Co-Utile Collaborative Anonymization of Microdata

Jordi Soria-Comas and Josep Domingo-Ferrer

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{jordi.soria,josep.domingo}@urv.cat

Abstract. In surveys collecting individual data (microdata), each respondent is usually required to report values for a set of attributes. If some of these attributes contain sensitive information, the respondent must trust the collector not to make any inappropriate use of the data and, in case any data are to be publicly released, to properly anonymize them to avoid disclosing sensitive information. If the respondent does not trust the data collector, she may report inaccurately or report nothing at all. The reduce the need for trust, local anonymization is an alternative whereby each respondent anonymizes her data prior to sending them to the data collector. However, local anonymization by each respondent without seeing other respondents' data makes it hard to find a good trade-off minimizing information loss and disclosure risk. We propose a distributed anonymization approach where users collaborate to attain an appropriate level of disclosure protection (and, thus, of information loss). Under our scheme, the final anonymized data are only as accurate as the information released by each respondent; hence, no trust needs to be assumed towards the data collector or any other respondent. Further, if respondents are interested in forming an accurate data set, the proposed collaborative anonymization protocols are self-enforcing and co-utile.

Keywords: Information security and privacy; utility and decision theory; co-utility.

1 Introduction

A microdata file contains data collected from individual respondents. Because of the level of detail in the data, they can be useful for a variety of secondary analyses by third parties other than the data collector. However, releasing the original data is not feasible because it would lead to a violation of the privacy of respondents. Statistical disclosure control (SDC), a.k.a. statistical disclosure limitation, for microdata seeks to produce an anonymized version of the microdata file such that it enables valid statistical analyses but thwarts inference of confidential information about any specific individual.

The mainstream literature on SDC for microdata (*e.g.*see [7]) focuses on centralized anonymization, which features a trusted data collector. The data collector (e.g. National Statistical Institute) gathers original data from the respondents and takes care of anonymizing them. While avoiding the computational burden of anonymization is confortable to respondents, it has the downside that they need to trust the data collector.

Local anonymization is an alternative disclosure limitation paradigm suitable for scenarios where the respondents do not trust (or trust only partially) the data collector. Each respondent anonymizes her own data before handing them to the data collector. In comparison to centralized anonymization, local anonymization usually results in greater information loss. The reason is that each respondent needs to protect her data without seeing the other respondents' data, which makes it difficult for her to find a good trade-off between the disclosure risk limitation achieved and the information loss incurred.

1.1 Contribution and plan of this paper

To overcome the limitations of the centralized and the local anonymization paradigms, we propose the notion of *collaborative anonymization*, which is in line with the novel notion of co-utility [4]. Co-utility models the interaction among a set of peers, each one with a selfish goal, in which peers help each other rationally. For collaboration to arise rationally, the best strategy for a peer to reach her goal must be to help another peer in reaching his goal. The advantage of co-utility is that it leads to a system that works smoothly without the need of external enforcement.

The rest of the paper is organized in the following manner. Section 2 provides background on several notions that this paper builds on. Section 3 provides a brief review of related work. Section 4 lists the requirements of collaborative anonymization and justifies why it is rationally preferable to centralized and local anonymization. Section 5 describes a collaborative anonymization technique that hides each respondent within a group of respondents. Section 6 describes a collaborative anonymization technique that masks the value of the confidential data. Conclusions and future research issues are summarized in Section 7.

2 Background

2.1 *k*-Anonymity

k-Anonymity [14] is a privacy model that seeks to thwart re-identification of anonymized records. Central to k-anonymity is the notion of quasi-identifier attributes, also known as key attributes. Quasi-identifiers are attributes that, when considered separately, do not identify the respondent behind a record, but, which used in combination may allow an attacker to uniquely link that record to an external database containing identifiers (this database is the attacker's background knowledge). Such a unique linkage is called re-identification.

With above setting in mind, k-anonymity can be defined as follows.

Definition 1 (k-Anonymity). *A protected data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of values of quasi-identifier attributes, at least k records exist in the data set sharing that combination.*

If the quasi-identifiers considered by the data protector to enforce k-anonymity coincide with the quasi-identifiers that an attacker can use to link with his background knowledge, then k-anonymity reduces the probability of successful re-identification to $1/k$.

Of course, which attributes should be labeled as quasi-identifiers is debatable. At the very least, attributes that can be found in a public non-de-identified data sets (*e.g.* electoral rolls, phonebooks, etc.) must be taken as quasi-identifiers. However, this is not enough to prevent re-identification by attackers with additional knowledge.

2.2 Reverse Mapping

Reverse mapping [13, 3] is a post-masking technique that can be applied to any anonymized data set. The result is a *reverse-mapped* data set, constructed by taking each attribute of the anonymized data set at a time, and replacing the value of each record by the value in the original data set with equal rank.

Thus, reverse mapping requires knowing the marginal distribution of each of the attributes in the original data set. Hence, if the data collector wants to allow reverse mapping by parties other than himself, he must release those marginal distributions. And, for those distributions to be releasable, they must be assumed to be non-disclosive. The good news is that this is quite a reasonable assumption, as the distribution of an attribute essentially conveys statistical information (it is, in principle, unrelated to any specific individual). For the extreme cases in which a single value can be associated to a specific individual (*e.g.* the turnover of the largest company in a specific sector), prior masking of the marginal distribution would be needed (*e.g.* by top coding it).

The interesting point about the reverse mapping transformation is that it allows viewing any microdata anonymization method as being *functionally equivalent to permutation* (mapping the original data set to the reverse-mapped data set) *plus a small amount of noise* (mapping the reverse-mapped data set to the anonymized data set). The noise is necessarily small because it does not modify the ranks of the values: by construction, ranks in the reverse-mapped and the anonymized data set are the same. Therefore, *the essential anonymization principle turns out to be permutation.*

2.3 Co-Utility

Consider a set of self-interested peers (having each a utility function, that is, a specific goal or a defined preference relation between a set of possible outcomes) that act strategically (each peer acts to seek an outcome that maximizes her utility, according to her knowledge of the environment).

Co-utility [4] models a kind of interaction between the peers in which it is in the best interest of each of them to help another peer in reaching her goal. The primary advantage of a co-utile system is that it does not require any external mechanism to enforce a particular outcome or coordinate the actions of the peers.

Co-utility can be formalized using game theory. To guarantee a specific interaction outcome without external enforcement, the outcome must be self-enforcing; in game-theoretic terms, it must be an equilibrium. An outcome is an equilibrium if no agent (peer) has incentives to change her strategy in that outcome; in other words, provided that all other agents keep their strategies unchanged, no agent can increase her utility by modifying her strategy.

If the utility of an outcome for an agent depended on the preferences of another agent, attaining an equilibrium would require each agent to report her preferences. On one side, this would increase the complexity of the system, because an agent may report untruthful preferences if she believes that doing so is going to yield a better outcome for her. On the other side, gathering all agents' preferences by a specific party or agent should be avoided if we want a truly distributed interaction. Following the above rationale, we define games amenable to co-utility as those in which the utility of each agent is independent of the preferences of other agents.

Definition 2 (Co-utility amenable game). *Let G be a sequential Bayesian game for n agents. We say that G is a* co-utility-amenable game *if the utility of any agent is independent of the types of the other agents, i.e., $\forall\, i, j$, with $i \neq j$ and $\forall\, t_j, t'_j \in T_j$, we have that $u_i(s_1, \ldots, s_j, \ldots, s_n, t_1, \ldots, t_j, \ldots, t_n) = u_i(s_1, \ldots, s_j, \ldots, s_n, t_1, \ldots, t'_j, \ldots, t_n)$.*

Having defined a co-utility amenable game, we are ready to define when a protocol P that produces as output a strategy profile of the game is co-utile. An agent can be reluctant to play a strategy that is beneficial to herself if the strategy provides a much larger benefit to another agent. Because of that, different levels of co-utility can be distinguished, depending on whether agents maximize or just increase their utility by following the protocol. In *strict co-utility* each agent maximizes her utility and, thus, there is no reason for any agent to not to follow the protocol.

Definition 3 (Strict co-utility). *Let G be a co-utility amenable game for n agents. Let P be a self-enforcing protocol for G. We say P is a* strictly co-utile protocol *if $\forall\, i \in \{1, \ldots, n\}$, and $\forall\, s'_1 \in S_1, \ldots, s'_n \in S_n$ and $\forall\, t_1 \in T_1, \ldots, t_n \in T_n$, we have that $u_i(s_1, \ldots, s_n, t_1, \ldots, t_n) \geq u_i(s'_1, \ldots, s'_n, t_1, \ldots, t_n)$, where the outcome of P is $(s_1, \ldots, s_n)$.*

Designing co-utile protocols is usually a matter of finding a group of peers with a sufficiently aligned set of preferences. As the focus of this paper is on data anonymization, we consider a set of privacy-conscious peers that are required to report some data, for instance, to answer a certain survey. Now, a rational privacy-conscious peer will report false data or report no data at all *unless she has some interest in the pooled responses by all peer to be as accurate as possible.* Hence, we will make the assumption that all peers are interested in obtaining an accurate data set.

Under the above assumption, one possible approach to designing a co-utile protocol can be based on each peer hiding within a group of peers when reporting her record. Note that hiding one's identity when reporting one's record helps other peers in the group to hide their own identities. Conversely, it is hard to hide in a group where none of the other members is anonymous.

3 Related Work

This work seeks to empower each respondent to anonymize her own data while preserving utility as in the centralized paradigm.

Related works exist that consider privacy-conscious data set owners, rather than privacy-conscious respondents. When dealing with privacy-conscious data set owners, one faces a data integration problem where the data owners do not want to share data that are more specific than those in the final anonymized data set to be jointly obtained. In [17] a top-down generalization approach for two owners of vertically partitioned data sets is proposed. Both owners start with the maximum level of generalization, and they iteratively and collaboratively refine the generalization. In [8, 9] the same problem is tackled by using cryptographic techniques. In [10] the anonymization of horizontally partitioned data sets is considered. The main difference between the above proposals and our work is that the number of respondents is usually much greater than the number of data set owners (the latter are a small number in most realistic data integration settings). In our case, there is a different respondent for each data record being collected, which makes proposals oriented to a few data set owners unusable.

Among the related works specifically addressing respondent privacy, the local anonymization paradigm is closest to our approach in terms of trust requirements. Several local anonymization methods have been proposed. Many basic SDC techniques such as global recoding, top and bottom coding, and noise addition can be applied locally (check [7] for details on such techniques). There are, however, some techniques specifically designed for local anonymization that, in addition to helping a respondent to hide her response, allow the data collector to get an accurate estimation of the distribution of responses for groups of respondents. In randomized response [18], the respondent flips a coin before answering a sensitive dichotomous question (like "Have you taken drugs this month?"); if the coin comes up tails, the responder answers "yes", otherwise she answers truthfully. This protects the privacy of respondents, because the survey collector cannot determine whether a particular respondent's "yes" is random or truthful; but he knows that the "no" answers are truthful, so that he can estimate the real proportion of "no" as twice as much as the observed proportion of "no" (from which the real proportion of "yes" follows). FRAPP [1] can be seen as a generalization of random response. In FRAPP, the respondent reports the real value with some probability and, otherwise, it returns a random value from a known distribution. In AROMA [15] each respondent hides her confidential data within a set of possible confidential values drawn from some known distribution. In any case, to obtain an accurate result, the output of a query performed on the anonymized data must be adjusted according to the known distribution used to mask the actual data. While some kind of adjustment of the query results may also be needed in the centralized paradigm (*e.g.* when the generalization used for quasi-identifiers in a k-anonymous data set does not match the query), the randomness introduced by local anonymization makes the estimate less accurate than in centralized anonymization.

An advantage of local anonymization, though, is that the respondent is given some capability to decide the amount of anonymization required, which is likely to increase her disposition to provide truthful data (rather than fake data). Yet, most privacy models/techniques give uniform disclosure limitation guarantees to all respondents, which may not suit the different perceptions of disclosure risk of the various respondents. To address this concern, [20] proposed a privacy model in which each individual determines the amount of protection required for her data.

4 Collaborative Anonymization: Requirements and Justification

A problem with centralized anonymization is that, if a respondent does not trust the data collector to properly use and/or anonymize her data, she may decide to provide false data (hence causing a response bias) or not data at all (hence causing a non-response bias). Local (also known as independent) anonymization is an alternative that is not free from problems either. As argued in Section 2.2 above, permutation is essential to anonymization, but the permutation caused by a certain amount of masking depends not only on one's own record but on the values of the records of the other respondents. Hence, for a respondent anonymizing her own record in isolation it is hard to determine the amount of masking that yields a good trade-off between disclosure risk and information loss, *i.e.* that causes enough permutation but not more than enough permutation. A natural tendency is for each respondent to play it safe and overdo the masking, just in case, which incurs more information loss than necessary.

To deal with the above shortcomings of centralized and local anonymization, we propose a new paradigm that we call collaborative data anonymization. Consider a set of respondents $R_1, \ldots, R_m$ whose data are to be collected. Each respondent is asked to report information about a set of attributes (some of them containing confidential/sensitive information). Since respondents place limited trust on the data collector, they may refuse to provide the collector with non-anonymized data. A more realistic goal is to generate, in a collaborative and distributed manner, an anonymized data set that satisfies the following two requirements: (i) it incurs no more information loss than the data set that would be obtained with the centralized paradigm for the same privacy level, and (ii) neither the respondents nor the data collector gain more knowledge about the confidential/sensitive attributes of a specific respondent than the knowledge contained in the final anonymized data set.

In general, the motivations for a respondent to contribute her data are not completely clear. A rational respondent will only contribute if the benefit she gets from participating compensates her privacy loss. It is not in our hands to determine what the motivations of the respondents are. However, since our collaborative approach achieves the same data utility as the centralized approach while improving the respondent's privacy vs the data collector, any respondent willing to participate under the centralized approach should be even more willing to participate under our collaborative scheme. More precisely, we can distinguish several types of respondents depending on their interests in the collected data and in their own privacy:

- A respondent without any interest in the collected data set is better off by declining to contribute.
- A respondent who is interested in the collected data and has no privacy concerns can directly supply her data and needs no anonymization (neither local, nor centralized nor collaborative).
- A respondent who is interested in the collected data but has privacy concerns will prefer the collaborative approach to the centralized and the local approaches. Indeed, the collaborative approach outperforms the centralized approach in that the former offers privacy vs the data collector. Also, the collaborative approach outperforms the local approach in that it yields a collected anonymized data set with less information loss, that is, with higher utility.

Remark (co-utile anonymization). Note that the level of privacy protection obtained by a respondent affects the privacy protection that other respondents get. A basic approach for preserving the privacy of a specific respondent is based on hiding that respondent within a group of respondents. None of the respondents in such a group is interested making any of the respondents in the group re-identifiable, because that makes her own data more easily re-identifiable. For example, if one record in a k-anonymous group is re-identified, the probability of successful re-identification for the other group members increases from $1/k$ to $1/(k-1)$. This fact suggests that a respondent is interested not only in protecting her privacy, but also in helping other respondents in preserving theirs. This is the fundamental principle behind the notion of co-utility (see Section 2.3 above): the best strategy to attain one's goal is to help others in attaining theirs. The fact that privacy protection turns out to be co-utile ensures that respondents will be willing to collaborate with each other to improve the protection of all the group.

5 Collaborative *k*-Anonymity

This section describes how to generate a k-anonymous data set in a distributed manner, such that none of the respondents releases more information than the one available on her in the final k-anonymous data set. To this end, some communication between the respondents is needed to determine the k-anonymous groups.

In general, there can be several combinations of attributes in a data set that together act like a quasi-identifier, that is, such that each combination of attributes can be used to re-identify respondents; for example, one might have a quasi-identifier *(Age, Gender, Birthplace)* and another quasi-identifer *(Instruction_level, City_of_residence, Nationality)*. Without loss of generality and for the sake of simplicity, we will assume there is a single quasi-identifier that contains all the attributes that can potentially be used in record re-identification. Note that this is the worst-case scenario. Let QI be the set of attributes in this quasi-identifier.

Quasi-identifier attributes are usually assumed to contain no confidential information, that is, the set of quasi-identifier attributes is assumed to be disjoint from the set of confidential/sensitive attributes. This assumption is reasonable, because it is equivalent to saying that the attacker's background information does not include sensitive information on any respondent (indeed, the attacker wants *to learn* sensitive information, so it is reasonable to assume that he does not yet know it). Certainly, there might be special cases in which the attacker knows and uses sensitive data for re-identification, but we will stick to the usual setting in which this does not happen.

Since the attributes in QI are non-confidential, respondents can share their values among themselves and with the data collector, so that all of them get the complete list of QI attribute values. Based on that list, the data collector or any respondent can generate the k-anonymous groups. We propose to delegate the generation of the k-anonymous groups to the data collector. There are two main reasons for this:

- *Utility.* The actual k-anonymous partition chosen may have an important impact over analyses that can be accurately performed on the k-anonymous data. The data collector is probably the one who knows best (even if often only partially) the intended use of the data and, thus, the one who can make the most appropriate partition in k-anonymous groups.
- *Performance.* Generating the k-anonymous groups is the most computationally intensive part of k-anonymity enforcement. Hence, by delegating this task to the data collector, respondents relieve themselves from this burden.

When respondents have some interest in using the anonymized data set, it is plausible to assume that any respondent will rationally collaborate to generate it. The level of protection that a respondent in a given k-anonymous group gets is dependent on the level of protection that the other respondents in the group get: as justified above in Section 4, k-anonymization is co-utile.

On the other side, the data collector may try to deviate from the algorithm. Because the generation of the k-anonymous partition has been delegated to the data collector, respondents must make sure before reporting confidential information that the partition computed and returned by the data collector satisfies the requirements of k-anonymity. That is, each respondent must check that her k-anonymous group comprises k or more respondents.

After verifying the partition returned by the data collector, the respondent uploads to the data collector the quasi-identifier attribute values of her k-anonymous group together with her confidential data. This communication must be done through an anonymous channel (e.g. Tor [2]) to prevent anyone (the data collector, an intruder or anyone else) from tracking the confidential data to any respondent.

The above described steps to collaboratively generate a k-anonymous data set are formalized in Protocol 1.

Protocol 1

1. *Let $R_1, \dots, R_m$ be the set of respondents. Let (qi_i, c_i) be the quasi-identifier and confidential attribute values of R_i, for $i = 1, \cdots, m$.*
2. *Each R_i uploads her qi_i to a central data store so that anyone can query for qi_i.*
3. *The data collector generates a k-anonymous partition $\{P_1, \dots, P_p\}$ and uploads it to the central data store.*
4. *Each R_i checks that her k-anonymous group P_{R_i} contains k or more of the original quasi-identifiers.*
   If that is not the case, R_i refuses to provide any confidential data and exits the protocol.
5. *Each R_i sends (P_{R_i}, c_i) to the data collector through an anonymous channel.*
6. *With the confidential data collected, the data collector generates the k-anonymous data set.*

Protocol 1 is compatible with any strategy to generate the k-anonymous partition. Possible strategies include:

- *Methods reducing the detail of the quasi-identifier attributes.* Options here are generalization and supression [14, 11, 12], or microaggregation [5]).
- *Methods breaking the connection between quasi-identifier attributes and confidential attributes.* Among these we have Anatomy [19] (that splits the data into two tables, one containing the original quasi-identifier values and the other the original confidential attribute values, with both tables being connected through a group identifier attribute) and probabilistic k-anonymity [16] (that seeks to break the relation between quasi-identifiers and confidential attributes by means of a within-group permutation).

In fact, since the data collector and the respondents all know the exact values of the quasi-identifiers and the confidential attributes in each k-anonymous group, each of them can generate the k-anonymous data that suits her best.

In essence, the proposed protocol offers the same privacy protection as local anonymization (confidential data are only provided by the respondents in an anonymized form) while maintaining the data utility of centralized k-anonymization. At the respondents' side, there are only some minor additional communication and integrity checking costs.

We illustrate the steps of Protocol 1 for the respondents listed in the leftmost table of Figure 1. In Step 2 each respondent uploads her quasi-identifiers. The uploaded data are shown in the center-left table of Figure 1. At Step 3 the data collector analyzes the data uploaded in Step 2 and generates the partition in k-anonymous groups; this partition is shown in the center-right table of Figure 1.

In Step 4 each respondent checks that her group contains k or more of the quasi-identifier values uploaded in Step 2. Since this condition holds for all respondents in the example of the figure, respondents proceed to Step 5. In Step 5 each respondent uploads, through an anonymous channel, the group identifier she has been assigned together with her value for the confidential/sensitive attribute. The result is shown in the rightmost table of Figure 1. Here the layout of the rightmost table can be misleading: although we list in the i-th row the salary of R_i for $i = 1, \cdots 4$, any permutation of the four salaries could be listed (all four salaries in the P_1 group are indistinguishable). A similar comment holds for rows 5-8, in which we could list any permutation of the salaries in the P_2 group. At this point, the data collector (and the respondents) can generate the k-anonymous data set using the method they like best using that they see all tables in Figure 1 except the leftmost one.

	QI		Sensitive
	Zip	Age	Salary
R_1	13053	28	35000
R_2	13068	29	30000
R_3	13068	21	20000
R_4	13053	23	27000
R_5	14853	50	40000
R_6	14853	55	43000
R_7	14850	47	48000
R_8	14850	49	45000

Step 2	
Zip	Age
13053	28
13068	29
13068	21
13053	23
14853	50
14853	55
14850	47
14850	49

Step 3		
	Zip	Age
P_1	13053	28
	13068	29
	13068	21
	13053	23
P_2	14853	50
	14853	55
	14850	47
	14850	49

Step 5	
	Salary
P_1	35000
P_1	30000
P_1	20000
P_1	27000
P_2	40000
P_2	43000
P_2	48000
P_2	45000

Fig. 1. Distributed collaborative k-anonymization. Step numbers refer to Protocol 1. Each row in the leftmost table is only seen by the corresponding respondent. The other three tables are entirely seen by all respondents and the data collector.

Distributed anonymization based on hiding in a group via manipulation of the quasi-identifiers has an important flaw. An attacker may try to simulate one or more respondents, in order to gain more insight into the k-anonymous groups. To thwart this kind of attack, we need to make sure that every respondent has a verified identity, possibly by having all respondents registered with some trusted authority. If that is not feasible, some mitigation measures can be put in place to make it more difficult for an attacker to adaptively fabricate quasi-identifier values similar to a target respondent in order to track her:

- One option is for the data store manager (maybe the data collector) to unlock the access to the quasi-identifiers list (of Step 2) only after every respondent has uploaded her quasi-identifiers. In this way, the attacker must generate his quasi-identifier values without knowing the quasi-identifier values of the other respondents. This option has the shortcoming that respondents need to trust the data store manager to perform the above access control.
- An alternative that does not require trust in any central entity is to have each respondent upload a commitment (in the cryptographic sense, [6]) to

her quasi-identifiers before any actual quasi-identifier is uploaded. In this way, each respondent can check that none of the uploaded values was forged to target a specific respondent.

In the following section, we explore distributed anonymization based on masking the confidential attributes, rather than on hiding in a group via quasi-identifier manipulation.

6 Collaborative Masking of Confidential Data

Although k-anonymity is a popular privacy model, it has some important limitations. First of all, attribute disclosure is possible, even without re-identification, if the variability of the confidential attribute(s) within a k-anonymous group is small. Also, k-anonymity assumes that confidential attributes are not used in re-identification (*i.e.* that no confidential attribute is also a quasi-identifier), but this may not be the case if the attacker knows some confidential data. Moreover, we mentioned in the previous section that in our distributed generation of the k-anonymous data set, an attacker might simulate respondents to gain more insight into the k-anonymous groups. To deal with these issues, this section takes a different approach to generate the anonymized data set: instead of hiding within a group of respondents, each respondent masks her confidential data.

In this section, we relax the assumption that the set of quasi-identifier attributes and the set of confidential attributes are disjoint. The only assumption we make is that releasing the marginal distribution of a confidential attribute is not disclosive. What needs to be masked in the relation between a confidential attribute and any other attribute. Thus, we consider a data set with attributes $(A, C_1, \ldots, C_d)$ where C_j are confidential attributes for $j = 1, \cdots, d$ and A groups all non-confidential attributes.

Since the marginal distribution of confidential attributes is not disclosive, respondents can share the contents of each confidential attribute among themselves and with the data collector, so that all of them get the complete list of values for each confidential attribute. In this way, each respondent can evaluate the sensitivity of her value for each confidential attribute by taking into account the values of the other respondents for that attribute. From this sensitivity evaluation, the respondent can make a more informed decision regarding the amount of masking she needs to use.

Thus, we assume that each respondent R_i makes a decision about the amount of masking required for her confidential data and reports to the data collector the tuple $(a_i, c'_{1i} \ldots, c'_{di})$, where a_i is the original value of the non-confidential attributes and c'_{1i} the masked value of confidential attribute C_i. The fact that each respondent freely and informedly decides on the amount of masking required for her confidential data is a strong privacy guarantee (the respondent can enforce the level of permutation she wishes with respect to the original values). In fact, even if the data collector or any other entity recommend a specific amount of masking, respondents are free to ignore this recommendation. For a rational respondent, the selected level of masking is based on both privacy and utility considerations.

The reported masked data can be directly used to generate the masked data set. Better yet, by applying reverse mapping, the original marginal distribution of each confidential attribute can be recovered. This reverse mapping can be performed by the data collector and also by each respondent (because all respondents know the marginal distribution of the original attributes).

The previous discussion is formalized in Protocol 2.

Protocol 2

1. *Let $R_1, \ldots, R_m$ be the set of respondents. Let $(a_j, c_{j1} \ldots, c_{jd})$ be the attribute values of R_i.*
2. *For each confidential attribute C_j, each respondent R_i uploads c_{ij} to a central data store through an anonymous channel.*
3. *For each confidential attribute C_j, each respondent R_i analyzes all attribute values and decides on the amount of masking required for c_{ij}. Let c'_{ij} be the masked value.*
4. *Each respondent R_i uploads $(a_i, c'_{i1} \ldots, c'_{id})$ to the data store.*
5. *The data collector applies reverse mapping to the data uploaded in Step 4 in order to obtain the final anonymized data set. (The same can be done by each respondent.)*

Although the reasons why Protocol 2 is safe have already been presented in the discussion prior to the algorithm formalization, a more systematic analysis is presented in the following proposition.

Proposition 1. *At the end of Protocol 2, nobody learns information about any respondent R_i that is more accurate than the masked data reported by R_i in Step 4.*

Proof. Apart from the release of the masked data in Step 4, the only step in which R_i releases data is Step 2. Since the data released in Step 2 are not anonymized, we need to make sure they cannot be linked back to R_i.

Because the uploads in Step 2 are performed through an anonymous channel, there is no way for an attacker to track the data transfers to any particular respondent. What is more, since each c_{ij} is separately uploaded through the anonymous channel, there is no way for the attacker to link to one another the values c_{ij}, $j = 1, \cdots, d$ corresponding to the same respondent R_i (if the attacker could link such values, he could reconstruct the original record of R_i).

Finally, since by assumption releasing the marginal distribution of each confidential attribute is not disclosive, there is no risk in uploading each c_{ij} in Step 2. The reason is that each c_{ij} carries less information than the marginal distribution of attribute C_j. (The release of a c_{ij} could be problematic if C_j contains confidential information and, at the same time, can be used in re-identification, but assuming that the marginals are not disclosive rules out this situation). □

We illustrate the steps of Algorithm 2 for the respondents listed in the leftmost table of of Figure 2. We assume that Age and Salary are the confidential attributes. In Step 2 each respondent uploads to the central data store each of her values for the confidential attributes. Each respondent performs a separate upload through an anonymous channel for each of the confidential attributes. At the end of Step 2, the marginal distribution of the confidential attributes is available to the data collector and all respondents in the central data store, as illustrated in the center-left table of Figure 2. In Step 3 each respondent can analyze the marginal distributions and decide on the amount of masking required for each confidential attribute. In this example Age is masked by adding a random value between -5 and 5, and Salary is masked by adding a random value between -5000 and 5000. Of course, each respondent could have applied a different masking. In Step 4 each respondent uploads the masked confidential attributes together with the rest of attributes (the non-confidential ones). This upload need not be done through an anonymous channel, because all confidential data are masked. The data set uploaded by respondents to the central data store at the end of Step 4 is shown in the center-right table of Figure 2. In the final step, the data collector applies reverse mapping to each confidential attribute to recover the original marginal distributions, as illustrated in the rightmost table of Figure 2.

		Sensitive	
	Zip	Age	Salary
R_1	13053	28	35000
R_2	13068	29	30000
R_3	13068	21	20000
R_4	13053	23	27000
R_5	14853	50	40000
R_6	14853	55	43000
R_7	14850	47	48000
R_8	14850	49	45000

Step 2	
Age	Salary
28	35000
29	30000
21	20000
23	27000
50	40000
55	43000
47	48000
49	45000

Step 4		
Zip	Age	Salary
13053	29	37306
13068	24	27765
13068	18	18951
13053	19	28151
14853	51	36879
14853	50	42631
14850	52	45585
14850	49	40390

Step 5		
Zip	Age	Salary
13053	29	40000
13068	28	27000
13068	21	20000
13053	23	30000
14853	50	35000
14853	49	45000
14850	55	48000
14850	47	43000

Fig. 2. Distributed collaborative masking of the confidential attributes. Step numbers refer to Protocol 2. Each row in the leftmost table is only seen by the correspoding respondent. The other three tables are entirely seen by all respondents and the data collector.

7 Conclusions and Future Research

We have sketched two protocols for collaborative microdata anonymization. The first one assumes a clear separation between confidential attributes and quasi-identifiers, and seeks to attain k-anonymity. In the second one, no separation between quasi-identifiers and confidential attributes is assumed, and the goal is to sufficiently mask the confidential attributes.

Compared to local anonymization, collaborative anonymization incurs less information loss and achieves the same privacy vs the data collector. Compared to centralized anonymization, collaborative anonymization requires less trust in the data collector and achieves the same data utility. Therefore, collaborative anonymization should be preferred by rational respondents to both local and centralized anonymization.

In a survey, the motivations for respondents to report data and report them truthfully to the data collector are in general unclear. As a rule, a rational respondent is willing to participate only if the benefit she obtains is greater than the potential harm due to privacy loss. If respondents are interested in the collected data set and they wish it to be as accurate as possible, then collaborative anonymization protocols are co-utile.

Future work will be devoted to develope collaborative anonymization protocols for a broader range of privacy models (beyond k-anonymity) and disclosure limitation techniques.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: Templeton World Charity Foundation (grant TWCF0095/AB60 "CO-UTILITY"), Government of Catalonia (ICREA Acadèmia Prize to the second author and grant 2014 SGR 537), Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY"), European Commission (projects FP7 "DwB", FP7 "Inter-Trust" and H2020 "CLARUS"). The second author leads the UNESCO Chair in Data Privacy. The views in this paper are the authors' own and do not necessarily reflect the views of the Templeton World Charity Foundation or UNESCO.

References

1. S. Agrawal and J.R. Haritsa. A framework for high-accuracy privacy-preserving mining. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, pp. 193–204, 2005.
2. R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, SSYM'04, pp. 21–21, Berkeley, CA, USA, 2004. USENIX.
3. J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *CoRR*, abs/1501.04186, 2015.
4. J. Domingo-Ferrer, J. Soria-Comas, and Ciobotaru O. Co-utility: self-enforcing protocols without coordination mechanisms. In *Proceeding of the 5th International Conference on Industrial Engineering and Operations Management (IEOM 2015)*, 2015.
5. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min. Knowl. Discov.*, 11(2):195–212, 2005.

6. O. Goldreich. *Foundations of Cryptography: Vol. 1, Basic Tools*. Cambridge University Press, 2001.
7. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.
8. W. Jiang and C. Clifton. Privacy-preserving distributed k-anonymity. In *Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, DBSec'05, pp. 166–177, Berlin, 2005. Springer.
9. W. Jiang and C. Clifton. A secure distributed framework for achieving k-anonymity. *The VLDB Journal*, 15(4):316–333, 2006.
10. P. Jurczyk and L. Xiong. Distributed anonymization: achieving privacy for both data subjects and data providers. In *Proceedings of the 23rd Annual IFIP WG 11.3 Working Conference on Data and Applications Security XXIII*, pp. 191–207, Berlin, 2009. Springer.
11. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pp. 49–60, New York, NY, USA, 2005. ACM.
12. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22Nd International Conference on Data Engineering*, ICDE '06, Washington, DC, USA, 2006. IEEE Computer Society.
13. K. Muralidhar, R. Sarathy, and J. Domingo-Ferrer. Reverse mapping to preserve the marginal distributions of attributes in masked microdata. In Josep Domingo-Ferrer, editor, *Privacy in Statistical Databases*, LNCS 8744, pp. 105–116. Springer, 2014.
14. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proc. of the IEEE Symposium on Research in Security and Privacy*, 1998.
15. C. Song and T. Ge. Aroma: a new data protection method with differential privacy and accurate query answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 1569–1578, New York, NY, USA, 2014. ACM.
16. J. Soria-Comas and J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012)*, pp. 1–8. IEEE, 2012.
17. K. Wang, B. C. M. Fung, and G. Dong. Integrating private databases for data analysis. In *Proc. of IEEE Intl. Conf. on Intelligence and Security Informatics, ISI 2005*, Atlanta GA, pp. 171–182, 2005.
18. S. L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63+, 1965.
19. X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In *Proc. of the 32nd Intl. Conf. on Very Large Data Bases (VLDB 2006)*, pp. 139–150. 2006.
20. X. Xiao and Y. Tao. Personalized privacy preservation. In *Proc. of the 2006 ACM SIGMOD Intl. Conf. on Management of Data, SIGMOD '06*, pp. 229–240, New York, NY, USA, 2006.