propose a self-awareness protocol to facilitate the data collection in IoT-related applications. Instead of placing full trust on the utility provider (data collector), we allow each data owner (respondent) to learn the protection level provided by the data collector before the data submission process. We summarize our contributions as follows.

(i) We propose a privacy preserved approach to enable the respondents to learn about the anonymous protection level they will receive from the data collector before the data submission.

(ii) Our notion of self-awareness protection can be used to increase the confidence of respondents in the data collection process. Hence, respondents will feel comfortable to submit their genuine data while the data collector can ensure the usefulness of the collected data.

1.4. Organization. The rest of this chapter is organized as follows. The background and related work for this research are presented in Section 2. We describe the technical preliminaries of our solution in Section 3. We present our solution in Section 4 followed by analysis of correctness, privacy, efficiency, and discussion in Section 5. Our conclusion is in Section 6.

2. Background and Related Work

2.1. Privacy Paradigm in IoT. In 1973, the United States Department of Health, Education, and Welfare proposed Fair Information Practice Principles (FIPPs) as the guideline to assure fair practice and adequate data privacy protection. In particular, the guideline aims to protect the consumer rights such as how online entities should collect and use the personal data [21]. Five principles of FIPPs are as follows [22].

(1) There must be no personal data record-keeping systems whose very existence is secret.

(2) There must be a way for a person to find out what information about the person is in a record and how it is used.

(3) There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.

(4) There must be a way for a person to correct or amend a record of identifiable information about the person.

(5) Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

Based on the above principles, we now analyze the privacy protection in current IoT. Since data are collected automatically, it is hard for the data owners to ensure that their privacy can be protected. In most cases, utility providers will design a series of mechanisms to guarantee the privacy protection of the collected data. However, we found that data owners are generally not able to verify those mechanisms offered by the provider. Therefore, a self-awareness protocol should be available for automatic data collection process.

2.2. Anonymous Data Collection. In general, online data collection is a process which involves collaboration between a trusted party (data collector) and a number of data owners (respondents). Due to concerns regarding privacy, respondents might refuse to contribute their personal data or submit inaccurate data to the data collector. Therefore, the data collector needs to ensure the privacy of data submitted through a series of secure mechanisms. However, the protection level provided by the data collector is hard to be verified by the respondents.

Often, data collected from the respondents will be used for research or data analysis. The release of the collected data causes a privacy issue in data publishing, in particular, when it involves the republication of the same data in a given period [23]. There are two settings that can be observed when the data is released to the data recipient. If the data recipient is a third party, data must be released in an anonymous form without compromising the privacy of the respondents. Let us consider a scenario where a hospital (data collector) wishes to publish patients' records to a research institute (data recipient) for data analysis. In a common practice, all the explicit personal identity information (PII) such as name and social security number will be removed from the original dataset before it is released to the data recipient. However, removing PII does not preserve privacy.

Data anonymization is an interesting solution to protect the privacy of the respondents for this setting. Sweeney proposed k-anonymity model to address the linking attack [24]. The concept of k-anonymity [25] is such that each released data is indistinct from at least $(k-1)$ other data. However, k-anonymity is found vulnerable against background knowledge attacks by Machanavajjhala et al. [26].

In the literature, techniques such as (α, k)-anonymity [27, 28], l-diversity [26], and t-closeness [29] have been proposed to enhance the k-anonymity model. We note that these techniques assumed that k-anonymity has been achieved in the first place before applying additional techniques to enhance the anonymous protection of the released data. For instance, (α, k)-anonymity model assumed that all the released data adhere to k-anonymity. In addition, it requires that the frequency of the sensitive value in any quasi-identifier is less than α after the anonymization [27]. In the l-diversity model, the sensitive attribute in the k-anonymous table is well represented by l values such that each sensitive value is at most $1/l$. A survey of recent attacks and privacy models in data publishing can be found in [30].

In this paper, we consider the second setting where the data analysis is performed by the data collector. This scenario is more complex to deal with because the data collector has the full access to all raw data from the respondents. Therefore, we need to design a protocol to increase the confidence of the respondents before they submit their records to the data collector. In other words, respondents are aware of the protection level they received from the data collector after the data submission.

3. Related Works

Various self-oriented privacy protections have been proposed in the literature. Self-enforcing privacy (SEP) for e-polling was proposed in [31]. The idea of SEP is to enforce the pollster to protect the respondents' privacy by allowing the respondents to trace their data after the submission. If the pollster releases the poll results, the respondents can indict the pollster by using the evidence they obtained during the data collection process. A fair indictment scheme for SEP can be found in [32].

The most related research to our work in this paper is the respondent-defined privacy protection (RDPP) for anonymous data collection proposed in [33]. The basic idea of RDPP is to allow the respondents to specify the level of protection they require before providing any data to the data collector. For instance, a number of respondents (minimum threshold) must satisfy the constraint chosen by the respondent i before he agrees to submit the data. In their protocol, respondents are aware of the minimum level of privacy protection they will receive before submitting their dataset to the data collector. Instead of relying on the data collector to guarantee the privacy protection, the respondents are free to define their preferred protection level.

In this paper, we do not consider indictment for our protocol because the data analysis is done by the data collector. Instead of allowing the respondents to freely define their own privacies, we assume that respondents are willing to submit their data if the protection level offered by the data collector can be verified by them.

4. Technical Preliminaries

4.1. Homomorphic Encryption Scheme. We use homomorphic encryption scheme (i.e., Paillier [34]) as our primary cryptographic tool. Let $\mathrm{Enc}_{\mathrm{pk}}(m)$ denote the encryption of m with the public key, pk. Given two ciphertexts, $\mathrm{Enc}_{\mathrm{pk}}(m_1)$ and $\mathrm{Enc}_{\mathrm{pk}}(m_2)$, there exists an efficient algorithm $+_h$ to compute

Table 1: Sample medical dataset.

Patient	Gender	Age	Zip	Disease
Bob	Male	15	27892	Flu
Sam	Male	13	27886	Heart disease

$\text{Enc}_{\text{pk}}(m_1 + m_2)$. This additive property can be performed without the decryption key.

4.2. Definitions. Let us assume that there are n respondents $\mathcal{R} = \{\mathcal{R}^1, \mathcal{R}^2, \ldots, \mathcal{R}^n\}$ and a data collector $\mathcal{C}$. Each respondent i has a database $\mathcal{D}_i$ with m records. We denote T as the dataset collected by the data collector. Also, the dataset T consists of d quasi-identifier $\text{QID} = \{\text{QI}_1, \text{QI}_2, \ldots, \text{QI}_d\}$ and a sensitive attribute. Note that the quasi-identifier can be either categorical or continuous data while the sensitive attribute is a categorical data from its domain (e.g., disease).

A quasi-identifier (QI) is a minimal set of attributes in T that can be joined with external information to uniquely distinguish individual records [24]. Note that the quasi-identifier can be either categorical or continuous data while the sensitive attribute is a categorical data from its domain.

Definition 1 (quasi-identifier). A quasi-identifier (QI) is a minimal set of attributes that can uniquely distinguish tuples in T. The QI for Table 1 is {Gender, Age, Zip} and it can be generalized as {Male, 10–16, 278 ∗ ∗}.

Definition 2 (k-anonymity). T is said to satisfy k-anonymity with respect to QI if and only if each set of attributes in QI appears at least k occurrences in T.

Definition 3 (self-awareness privacy). Each respondent i is said to achieve self-awareness privacy if he learns the protection level (e.g., k-anonymity) provided by the data collector. At the end of the protocol execution, each respondent remains anonymous to others and the data collector is not able to identify any of the respondents with probability more than 0.5.

4.3. Components. Our self-awareness data collection protocol consists of the following three components.

(i) *Data collector*: an authorized party who wants to collect data from a group of respondents via wired or wireless network.

(ii) *Respondent*: participant in the data collection process who is also a candidate to submit his/her record to the data collector.

(iii) *The onion router* (Tor): an anonymous network used to conceal the respondent's privacy such that the agency cannot monitor the activity flows of any respondent.

We show the interactions among the components in our solution in Figure 1. We assume that the respondents and the data collector are equipped with ubiquitous sensors to detect, communicate, and execute the protocol.

4.4. Adversary Model. We assume that both the data collector and the respondents are semihonest players (also known as honest-but-curious). Semihonest players follow the protocol faithfully but may try to discover extra information during the protocol execution.

In our protocol design, the data collector must follow the protocol faithfully in order to ensure that all respondents are willing to participate in the data collection process. For the same reason, all respondents should be semihonest in order to ensure that the privacy protection level offered by the data collector can be achieved.

4.5. Notations Used. The notations used hereafter in this paper are summarized in Notations section.

5. Self-Awareness Data Collection Protocol

5.1. Protocol Idea. The basic idea of our protocol is to allow the respondents to know the protection level they will receive from the data collector before the data submission process [35]. In our design, the data collector will release a set of quasi-identifiers $\text{QID} = \{\text{QI}_1, \text{QI}_2, \ldots, \text{QI}_n\}$ for T and define a protection level it wants to provide to the respondents (e.g., a threshold k). Note that a larger k will make the respondents feel more comfortable to submit their records. We also require the respondents to collaborate together to find the number of records in $(\mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_n)$ which met the quasi-identifier determined by the data collector. We assume that the communication between the data collector and the respondents is via a mixture network such as Tor [14]. Note that the communication (respondents and data collector) and collaboration (among respondents) in our solution are run automatically. We show the overview of our proposed solution in Figure 1.

In the following sections, we will describe our self-awareness data collection protocol in details.

5.2. Our Protocol. In order to participate in the data collection process, all players can precompute some information to be used during the protocol execution. For example, each respondent i can generate a cryptographic key pair $(\text{pk}_i, \text{pr}_i)$ where pk_i is the public key and pr_i is the corresponding private key. Next, the respondents encrypt their personal identifiable information (PII) such as name or social security number by using the pk_i. The encrypted PII will be used as the public identity $\mathcal{I}_i$ of the respondent i. This public identity is important for other respondents to identify the owner of a given public key. Each respondent then submits his public identity and encryption key to the data collector via a Tor network. Let us assume there are n respondents who participate in the data collection process and, hence, the data collector will receive n submissions $(\mathcal{I}_1, \text{pk}_1), (\mathcal{I}_2, \text{pk}_2), \ldots, (\mathcal{I}_n, \text{pk}_n)$ from the respondents.

Before the data collection begins, the data collector is required to define a set of m quasi-identifiers denoted as $\text{QID} = \{\text{QI}_1, \text{QI}_2, \ldots, \text{QI}_m\}$ for the dataset T to be collected and determine the protection level (e.g., k value) for the respondents.

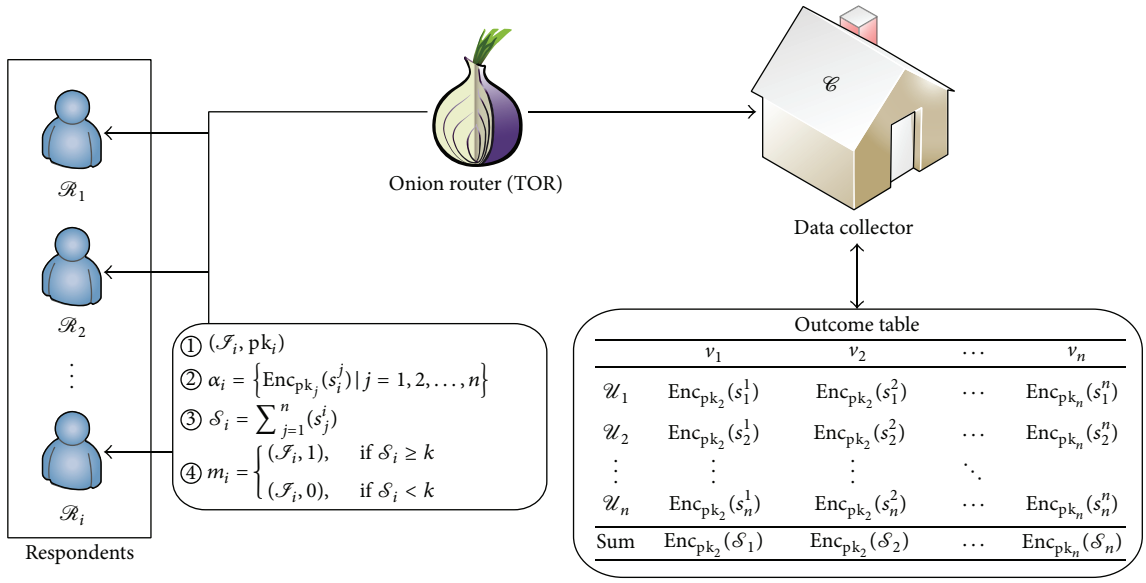

Figure 1: Overview of the proposed solution.

Table 2: Outcome table released by the data collector.

	v_1	v_2	$\cdots$	v_n
$\mathscr{U}_1$	$\mathrm{Enc}_{\mathrm{pk}_1}(s_1^1)$	$\mathrm{Enc}_{\mathrm{pk}_2}(s_1^2)$	$\cdots$	$\mathrm{Enc}_{\mathrm{pk}_n}(s_1^n)$
$\mathscr{U}_2$	$\mathrm{Enc}_{\mathrm{pk}_1}(s_2^1)$	$\mathrm{Enc}_{\mathrm{pk}_2}(s_2^2)$	$\cdots$	$\mathrm{Enc}_{\mathrm{pk}_n}(s_2^n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathscr{U}_n$	$\mathrm{Enc}_{\mathrm{pk}_1}(s_n^1)$	$\mathrm{Enc}_{\mathrm{pk}_2}(s_n^2)$	$\cdots$	$\mathrm{Enc}_{\mathrm{pk}_n}(s_n^n)$
SUM	$\mathrm{Enc}_{\mathrm{pk}_1}(\mathscr{S}_1)$	$\mathrm{Enc}_{\mathrm{pk}_2}(\mathscr{S}_2)$	$\cdots$	$\mathrm{Enc}_{\mathrm{pk}_n}(\mathscr{S}_n)$

To initiate the protocol, the data collector first randomly assigns a public key pk_i for each $\mathrm{QI}_j \in \mathrm{QID}$. If $|\mathrm{QID}| > n$, the same public key can be assigned to more than one quasi-identifier. Otherwise, the data collector selects m/n of the public keys for the assignment. For simplicity, we will assume that the size for both quasi-identifier and public key is equal (i.e., $m = n$) and $\ell = \{(\mathrm{pk}_1, \mathrm{QI}_1), (\mathrm{pk}_2, \mathrm{QI}_2), \ldots, (\mathrm{pk}_n, \mathrm{QI}_n)\}$. Next, the data collector publishes $(\mathscr{I}, \ell)$ to a shared location (e.g., a webpage):

$$(\mathscr{I}, \ell) = \{(\mathscr{I}_1, (\mathrm{pk}_1, \mathrm{QI}_1)), (\mathscr{I}_2, (\mathrm{pk}_2, \mathrm{QI}_2)), \ldots, (\mathscr{I}_n, (\mathrm{pk}_n, \mathrm{QI}_n))\}. \quad (1)$$

Based on the information from (1), each respondent i retrieves ℓ to examine if his records in $\mathscr{D}_i$ match any of the quasi-identifiers $\mathrm{QI}_j \in \mathrm{QID}$. At this phase, each respondent i maintains a scores list for QID, $\{s_i^1, s_i^2, \ldots, s_i^n\}$. We denote s_i^j as the score determined by the respondent i for QI_j. The respondent raises each score by 1 when a record in $\mathscr{D}_i$ matches the quasi-identifier. Upon the completion, the respondent i encrypts each s_i^j by using the public key pk_j assigned to the quasi-identifier QI_j. The encrypted scores list computed by each respondent i can be represented as $\alpha_i = \{\mathrm{Enc}_{\mathrm{pk}_1}(s_i^1), \mathrm{Enc}_{\mathrm{pk}_2}(s_i^2), \ldots, \mathrm{Enc}_{\mathrm{pk}_n}(s_i^n)\}$. Then, all the respondents send α_i to the data collector and a shared location. Note that this location can be a separate space that is not shared with the data collector.

Upon receiving α_i from all the respondents, the data collector performs the following tasks.

(1) *Aggregates the scores determined by all respondents for each* QI_i. The data collector performs this computation in an encrypted form by using the additive property of the Paillier cryptosystem. The output of the aggregation can be represented as

$$\mathrm{Enc}_{\mathrm{pk}_j}(\mathscr{S}_i) = \mathrm{Enc}_{\mathrm{pk}_j}(s_1^j) +_h \mathrm{Enc}_{\mathrm{pk}_j}(s_2^j) +_h \cdots +_h \mathrm{Enc}_{\mathrm{pk}_j}(s_n^j). \quad (2)$$

(2) *Publishes an outcome table*. The data collector publishes the scores for each QI_j in an outcome table as shown in Table 2. In Table 2, each row $(\mathscr{U}_i)$ represents the encrypted scores received from each respondent i while the column (v_i) shows the encrypted scores for each quasi-identifier QI_j. Note that all the data in v_j are encrypted by using the same public key pk_j. Therefore, only the respondent who has been assigned the QI_j can decrypt $\mathrm{Enc}_{\mathrm{pk}_j}(\mathscr{S}_i)$ to learn the number of matched records $(\mathscr{S}_i)$ for QI_j.

After the data collector releases the outcome table, the respondents need to verify that the data released are genuine. For instance, each respondent i verifies that the encrypted scores list α_i submitted to the data collector appears as one of the rows in Table 2. If the respondent fails to verify the data, he or she then issues a decision message m_i with a random value.

Let us assume all the respondents successfully verify the data in Table 2. Next, each respondent i retrieves v_i

Self-Awareness Data Collection Protocol

Phase 1: Public Key and Public Identity Submissions

The data collector broadcasts a submission request to n respondents. Each $\mathscr{R}_i$ generates a cryptographic key pair $(\mathrm{pk}_i, \mathrm{pr}_i)$ and a public identity $\mathscr{I}_i$ by encrypting its personal identifiable information (PII). Note that the respondents can precompute the cryptographic key pair and the PII in an offline mode. Next, each $\mathscr{R}_i$ sends $(\mathscr{I}_i, \mathrm{pk}_i)$ to $\mathscr{C}$ via the Tor network.

Phase 2: Satisfaction Scores Computation

The data collector $\mathscr{C}$ generates QID, decides a threshold k and assigns a public key for each QI_i. Next, it broadcasts the information to all respondents. Each $\mathscr{R}_i$ examines if his record in $\mathscr{D}_i$ satisfy QID. For each satisfy case, the $\mathscr{R}_i$ increases the constraint score s_i^j by 1. We denote s_i^j as the score determines by $\mathscr{R}_i$ for QI_j. Next, each $\mathscr{R}_i$ encrypts $\{s_i^j \mid j = 1, 2, \ldots, n\}$ by using the public key pk_j to produce $\alpha_i = \{\mathrm{Enc}_{\mathrm{pk}_j}(s_i^j) \mid j = 1, 2, \ldots, n\}$. Each $\mathscr{R}_i$ then anonymously sends α_i to $\mathscr{C}$ and a shared location.

Phase 3: Scores List Verification

The data collector $\mathscr{C}$ computes and publishes an outcome table. Each $\mathscr{R}_i$ examines if the published scores list is same as the original list he sent to $\mathscr{C}$. If the list has been modified, the respondent will not participate in the next phase.

Phase 4: Satisfaction Score Checking

Each $\mathscr{R}_j$ retrieves and decrypts $\{\mathrm{Enc}_{\mathrm{pk}_j}(s_j^i) \mid i = 1, 2, \ldots, n\}$. Next, it computes $\mathscr{S}_j = \sum_{i=1}^{n}(s_j^i)$ as the satisfaction score for QI_j. If the satisfaction score $\mathscr{S}_j$ is at least with k_j occurences (e.g., $\mathscr{S}_j \geq k_j$), the $\mathscr{R}_j$ sends $m_i = (\mathscr{I}_i, 1)$ to $\mathscr{C}$. Otherwise, $m_i = (\mathscr{I}_i, 0)$ will be sent to $\mathscr{C}$.

Phase 5: Data Submission

The respondents submit his record to $\mathscr{C}$ with the confidence that their privacy protection is achieved at k-anonymity level.

Algorithm 1: Self-Awareness data collection protocol.

(based on his public identity $\mathscr{I}_i$) and decrypts all encrypted data by using the private key pr_i. After the decryption, the respondents must ensure that the aggregated score $\mathrm{Enc}_{\mathrm{pk}_j}(\mathscr{S}_i)$ computed by the data collector is correct. The respondents can verify this by computing $\mathscr{S}_i = \sum_{j=1}^{n}(s_j^i)$ from the decrypted scores and then compare it with the decrypted result of $\mathrm{Enc}_{\mathrm{pk}_j}(\mathscr{S}_i)$. Lastly, each respondent i compares $\mathscr{S}_i$ with the threshold k determined by the data collector. If the number of matched records $\mathscr{S}_i$ is greater than the threshold value (e.g., $\mathscr{S}_i \geq k$), we assume that the respondent will submit his records to the data collector. Otherwise, the respondent will abort from the data collection process.

At the final phase, each respondent i sends a decision message m_i to the shared location. If the decision message m_i is set to 1, this indicates that $\mathscr{S}_i \geq k$. Therefore, the respondents should submit their records to the data collector. Otherwise, if m_i is set to 0, the respondents should not reveal any record to the data collector.

We summarize our self-awareness data collection protocol in Algorithm 1.

6. Analysis and Discussion

6.1. Analysis of Correctness. In this paper, we assume that both the data collector and the respondents are semihonest players. The semihonest model is realistic in our solution. If both players follow the protocol faithfully, each respondent can ensure that he will achieve the protection level offered by the data collector (e.g., k-anonymity). At the same time, the data collector can guarantee that the datasets collected are useful for analysis.

During the protocol execution, all respondents are required to verify (1) the encrypted scores released by the data collector are genuine and (2) the aggregated score for each QI_j computed by the data collector is correct. The first verification is to ensure that the data collector has received all data computed by the respondents correctly while the second verification is useful for the respondents to detect a malicious data collector.

In our protocol design, the data collector needs to define a protection level (e.g., k value) before the data collection begins. The data collector can define the same protection level for all QI_j or define difference in anonymous levels k_i for each $\mathrm{QI}_i \in \mathrm{QID}$. For the latter case, the respondents can perform the same steps to verify each value of k_i.

6.2. Analysis of Privacy. The privacy analysis of our protocol depends on how much information has been revealed during the protocol execution. In general, our solution should protect the privacy of the respondents. This leads to the following two requirements: (1) the data collector should not be able to infer any sensitive information of the respondents from the data collected and (2) the respondents are aware of

the data they submit and the protection level they will receive from the data collector.

In our protocol design, we utilize Tor network to prevent direct communication between the data collector and the respondents. This approach will not allow the data collector to track the identity of any respondent. Also, we assume that each respondent has no knowledge about the profile of other respondents, but the number of respondents in the protocol is known publicly.

The unique identity $\mathcal{I}_i$ of each respondent will not leak the profile of any respondent because they are in an encrypted form. The data collector is not able to decrypt α_i in the absence of private keys from the respondents. Further, our protocol ensures that no party (including the data collector) can learn the encrypted score in the outcome table before the decryption. Note that only the respondent who has the private key can perform the decryption.

To prevent possible collusions between the data collector and other respondents, we assume that all data transmissions are performed via an anonymous communication channel (e.g., Tor network). This can ensure that the profile of each respondent remains anonymous from others.

The shared location (e.g., web page or web folder) used in our protocol is to allow the respondents to learn the decisions made by others and to detect a malicious data collector. Each respondent notifies others about the verification result by using a decision message m. Since the decision message only reveals the public identity of the respondents, we can assume that the profile of the respondents remains hidden from others.

6.3. Analysis of Efficiency. The complexity of our protocol is dominated by the cryptographic operations (encryption and decryption) performed by respondents. We implement our protocol in Java and ran it on a single computer with a 2 GHz CPU and a 2 GB RAM. The performance evaluation is shown in Figure 2. Each respondent performs the same amount of cryptographic operations in our experiment.

6.4. Discussion. In this paper, we assume that the size of the public keys (or the number of respondents) and the quasi-identifier is equal (e.g., $|\mathcal{R}| = |\text{QID}| = n$). However, our protocol works correctly for unequal cases. The owner of the public key only performs the decryption and computes $\mathcal{S}_i$ at the end of the protocol execution. A respondent may not be involved in the final phase if his public key is not selected by the data collector (for cases when $|\mathcal{R}| > n$). Otherwise, a respondent needs to repeat final phase for several times if his public key is assigned to more than one QI_j.

7. Conclusion and Future Work

In this paper, we presented a self-awareness protocol for IoT data collection. Since the release of raw data to the data collector has a high risk to compromise privacy of the respondents, we aim to increase confidence of the respondents before they submit their records to the data collector. Our self-awareness protocol allows each respondent to help others in order to preserve his own privacy. At the same time, the final collected data should adhere to the protection level promised by the data collector before the data collection begins. Also, our solution can be extended to support indictment scheme (when the data is released to a third party) because the respondents have evidence (e.g., value of k) to indict a malicious data collector.

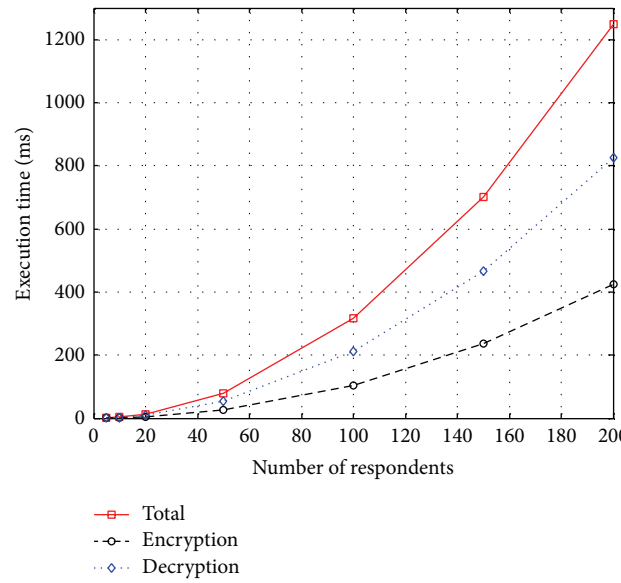

Figure 2: Performance of the proposed solution.

Notations

$\mathcal{R}_i$: Respondent i
$|\mathcal{R}|$: Size of the respondents
T: Dataset collected by the data collector
$\mathcal{D}_i$: Local database of respondent i
k: Anonymous protection level
QID: Quasi-identifier set determined by the data collector
|QID|: Size of the quasi-identifier
QI_i: ith quasi-identifier in QID
$\mathcal{I}_i$: Public identity of the respondent i
s_i^j: Score determined by the respondent i for QI_j
$\mathcal{S}_i$: Satisfaction score of QI_i
pk_i: Public key of respondent i
pr_i: Private key of respondent i
$\text{Enc}_{\text{pk}_i}(\cdot)$: Encryption operation by using pk_i
$\text{Dec}_{\text{pr}_i}(\cdot)$: Decryption operation by using pr_i
m_i: Decision message from respondent i.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] N. Y. Yen and S. Y. F. Kuo, "An integrated approach for internet resources mining and searching," *Journal of Convergence*, vol. 3, pp. 37–44, 2012.

[2] S. K. Dhurandher, M. S. Obaidat, and M. Gupta, "An acoustic communication based AQUA-GLOMO simulator for underwater networks," *Human-Centric Computation and Information Sciences*, vol. 2, article 3, 2012.

[3] B. Singh and D. K. Lobiyal, "A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks," *Human-Centric Computation and Information Sciences*, vol. 2, article 13, 2012.

[4] G. H. S. Carvalho, I. Woungang, A. Anpalagan, and S. K. Dhurandher, "Energy-efficient radio resource management scheme for heterogeneous wireless networks: a queueing theory perspective," *Journal of Convergence*, vol. 3, no. 4, pp. 15–22, 2012.

[5] A. U. Bandaranayake, V. Pandit, and D. P. Agrawal, "Indoor link quality comparison of IEEE 802.11a channels in a multi-radio mesh network testbed," *Journal of Information Processing Systems*, vol. 8, no. 1, pp. 1–20, 2012.

[6] J. K.-Y. Ng, "Ubiquitous healthcare: healthcare systems and applications enabled by mobile and wireless technologies," *Journal of Convergence*, vol. 3, no. 2, pp. 15–20, 2012.

[7] F. Barigou, B. Atmani, and B. Beldjilali, "Using a cellular automaton to extract medical information from clinical reports," *Journal of Information Processing Systems*, vol. 8, no. 1, pp. 67–84, 2012.

[8] R. H. Weber, "Internet of things—new security and privacy challenges," *Computer Law and Security Review*, vol. 26, no. 1, pp. 23–30, 2010.

[9] N. Zhong, J. H. Ma, R. H. Huang et al., "Research challenges and perspectives on wisdom web of things (W2T)," *The Journal of Supercomputing*, vol. 64, no. 3, pp. 862–882, 2013.

[10] J. Ma, "Smart u-things-challenging real world complexity," *IPSJ Symposium Series*, vol. 19, pp. 146–150, 2005.

[11] J. Ma, "Smart u-things and ubiquitous intelligence," in *Proceedings of the 2nd International Conference on Embedded Software and Systems*, p. 776, Springer, Xi'an, China, 2005.

[12] J. Ma, Q. Zhao, V. Chaudhary et al., "Ubisafe computing: vision and challenges (I)," in *Proceedings of the 3rd International Conference on Autonomic and Trusted Computing*, pp. 386–397, Springer, Wuhan, China, 2006.

[13] J. Ma, L. T. Yang, B. O. Apduhan, R. Huang, L. Barolli, and M. Takizawa, "Towards a smart world and ubiquitous intelligence: a walkthrough from smart things to smart hyperspaces and UbicKids," *International Journal of Pervasive Computing and Communications*, vol. 1, no. 1, pp. 53–68, 2005.

[14] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: the second-generation onion router," in *Proceedings of the 13th conference on USENIX Security Symposium*, vol. 13, p. 21, USENIX Association, San Diego, Calif, USA, 2004.

[15] M. Edman and B. Yener, "On anonymity in an electronic society: a survey of anonymous communication systems," *ACM Computing Surveys*, vol. 42, no. 1, article 5, 2009.

[16] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–88, 1981.

[17] K. Peng, "Attack and correction: how to design a secure and efficient mix network," *Journal of Information Processing Systems*, vol. 8, no. 1, pp. 175–190, 2012.

[18] B. Li, E. Erdin, M. H. Gunes, G. Bebis, and T. Shipley, "An analysis of anonymity technology usage," in *Proceedings of the 3rd International Conference on Traffic Monitoring and Analysis*, pp. 108–121, Springer, Vienna, Austria, 2011.

[19] J. Domingo-Ferrer, "Coprivacy: towards a theory of sustainable privacy," in *Proceedings of the International Conference on Privacy in Statistical Databases*, pp. 258–268, Springer, Corfu, Greece, 2010.

[20] J. Domingo-Ferrer, "Coprivacy: an introduction to the theory and applications of co-operative privacy," *SORT: Statistics and Operations Research Transactions*, pp. 25–40, 2011.

[21] T. Teraoka, "Organization and exploration of heterogeneous personal data collected in daily life," *Human-Centric Computing and Information Sciences*, vol. 2, article 1, 2012.

[22] Privacy Online: A Report to Congress. Federal Trade Commission, 1998.

[23] K.-S. Wong and M. Kim, "Secure re-publication of dynamic big data," in *Cyberspace Safety and Security*, G. Wang, I. Ray, D. Feng, and M. Rajarajan, Eds., vol. 8300, pp. 468–477, Springer International Publishing, 2013.

[24] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[25] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, p. 188, ACM, Seattle, Wash, USA, 1998.

[26] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, article 3, 2007.

[27] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 754–759, ACM, Philadelphia, Pa, USA, 2006.

[28] R. C.-W. Wong, Y. Liu, J. Yin, Z. Huang, A. W.-C. Fu, and J. Pei, "(alpha, k)-anonymity based privacy preservation by lossy join," in *Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-Age Information Management Conference on Advances in Data and Web Management*, pp. 733–744, Springer, Huang Shan, China, 2007.

[29] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian, "t-Closeness: privacy beyond k-anonymity and ℓ-diversity," in *Proceedings of the 23rd International Conference on Data Engineering (ICDE '07)*, pp. 106–115, Istanbul, Turkey, April 2007.

[30] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: a survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, article 14, 2010.

[31] P. Golle, F. McSherry, and I. Mironov, "Data collection with self-enforcing privacy," in *Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS '06)*, pp. 69–78, ACM, Alexandria, Va, USA, November 2006.

[32] M. Stegelmann, "Towards fair indictment for data collection with self-enforcing privacy," in *Security and Privacy—Silver Linings in the Cloud*, K. Rannenberg, V. Varadharajan, and C. Weber, Eds., vol. 330, pp. 265–276, Springer, Berlin, Germany, 2010.

[33] R. Kumar, R. Gopal, and R. Garfinkel, "Freedom of privacy: anonymous data collection with respondent-defined privacy

protection," *INFORMS Journal on Computing*, vol. 22, no. 3, pp. 471–481, 2010.

[34] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proceedings of the 17th International Conference on Theory and Application of Cryptographic Techniques*, pp. 223–238, Springer, Prague, Czech Republic, 1999.

[35] K.-S. Wong and M. Kim, "Privacy-preserving data collection with self-awareness protection," in *Frontier and Innovation in Future Computing and Communications*, J. J. Park, A. Zomaya, H.-Y. Jeong, and M. Obaidat, Eds., vol. 301, pp. 365–371, Springer, Amsterdam, The Netherlands, 2014.