Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Towards a respondent-preferred k_i-anonymity model*

Kok-Seng WONG[1], Myung Ho KIM[†‡2]

([1]*School of Computer Science and Engineering, Soongsil University, Seoul 06978, Korea*)

([2]*School of Software, Soongsil University, Seoul 06978, Korea*)

[†]E-mail: kmh@ssu.ac.kr

Received Nov. 16, 2014; Revision accepted June 6, 2015; Crosschecked Aug. 6, 2015

Abstract: Recently, privacy concerns about data collection have received an increasing amount of attention. In data collection process, a data collector (an agency) assumed that all respondents would be comfortable with submitting their data if the published data was anonymous. We believe that this assumption is not realistic because the increase in privacy concerns causes some respondents to refuse participation or to submit inaccurate data to such agencies. If respondents submit inaccurate data, then the usefulness of the results from analysis of the collected data cannot be guaranteed. Furthermore, we note that the level of anonymity (i.e., k-anonymity) guaranteed by an agency cannot be verified by respondents since they generally do not have access to all of the data that is released. Therefore, we introduce the notion of k_i-anonymity, where k_i is the level of anonymity preferred by each respondent i. Instead of placing full trust in an agency, our solution increases respondent confidence by allowing each to decide the preferred level of protection. As such, our protocol ensures that respondents achieve their preferred k_i-anonymity during data collection and guarantees that the collected records are genuine and useful for data analysis.

Key words: Anonymous data collection, Respondent-preferred privacy protection, k-anonymity
doi:10.1631/FITEE.1400395 **Document code:** A **CLC number:** TP309

1 Introduction

Data collection and data publishing are two related processes where the protection of a respondent's privacy is a concern. In general, data collection involves collaboration between an authorized party (such as an individual, a private company, or a government agency) that has permission to collect personal data and a given group of respondents. The respondents then participate by answering questions from the agency. Often, data is collected for a specific purpose, such as to study a survey objective, to obtain feedback from customers regarding products and services, or to discover knowledge. The agency then releases the data in an anonymous form during the data publishing process.

Consider a scenario in which a health insurance company (agency) would like to collect some medical data from several medical institutions (respondents) to develop new insurance policies. Since medical data is highly sensitive information, we do not want to reveal to the agency the identity of any of the data owners. In this scenario, there are two main paradigms that can be used to protect the patient's privacy. The first paradigm relies on a respondent's trust in the agency, while the second is dependent on the respondent's anonymity. If respondents are not comfortable with the agency, they may refuse to provide accurate data, and if the submitted data is not genuine, we can foresee a data utility problem because the results of the analysis will be inaccurate. With respect to the second paradigm, we need to prevent a re-identification problem. In other words, the agency should not be able to link any of the collected data to the real identity of any of the patients. A new paradigm

‡ Corresponding author

* Project supported by the Basic Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2014R1A1A2058695)

ORCID: Myung Ho KIM, http://orcid.org/0000-0002-1933-7987

for medical data sharing has been proposed by Diamond *et al*. (2009).

After the data collection process, the agency may publish some or all of the collected data for data analysts to analyze or for research purposes. The identity of each respondent should remain protected when the data is released by the agency. It is common practice to remove all explicit personal identifiable information (PII), such as the social security number and name, from the original dataset before releasing data to other parties. However, removing PII does not preserve privacy because a set of quasi-identifiers (i.e., age, gender, blood type, and religion) could be used with external information, such as a voter list, to identify individuals (Sweeney, 1997). For example, 87% of the population of the USA were found to be under a linking attack where identities could be determined by using only quasi-identifiers such as gender, zip code, and date of birth (Sweeney, 2002).

Data anonymization is an interesting solution to protect the identities of respondents during the data publishing process. Samarati (2001) proposed a k-anonymity model to address the linking attack. The concept of k-anonymity is that each datum released is indistinct from at least $k-1$ other data. Several approaches have since been proposed to improve the k-anonymity model, and we will discuss some of these in Section 2.

In most of the literature, data collection agencies assume that all respondents are comfortable with submitting their data. This assumption is not realistic due to the rise in privacy concerns, which has caused respondents to refuse to participate or to submit inaccurate data to agencies. If the respondents submit inaccurate data, we cannot guarantee the usefulness of the data that has been collected. Furthermore, we note that the level of anonymity (i.e., k-anonymity) that is guaranteed by the agency cannot be verified by the respondents. It is therefore difficult for respondents to ensure that their privacy is protected at level k since they generally do not have access to all of the data collected from the agency. In other words, the actual level of anonymity is fully dependent on the honesty of the agency.

We aim to address two challenges in this study. First, we want to protect the identity of each respondent from the agency before and after data collection. Second, and more importantly, we want to guarantee the usefulness of the collected data by increasing respondent confidence. In particular, each respondent is aware of the number of respondents that have met his/her constraint before submitting data to the agency. Hence, respondents are more likely to submit genuine records when they achieve their preferred level of anonymous protection.

The first challenge can be solved by using anonymity technology such as the onion routing (Tor) (Dingledine *et al.*, 2004), an anonymous proxy server (Edman and Yener, 2009), or a mixed network (Chaum, 1981). These technologies are still under active study and focus mainly on network traffic analysis, anonymous communication channels, and private information retrieval. Since our aim in this paper is not to design any specific anonymity technology, we refer readers to Edman and Yener (2009) and Li *et al.* (2011) for a description of the use of these technologies.

The second challenge requires each respondent should help others preserve their own privacy. This idea is motivated by the concept of co-privacy (Domingo-Ferrer, 2010; 2011). Co-privacy (or cooperative privacy) considers the best option for a party to obtain privacy protection by helping other parties achieve theirs. The formal definition of co-privacy and its generalizations are given by Domingo-Ferrer (2010).

In this study, we propose a respondent-preferred anonymous protection model that facilitates privacy protection for respondents. Instead of placing their full trust in an agency, we allow each respondent i to decide the preferred k_i-anonymity during data collection without interference from the agency. Our solution can be used by the agency to apply additional models that improve the anonymous protection of each respondent before releasing data for analysis. We summarize our contributions as follows:

1. We propose a new approach called k_i-anonymity, where k_i is the level of anonymity preferred by respondent i. The respondents are therefore able to determine their anonymous protection levels before submitting their data to the agency. Note that the value of k_i is unknown to other parties.

2. Our notion of a respondent-preferred anonymous protection level aims to increase the confidence of the respondents with respect to their privacy during the data collection process. Hence,

each respondent submits genuine data to the agency while the agency can ensure the usefulness of the collected data.

3. Our protocol conforms to a Nash equilibrium model where a rational player who deviates from the computation cannot gain additional benefits from honest players.

2 Background and definitions

2.1 Preliminaries

In this section, we first explain the concept of a quasi-identifier and *k*-anonymity. Then we state the definitions of Nash equilibrium and co-privacy proposed by Domingo-Ferrer (2010). We determine the table of respondents (T) and the table that consists of all records collected by the agency (T'). For simplicity, we further assume that each respondent comprises only one tuple in T.

Definition 1 (Quasi-identifier) A quasi-identifier (QI) is a set of attributes that can uniquely distinguish tuples in T.

Definition 2 (*k*-anonymity requirement) Each release of data must be such that every tuple of QI in T' can be indistinctly matched to at least k respondents.

Definition 3 (*k*-anonymity) T' is said to satisfy the *k*-anonymity with respect to QI if, and only if, each set of attributes in QI has at least k occurrences in T'.

Definition 2 is a requirement that is used to define the *k*-anonymity such that no released data can be linked to the identity of the respondent. Under Definition 3, each respondent among $k-1$ other respondents is anonymous.

In this paper, we consider our protocol a game $\Gamma(P, A, U)$, such that $P=\{P_i\}_{i\in N}$, where $N=\{1, 2, \ldots, n\}$ is a finite set of players, $A=\times_{i\in N}A_i$, where A_i is a set of possible actions of P_i, and $U=\{u_i\}_{i\in N}$ is the utility function set of the players. The game is played between a data collector C and a set of n respondents $\{R^1, R^2, \ldots, R^n\}$. In general, the outcome a (e.g., the payoff) of the game for each player will be different. The utility function u_i of player i shows the player's preference over outcome a. In other words, P_i prefers an outcome a over outcome a' if, and only if, $u_i(a)>u_i(a')$. Each player has a set of possible strategies (i.e., σ_i). We denote $\bar{\sigma}=(\sigma_0, \sigma_1, ..., \sigma_n)$ as the joint strategies that are selected by all players. Note that σ_0 represents the strategies selected by C. Furthermore, we use $\bar{\boldsymbol{\sigma}}_{-i}$ to denote a vector consisting of all strategies in $\bar{\sigma}$ except σ_i.

Definition 4 (Nash equilibrium) Let $\Gamma(P, A, U)$ be as above, and let $u_i(\bar{\sigma})$ denote player i's expected utility if $\bar{\sigma}$ is played. A joint strategy $\bar{s}$ is a Nash equilibrium if $u_i(\bar{\boldsymbol{\sigma}}_{-i}, \sigma_i)\geq u_i(\bar{\boldsymbol{\sigma}}_{-i}, \sigma_i')$. In other words, no player i has any incentive to change his/her strategy to an alternative strategy σ_i'.

Definition 5 (Co-privacy) Let Π be a protocol with n respondents R^1, R^2, …, R^n and a data collector C. Each respondent reveals partial information from his/her private data (e.g., a constraint c_i) to others before the protocol begins. Protocol Π is said to be co-private if all respondents help respondent i achieve his/her preferred utility function.

Definition 5 states that if a privacy-preservation problem can be solved by a co-private protocol, the advantage is that it is in a player's rational privacy interest to help other players preserve their privacy. Note that co-privacy has been proven a Nash equilibrium.

2.2 *k*-anonymity model

For ease of explanation, consider the sample of medical data in Table 1. PatientID is the personal identifiable information (PII) that is used to uniquely identify each patient. The quasi-identifier is QI={Gender, Age, Zip code}, and the test result is a sensitive attribute. To protect the profile of each patient, the PatientID will be removed by each respondent before sending his/her record to the agency for simplicity. We assume that each respondent holds only one tuple in Table 1.

After the data collection process, the agency releases the collected data for analysis. Techniques such as generalization and suppression (Sweeney 2002) can be used to provide *k*-anonymity in the released data. For example, the last two digits of the zip code can be replaced by unknown values ('*'). Table 2 shows a sample table (RT) released by an agency during the data publishing process.

In Table 2, the values for the quasi-identifier for each t_i in RT appear at least twice. For instance, t_1[QI]=t_8[QI]=t_9[QI]=t_{10}[QI], t_2[QI]=t_3[QI], t_4[QI]=t_5[QI], and t_6[QI]=t_7[QI]. Therefore, we say that

Table 2 adheres to k-anonymity (k=2). In addition, we observe that the value of each $A_i \in$QI in RT[QI] appears at least k times. For example, |RT[Gender='Male']|=5, |RT[Gender='Female']|=5, and |RT[Age='61']|=2.

Table 1 Sample medical dataset

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

Table 2 Table RT adhering to k-anonymity (k=2)

PII	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

PII: personal identifiable information. M: male; F: female

2.3 Enhanced k-anonymity models

Recently, several attacks, including the minimality attack (Wong *et al*., 2007a), inference attack (Wong *et al*., 2007b), skewness and similarity attack (Li *et al*., 2007), and homogeneity and background knowledge attack (Machanavajjhala *et al*., 2007), have been identified in the k-anonymity model. Therefore, a table released during data publishing that satisfies k-anonymity cannot provide sufficient protection.

Techniques such as (α, k)-anonymity (Wong *et al*., 2006; 2007b), l-diversity (Machanavajjhala *et al*., 2007), and t-closeness (Li *et al*., 2007) have been proposed to improve the k-anonymity model. These techniques assume that k-anonymity has been achieved before applying additional techniques to improve the anonymity of the released data. For instance, the (α, k)-anonymity model assumes that all released data adheres to k-anonymity. In addition, the frequency of the sensitive value in any quasi-identifier is required to be less than α after anonymization (Wong *et al.*, 2006). In the l-diversity model, sensitive attributes in the k-anonymous table are well represented by the value of l such that each sensitive value is at most 1/l.

3 Related work

Many techniques have been proposed to allow an agency to collect sensitive data from a vast number of respondents. One of the possible solutions is to use a trusted third party (TTP). The TTP will collect data from all respondents and then provide the query service to the agency. This solution is straightforward but cannot guarantee respondents protection from leakage of the information to external adversaries or to the agency during the query process. Furthermore, a fully trusted party is not always available.

When dealing with untrusted parties, respondents may not want to provide sensitive data or may submit inaccurate data to the agency. We can address this problem by using the randomized response technique proposed by Warner (1965). The original idea for this technique was to make it difficult for the untrusted party to learn the exact response of the respondent. Statistical techniques are applied to randomly change the original data provided by the respondent before submission. A randomization technique, such as data perturbation, introduces some noise to protect the privacy of the respondents. This approach has received a great amount of attention for privacy preserving data mining (PPDM) (Agrawal and Srikant, 2000; Du and Zhan, 2003; Kargupta *et al*., 2003). In PPDM, the perturbed data is used to extract hidden information, such as the association rule (Evfimievski *et al*., 2002) and decision trees (Agrawal and Srikant, 2000; Du and Zhan, 2003), without compromising the privacy of the sensitive

data. This technique is somewhat efficient, but there is a tradeoff between the respondent's privacy and the accuracy of analysis results (Zhang *et al*., 2005).

Kumar *et al.* (2010) proposed a first respondent-defined privacy protection (RDPP) scheme to anonymously collect data. The basic idea of RDPP is to allow respondents to specify the level of protection they require before providing any data to the agency. For instance, a number of respondents (minimum threshold) must satisfy a constraint chosen by respondent i before he/she agrees to submit data. In their protocol, the respondents are aware of the minimum level of privacy protection they will receive before submitting their dataset to the agency. Instead of relying on the agency to guarantee the protection of their privacy, the respondents are free to define their preferred protection level. The RDPP protocol can be summarized as follows:

Phase 1: anonymous constraint submission. Each respondent i (with a private identifier, id_i) determines a constraint c_i together with a threshold value τ_i, and sends (id_i, c_i, τ_i) to the agency. The agency then shuffles these messages and posts them to a shared location (i.e., a website).

Phase 2: constraint satisfaction. Each respondent i examines the constraints and anonymously sends (id_i, $x_i \in c_j$) to the agency. Note that x_i is the dataset of respondent i that satisfies c_j.

Phase 3: invitation determination. The agency determines the optimal set $\mathcal{I}$ of the respondents for each constraint. All private identifiers in $\mathcal{I}$ will be posted to the shared location.

Phase 4: anonymous data submission. The respondents in $\mathcal{I}$ recognize their private identifiers and submit their dataset to the agency.

The RDPP gives us a new research direction to anonymously collect data. However, there are some limitations in using the proposed scheme:

1. Each respondent is required to reveal his/her record and the threshold to the agency before the data collection process. The agency decides which respondents to invite to submit their data.

2. The threshold of each respondent might disclose new information to C.

3. The scheme is designed to satisfy a single threshold. If the respondent wants to change his/her threshold, he/she needs to restart the scheme.

4 Our protocol

4.1 Protocol idea and design goals

The idea behind the design of our protocol is to allow each respondent to learn the number of occurrences (satisfaction score) of his/her records in T'. Our protocol uses the following main components:

Agency C: an authorized party who wants to collect data from a group of respondents.

Respondent R^i: the participant in the data collection process, who is also a candidate to submit his/her record to the agency.

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

We assume n respondents participate in the protocol, and let c_i denote the constraint chosen by R^i. We further assume that the total number of respondents n is public knowledge. Other respondents will help R^i examine if their records satisfy constraint c_i. Agency C is responsible for managing the satisfaction table of the constraints that consists of the scores for each constraint. Then each respondent i computes the satisfaction score S_i of his/her record based on the scores in the satisfaction table of the constraints. If the satisfaction score is higher than k_i occurrences, respondent i will submit his/her record to the agency. Otherwise, his/her records will not be revealed to the agency.

We assume that all the respondents are willing to help each other to gain a mutual benefit. In other words, our protocol requires each respondent help others preserve their privacy. This assumption is important in order for each respondent to learn the correct satisfaction score and to prevent leakage of sensitive information to the agency. This idea was first suggested by Wong and Kim (2014a) to provide self-awareness protection during the data collection process. Also, they considered it for use in Internet of things (IoT) data collection (Wong and Kim, 2014b).

At the end of the protocol execution, the agency may use the data collected from those who achieved their preferred level of anonymity for the following purposes:

1. To publish the data: provide the data that has been collected to external parties who are interested in obtaining the information.

2. For self-interest: analyze the dataset without disclosing any information to external parties.

3. To support a query-answer: help respondents answer statistical queries (e.g., mean, median, sum, and standard deviation).

The different uses for the collected data lead to different security and privacy concerns. In this paper, we assume that the agency will use the collected data for a local analysis to support a query-answer system. Hence, it will not release any data to an external party. Note that only those who have achieved their preferred level of anonymity can send a query to the agency. This is a fair condition because they have contributed data during the data collection process. For instance, we can think of a situation where a group of taxpayers (i.e., a corporation) submits private tax information to an authorized tax agent. The tax agent then analyzes the collected data and answers queries in terms of tax statistics (e.g., tax revenues). Other corporations who do not participate during data collection cannot request the tax agent to answer their queries.

Based on the ideas that have been discussed, we now formally define our respondent-preferred k_i-anonymity as follows:

Definition 6 (Constraint c_i) It is a possible set of constraints (for each attribute in QI) defined by respondent i to be satisfied by tuples in T'.

Definition 7 (Respondent-preferred k_i-anonymity) Each respondent i is said to prefer k_i-anonymity with respect to c_i if the attributes in QI that satisfy all constraints in c_i appear at least with k_i occurrences in T'.

Definition 7 states that each respondent is able to choose his/her preferred k_i-anonymity protection. The protection level should be independent such that two respondents may have the same or different protection levels, but they do not know the anonymity level preferred by each other.

4.2 Unique identity and constraint generations

To initiate the data collection process, the agency broadcasts a submission request to n respondents, and the submission request consists of information on T' (e.g., attribute information).

To participate in the data collection process, each respondent i generates a cryptosystem key pair (pk_i, pr_i) and a unique identity $\mathcal{I}_i$. This unique identity does not contain any profile information of the respondent and is used to retrieve the constraint scores published by the agency. Note that the unique identity can be generated by encrypting the PII (i.e., the PatientID).

Next, each respondent i generates a constraint c_i for his/her record in T' by specifying a set of rules such as {Male; 40–80; 156**}. This constraint indicates that the record is from a male patient aged 40 to 80 who lives in area with a zip code that begins with 156. Note that some of the constraints chosen by the respondents may be the same. To obtain a highly sensitive record, the respondent can set a more general constraint whereas for a less sensitive record, a more specific constraint can be used. Operators such as AND, OR, and ANY can be used to define a constraint. Each respondent i sends the constraint message ($\mathcal{I}_i$, pk_i, c_i) anonymously to the agency via the Tor network.

4.3 Constraint satisfaction checking

After receiving all constraint messages from all respondents R^1, R^2, …, R^n, agency C publishes ($\mathcal{I}_1$, pk_1, c_1), ($\mathcal{I}_2$, pk_2, c_2), …, ($\mathcal{I}_n$, pk_n, c_n) online. By checking the unique identity $\mathcal{I}_i$ in the published list, each respondent i first verifies that his/her constraint c_i has reached the agency unchanged. If the constraint is different from the submitted version, the respondent has a choice of either re-submitting data or quitting the data collection process.

Assume all constraints have correctly reached the agency. Each respondent i then examines if his/her record in T' satisfies the constraints of the other respondents. Respondent i increases s_j^i by 1 when his/her record satisfies constraint c_j. We denote s_j^i for j=1, 2, …, n as the constraint scores for c_j that determine each R^i. Next, each R^i encrypts s_1^i, s_2^i, ..., s_n^i by using a public key pk_j of the corresponding constraint owner to produce $\mathrm{Enc}_{\mathrm{pk}_1}(s_1^i)$, $\mathrm{Enc}_{\mathrm{pk}_2}(s_2^i)$, ..., $\mathrm{Enc}_{\mathrm{pk}_n}(s_n^i)$. Each R^i sends $\{\mathcal{I}_i, \varepsilon_i\}$ to the agency, where

$$\varepsilon_i = \{\mathrm{Enc}_{\mathrm{pk}_1}(s_1^i), \mathrm{Enc}_{\mathrm{pk}_2}(s_2^i), ..., \mathrm{Enc}_{\mathrm{pk}_n}(s_n^i)\}.$$

4.4 Computation of satisfaction scores

At this point, the agency manages a constraint satisfaction table (Table 3). The columns in Table 3

represent the encrypted scores for each constraint, whereas each row represents the encryption scores computed by each respondent.

Each R^i first verifies that the published ε_i is the same as the original list sent to the agency. If the list has been modified, the respondent will quit the data collection process. In the next step, each R^i retrieves the ciphertexts that correspond to his/her unique identity $\mathcal{I}_i$. For instance, respondent R^1 with a unique identity $\mathcal{I}_1$ retrieves $\mathrm{Enc}_{\mathrm{pk}_1}(s_1^1)$, $\mathrm{Enc}_{\mathrm{pk}_1}(s_1^2)$, …, $\mathrm{Enc}_{\mathrm{pk}_1}(s_1^n)$ from Table 3.

Table 3 Constraint satisfaction table

Respondent	ε_1	ε_2	…	ε_n
1	$\mathrm{Enc}_{\mathrm{pk}_1}(s_1^1)$	$\mathrm{Enc}_{\mathrm{pk}_1}(s_1^2)$	…	$\mathrm{Enc}_{\mathrm{pk}_1}(s_1^n)$
2	$\mathrm{Enc}_{\mathrm{pk}_2}(s_2^1)$	$\mathrm{Enc}_{\mathrm{pk}_2}(s_2^2)$	…	$\mathrm{Enc}_{\mathrm{pk}_2}(s_2^n)$
…	…	…	…	…
n	$\mathrm{Enc}_{\mathrm{pk}_n}(s_n^1)$	$\mathrm{Enc}_{\mathrm{pk}_n}(s_n^2)$	…	$\mathrm{Enc}_{\mathrm{pk}_n}(s_n^n)$

After decryption, respondent R^1 computes the satisfaction score, $S_1=\sum_{j=1}^{n}(s_1^j)$, for his/her constraint c_1. At this stage, if S_1 is the same as or greater than a preferred anonymous protection level (e.g., k_i), we assume that R^1 will submit his/her records to C. If the respondent's preferred protection level has not been achieved ($S_i<k_i$), then R^1 will not reveal his/her records to C. For both cases, a decision message m_i will be sent to the agency as follows:

$$m_i=\begin{cases}(\mathcal{I}_i, 1), & S_i\geq k_i,\\ (\mathcal{I}_i, 0), & S_i<k_i.\end{cases}$$

4.5 Updates for the constraint satisfaction table

When a decision message $m_i=(\mathcal{I}_i, 0)$ is received from R^i, the agency updates the constraint satisfaction table by removing the constraint c_i (the row with respondent i's unique identity $\mathcal{I}_i$) and ε_i (column) from Table 3. The agency releases the updated constraint satisfaction table online, and all of the remaining respondents recalculate the satisfaction score of their constraint of choice. A summary of our protocol is shown in Algorithm 1 and the workflow of our proposed solution in Fig. 1.

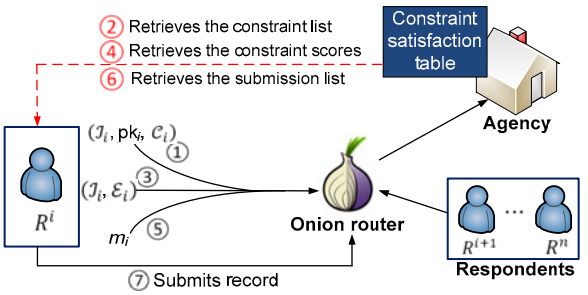

Fig. 1 Workflow of the proposed solution
Steps 6 and 7 are performed only when the anonymity level preferred by the respondent has been achieved

Algorithm 1 Respondent-preferred k_i-anonymity

Phase 1: constraint submission. The agency broadcasts a submission request to n respondents. Each respondent i generates a cryptographic key pair (pk_i, pr_i) and specifies its constraint c_i. Next, R^i sends ($\mathcal{I}_i$, pk_i, c_i) to the agency via the Tor network. Note that $\mathcal{I}_i$ is the unique identity of R^i. The agency then publishes $\{(\mathcal{I}_i, \mathrm{pk}_i, c_i)|i=1, 2, \ldots, n\}$ online.

Phase 2: constraint score computation. Each respondent i examines if his/her record in t_i satisfies the constraint c_j. For the case where it is satisfied, R^i increases the constraint score s_j^i by 1. The owner of the constraint will also determine its constraint score (i.e., the owner of c_j sets the value for s_j^i). Next, R^i encrypts $\{s_j^i \mid j=1, 2, ..., n\}$ by using the public key pk_j of the corresponding constraint owner to produce $\varepsilon_i=\{\mathrm{Enc}_{\mathrm{pk}_j}(s_j^i) \mid j=1, 2, ..., n\}$. At the end of this phase, R^i anonymously sends ($\mathcal{I}_i$, ε_i) to the agency.

Phase 3: score list verification. The agency publishes ($\mathcal{I}_i$, ε_i) online. Each R^i examines if the published constraint score list remains the same as the original list he/she sent to the agency. If the list has been modified, the respondent will not participate in the next phase.

Phase 4: preferred anonymity checking. Each R^j retrieves and decrypts $\{\mathrm{Enc}_{\mathrm{pk}_j}(s_j^i) \mid j=1, 2, ..., n\}$. Next, it computes $S_j=\sum_{i=1}^{n}s_j^i$ as the satisfaction score for the chosen constraint. If the satisfaction score S_j has at least k_i occurrences, R^j sends ($\mathcal{I}_j$, 1) to the agency. Otherwise, ($\mathcal{I}_j$, 0) will be sent to the agency.

Phase 5: satisfaction score table update. The agency needs to update the satisfaction score table when one or more respondents have a response with ($\mathcal{I}_j$, 0). The update process is performed by removing the constraint and the constraint scores of the respondent. This process will be repeated until all the respondents have confirmed their decisions to the agency.

Phase 6: record submission. The agency releases the unique identity of the respondents who are willing to submit their records. Those on the list will send their records to the agency with the confidence that their desired level of privacy protection has been achieved at the k_i-anonymity level.

5 Analysis and discussion

5.1 Correctness and anonymity analysis

At the end of Phase 1 in Algorithm 1, the agency publishes the constraint with the unique identity of each respondent. Each respondent i then determines if his/her constraint has reached the agency unchanged by checking his/her identity $\mathcal{I}_i$. If the published constraint is different from the submitted version, the respondent has the choice to either resubmit his/her constraint message or quit the data collection process. This step is important to ensure that the satisfaction scores that are computed are correct.

In Phase 3, the agency must publish ε_i correctly to ensure that all respondents continue to participate in the protocol.

In Phase 4, the agency updates the constraint satisfaction table when one or more respondents have decided not to submit their data. For every round of computation, the satisfaction score either remains unchanged or is reduced. If the satisfaction score has been reduced, the respondents who agreed in the previous round may want to make a new decision. Once all respondents have confirmed their decisions (with no further changes), the agency publishes the private identities of those respondents who want to submit their records. If one of the respondents fails to communicate with the agency during the protocol execution (due to some unexpected problems such as a network failure), the correctness of our protocol will not be affected.

For our protocol design, we assume that respondents are communicating with the agency using the Tor network. Since the Tor network is designed to conceal personal identity information, the profile of the respondents will remain anonymous to the agency. Furthermore, we can ensure that the data collected by the agency already adheres to the k_i-anonymity. The agency can apply additional steps to improve the anonymity of the collected data before it is released during the data publishing process.

5.2 Privacy analysis

The privacy analysis of our protocol depends on how much information has been revealed during the execution of the protocol. In our protocol, we aim to protect two sensitive pieces of information: (1) the profile of the respondents and (2) their preferred anonymous protection levels k_i ($i\geq 2$). In particular, we want to prevent any party from inferring this sensitive information from intermediate results or the final output. In the constraint generation phase, each respondent needs to ensure that his/her constraint does not contain any specific information or reveals too much information, allowing other parties to infer his/her identity. Therefore, a party who cross-examines the published constraints cannot learn any specific information about the respondents. For instance, if the agency colludes with some respondents, the agency cannot determine who owns any of the collected records. In addition, the unique identity $\mathcal{I}_i$ of each respondent will not serve to leak the profile of any respondent because these are stored in an encrypted form.

In Phase 3, the agency is not able to decrypt ε_i in the absence of private keys from the respondents. Furthermore, our protocol ensures that no party (including the agency) can identify the satisfaction score of any respondent because the computation for S_i requires the private key of the constraint owner. In our protocol, respondents do not interact directly with each other. This design prevents possible collusion among the respondents to learn sensitive information about others. Since we assume that all data transmissions have been performed using an anonymous communication channel (e.g., the Tor network), we can ensure that the identity of each respondent remains anonymous to others. At the same time, the final data that is collected should adhere to the preferred protection level of each respondent who submits his/her record.

5.3 Security analysis

In this subsection, we will analyze our protocol using a Bug model (Bella *et al*., 2005). Bug is a realistic threat model that divides participants (the agency and respondents) into three categories according to the following behaviors:

Good: all participants follow the protocol faithfully.

Bad: the participant does not follow the protocol properly and may act maliciously to obtain a personal advantage.

Ugly: the participant changes his/her behavior to 'good' or 'bad' during the execution of the protocol. For instance, the participant may change to 'Bad' behavior when he/she sees a chance to gain advantage from the protocol outcome.

When agency C is a 'good' participant, it will release correct quasi-identifier information in Phase 1 and will publish a correct and updated constraint satisfaction table. If C is 'bad', it may release fake quasi-identifier information in Phase 1 to mislead respondents so that they will submit sensitive records during Phase 2. Similarly, a 'bad' C may try to modify or encrypt a fake satisfaction score during the cooperative phase. However, the respondents are able to detect this 'bad' behavior because for each release, they will first verify if their submissions have reached C unchanged by checking the released data.

During the protocol execution, an 'ugly' C may first follow the protocol faithfully (e.g., release correct quasi-identifier information) but then change to 'bad' (e.g., modifying the satisfaction scores) during the cooperative phase to collect sensitive records from the respondents. After the data is collected, a 'bad' or 'ugly' agency may sell or publish data to a third party. During our protocol design, the respondents can choose their preferred anonymous protection level before they submit their sensitive records to C. Therefore, respondents can be confident that their data published by a 'bad' or 'ugly' C still adheres to k_i-anonymity.

When respondents are 'good' participants, they will submit their records accordingly. In particular, respondents will submit sensitive records when their preferred anonymous protection level has been achieved. If they are 'bad', they may not want to reveal any record to C by claiming that all (or the majority) of their data contains sensitive records and that their preferred anonymous protection levels have not yet been achieved. It is clear that by acting in such a manner, a 'bad' respondent can avoid submitting any data. 'Ugly' respondents submit only non-sensitive records during Phase 1 and do not participate in subsequent phases.

The use of the Bug model in our security analysis is appropriate because it allows situations in which participants do not collude but act maliciously to obtain a personal benefit (Bella *et al*., 2011). For instance, a 'bad' or 'ugly' C does not need to collaborate with any respondent to achieve its goal. In addition, we can see that respondents have no interest in collaborating with others because their ultimate goal is to achieve their preferred anonymous protection level.

5.4 Efficiency analysis

The complexity of our protocol is dominated by cryptographic operations (encryption and decryption) that have been performed by respondents. We implemented the cryptographic operations, based on Paillier (1999), in Java and ran the program on a single computer with a 2 GHz CPU and 2 GB of RAM. As expected, the respondent's computation time was proportional to the constraint size (Fig. 2). Note that the constraint size is the number of constraints processed by each respondent.

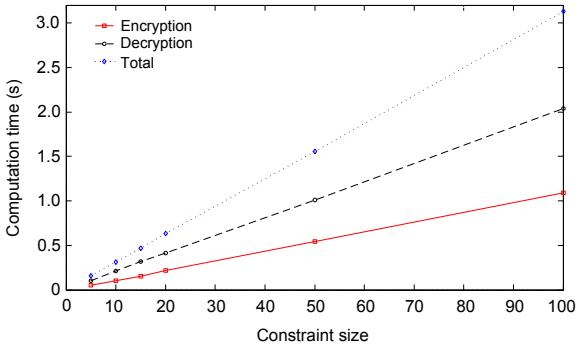

Fig. 2 Computation costs incurred by the respondent

Fig. 2 shows the computation costs incurred by the respondent in Phase 2, in which each respondent needs to examine the constraints that were published by the agency. In other words, they will conduct $n-1$ queries on their database. However, we note that it is

possible to simplify this operation by applying some simple filtering algorithms.

5.5 Comparison with the *k*-anonymity model

Our notion of k_i-anonymity is different from the concept of a conventional *k*-anonymity model. Unlike conventional *k*-anonymity, we do not use the same anonymity level (e.g., *k*-anonymity) for all the respondents. Instead, we allow each respondent to decide his/her preferred level of protection. Our objective is to increase the confidence of the respondents after they have submitted their data to the agency.

The ultimate aim of a conventional *k*-anonymity model is to release data with a scientific guarantee that individuals who have submitted data cannot be re-identified while ensuring that the data remains practically useful (Sweeney, 2002). However, the data owners are generally not able to verify the protection level that is offered (e.g., *k*-anonymity) by the agency to third parties.

The solution proposed in this paper is based on the *k*-anonymity model that is often used as the basic component to construct other anonymity models, such as *t*-closeness and *l*-diversity. We extend *k*-anonymity to k_i-anonymity to allow each respondent to choose his/her preferred level of protection. Our protocol design allows each respondent to verify whether their preferred level of protection has been achieved before submitting their data to *C*. In future work, we aim to extend other anonymity models, such as *t*-closeness and *l*-diversity, to support respondent-preferred protection (e.g., t_i-closeness and l_i-diversity).

In terms of privacy protection, our k_i-anonymity model achieves the same level of protection as a conventional *k*-anonymity model. However, our solution increases the confidence of the respondents by allowing them to determine their preferred level of protection. This property is very important since the rise in privacy concerns has caused some respondents to refuse to participate and others to submit inaccurate data to agencies. If respondents submit inaccurate data, we cannot guarantee the usefulness of the results obtained from the data analysis.

5.6 Discussion

In our protocol, each respondent compares the satisfaction score of his/her constraint with k_i. If the result of the comparison has been satisfied, respondent *i* will inform the agency about his/her decision to submit the data. However, if the respondent decides to increase the value of k_i, he/she can still send a new decision to the agency before the data submission phase.

As discussed in Section 3, the RDPP solution requires all respondents reveal their threshold values (preferred protection level) to the agency. The agency then computes and decides which respondents to invite to submit their data. In RDPP, the respondents can define only one threshold value for each protocol execution. In other words, if the respondent wants to change his/her threshold value, he/she needs to carry out the process from the beginning. Therefore, our solution differs from that proposed by Kumar *et al.* (2010) in two respects: (1) we do not reveal the preferred anonymous protection level k_i of each respondent *i* to the agency, and (2) our protocol allows the respondents to change the value of k_i during the protocol execution.

In this paper, we do not consider differential privacy (Dwork, 2008) for our threat model because we do not intend to introduce noise into the respondents' records. Rather, our main objective is to allow the agency to receive genuine records from the respondents. A discussion of syntactic approaches (e.g., *k*-anonymity) and differential privacy was given by Clifton and Tassa (2013).

The goal of a rational multiparty computation framework is to relax the requirements of malicious and semi-honest models. The malicious model must protect against all deviations from the protocol specification, including actions that do not give an advantage to an adversary. The secure protocols in the semi-honest model achieve a greater efficiency, but suffer from the strong assumption that parties will not deviate from the protocol even if they may benefit from doing so. As we have described in Section 5, our framework requires only that parties follow the protocol if such an action constitutes rational behavior. We argue that the assumption of rationality is more plausible than the blind obedience required in the semi-honest model, and the resulting protocols will be more efficient than their malicious model counterparts by preventing arbitrary (i.e., non-rational) actions. Perhaps most critically, protocols that are secure under a malicious model do not prevent a party

from lying about their input, and rational behavior provides a means to incorporate this into the discussion by ensuring that the results reflect the true data.

Recently, a concept known as co-utility (a generalization of co-privacy) was proposed by Domingo-Ferrer *et al*. (2015). Two types of self-enforcing protocols have been proposed: (1) a coordination protocol and (2) a co-utile protocol. We note that our solution in this paper can be categorized as a coordination protocol, and in future work, we aim to design a co-utile protocol to collect data.

6 Conclusions

In this paper, we introduced the concept of k_i-anonymity that allows respondents to choose their preferred anonymous protection level before submitting data to an agency. Our protocol ensures that the respondents achieve their preferred k_i-anonymity before the data is submitted, and that the usefulness of the collected data (i.e., the data utility) can be guaranteed. In our protocol, the agency helps respondents compute a satisfaction score for each constraint through encrypted means. The respondents then need to ensure that no sensitive information can be inferred from their constraints. Each respondent learns the satisfaction score only for his/her constraint without having any knowledge of that of the others. At the end of data collection, the data released for each respondent will achieve k_i-anonymity. Unlike the RDPP scheme, we do not need the respondents to reveal their preferred value of k_i to know whether their constraints have been satisfied by others. Instead, we design our scheme such that the respondent will learn the number of respondents that have satisfied their preferred anonymity level before making a decision to provide data to the agency.

References

Agrawal, R., Srikant, R., 2000. Privacy-preserving data mining. Proc. ACM SIGMOD Int. Conf. on Management of Data, p.439-450. [doi:10.1145/342009.335438]

Bella, G., Bistarelli, S., Massacci, F., 2005. Retaliation: can we live with flaws? *NATO Sec. Sci. Ser. D*, **6**:3-14.

Bella, G., Giustolisi, R., Riccobene, S., 2011. Enforcing privacy in e-commerce by balancing anonymity and trust. *Comput. Secur.*, **30**(8):705-718. [doi:10.1016/j.cose.2011.08.005]

Chaum, D.L., 1981. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM*, **24**(2):84-90. [doi:10.1145/358549.358563]

Clifton, C., Tassa, T., 2013. On syntactic anonymity and differential privacy. Proc. IEEE 29th Int. Conf. on Data Engineering Workshops, p.88-93. [doi:10.1109/ICDEW.2013.6547433]

Diamond, C.C., Mostashari, F., Shirky, C., 2009. Collecting and sharing data for population health: a new paradigm. *Health Aff.*, **28**(2):454-466. [doi:10.1377/hlthaff.28.2.454]

Dingledine, R., Mathewson, N., Syverson, P., 2004. Tor: the second-generation onion router. Proc. 13th Conf. on USENIX Security Symp., p.21.

Domingo-Ferrer, J., 2010. Coprivacy: towards a theory of sustainable privacy. Proc. Int. Conf. on Privacy in Statistical Databases, p.258-268. [doi:10.1007/978-3-642-15838-4_23]

Domingo-Ferrer, J., 2011. Coprivacy: an introduction to the theory and applications of co-operative privacy. *Stat. Oper. Res. Trans.*, Special issue, p.25-40.

Domingo-Ferrer, J., Soria-Comas, J., Ciobotaru, O., 2015. Co-utility: self-enforcing protocols without coordination mechanisms. Proc. Int. Conf. on Industrial Engineering and Operations Management, arXiv:1503.02563.

Du, W., Zhan, Z., 2003. Using randomized response techniques for privacy-preserving data mining. Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.505-510. [doi:10.1145/956750.956810]

Dwork, C., 2008. Differential privacy: a survey of results. Proc. 5th Int. Conf. on Theory and Applications of Models of Computation, p.1-19. [doi:10.1007/978-3-540-79228-4_1]

Edman, M., Yener, B., 2009. On anonymity in an electronic society: a survey of anonymous communication systems. *ACM Comput. Surv.*, **42**(1), Article 5. [doi:10.1145/1592451.1592456]

Evfimievski, A., Srikant, R., Agrawal, R., *et al.*, 2002. Privacy preserving mining of association rules. Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.217-228.

Kargupta, H., Datta, S., Wang, Q., *et al.*, 2003. On the privacy preserving properties of random data perturbation techniques. Proc. 3rd IEEE Int. Conf. on Data Mining, p.99-106. [doi:10.1109/ICDM.2003.1250908]

Kumar, R., Gopal, R., Garfinkel, R., 2010. Freedom of privacy: anonymous data collection with respondent-defined privacy protection. *INFORMS J. Comput.*, **22**(3):471-481. [doi:10.1287/ijoc.1090.0364]

Li, B., Erdin, E., Güneş, M.H., *et al.*, 2011. An analysis of anonymity technology usage. Proc. 3rd Int. Conf. on Traffic Monitoring and Analysis, p.108-121.

Li, N., Li, T., Venkatasubramanian, S., 2007. *T*-closeness: privacy beyond *k*-anonymity and *l*-diversity. Proc. 23rd Int. Conf. on Data Engineering, p.106-115. [doi:10.1109/ICDE.2007.367856]

Machanavajjhala, A., Kifer, D., Gehrke, J., *et al.*, 2007. *L*-diversity: privacy beyond *k*-anonymity. *ACM Trans.*

Knowl. Discov. Data, **1**(1), Article 3. [doi:10.1145/1217299.1217302]

Paillier, P., 1999. Public-key cryptosystems based on composite degree residuosity classes. Proc. 17th Int. Conf. on Theory and Application of Cryptographic Techniques, p.223-238. [doi:10.1007/3-540-48910-X_16]

Samarati, P., 2001. Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, **13**(6): 188-200. [doi:10.1109/69.971193]

Sweeney, L., 1997. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics*, **25**(2-3): 98-110. [doi:10.1111/j.1748-720X.1997.tb01885.x]

Sweeney, L., 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzz. Knowl.-Based Syst.*, **10**(5):557-570. [doi:10.1142/S0218488502001648]

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.*, **60**(309):63-69. [doi:10.1080/01621459.1965.10480775]

Wong, K.S., Kim, M.H., 2014a. Privacy-preserving data collection with self-awareness protection. *In*: Park, J.J., Zomaya, A., Jeong, H.Y., *et al.* (Eds.), Frontier and Innovation in Future Computing and Communications. Springer, Netherlands, p.365-371. [doi:10.1007/978-94-017-8798-7_44]

Wong, K.S., Kim, M.H., 2014b. Towards self-awareness privacy protection for Internet of things data collection. *J. Appl. Math.*, **2014**:827959.1-827959.9. [doi:10.1155/2014/827959]

Wong, R.C.W., Li, J., Fu, A.W.C., *et al.*, 2006. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.754-759. [doi:10.1145/1150402.1150499]

Wong, R.C.W., Fu, A.W.C., Wang, K., *et al.*, 2007a. Minimality attack in privacy preserving data publishing. Proc. 33rd Int. Conf. on Very Large Data Bases, p.543-554.

Wong, R.C.W., Liu, Y., Yin, J., *et al.*, 2007b. (α, k)-anonymity based privacy preservation by lossy join. Proc. Joint 9th Asia-Pacific Web Conf. on Advances in Data and Web Management and 8th Int. Conf. on Web-Age Information Management, p.733-744. [doi:10.1007/978-3-540-72524-4_75]

Zhang, N., Wang, S., Zhao, W., 2005. A new scheme on privacy-preserving data classification. Proc. 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining, p.374-383. [doi:10.1145/1081870.1081913]