A General Proximity Privacy Principle

Ting Wang, Shicong Meng, Bhuvan Bamba, Ling Liu, Calton Pu

College of Computing, Georgia Institute of Technology

{twang, smeng, bhuvan, lingliu, calton}@cc.gatech.edu

I. Introduction

Recent years have witnessed ever-increasing concerns about individual privacy in numerous data dissemination applications that involve private personal information, e.g., medical data or census data. Typically, such *microdata* is stored in a relational table, each record corresponding to an individual, which can be divided into three sub-categories: (1) *identifier* attribute, e.g., social security number, which can explicitly identify an individual, and is usually removed from the microdata for publication; (2) *quasi-identifier* (QI) attributes, e.g., age, zip-code, and birth-date, whose values in combination can potentially identify an individual, and are usually available from other sources (e.g., voter registration list); (3) *sensitive* (SA) attribute, e.g., disease, which is the private information to be protected for the individuals.

To address the privacy concerns, a plethora of work has been done on anonymized data publication [3], [4], [8], [9], [10], [11], [12], [13], [14], [16], [17], [18], [19], [20], [21], [22], [23], aiming at ensuring that no adversary can accurately infer the SA-value of an individual, based on the published data and her background knowledge. In particular, a majority of the efforts focus on addressing the *linking attacks*: the adversary possesses the exact QI-values of the victim, and attempts to discover his/her SA-value from the published data. A popular methodology of thwarting such attacks is *generalization* [18]: after partitioning the microdata into a set of disjoint subsets of tuples, called *QI-group*, generalization transforms the QI-values in each group to a uniform format, so that all the tuples belonging to the same QI-group G are indistinguishable in terms of QI-values.

Example 1. Consider the example of publishing uncertainty sensitive data as shown in Table I: `Age` and `Zip` are QI-attributes, and `Disease` is an uncertainty sensitive attribute, which follows the *x-relation* probabilistic model [1]: each `Disease` value is a discrete probability distribution over a set of alternative diseases, indicating the possibility of the individual's suffering of each specific disease.

The generalization over the microdata produces two QI-groups, as indicated by their group IDs (GID), and transforms the QI-values in each group to a unified format. The adversary who knows *Kevin*'s QI-values can no longer uniquely determine his SA-value: each `Disease` value in the first group may belong to him, therefore without further information, the adversary can associate *Kevin* with each specific `Disease` value with probability only 20%.

Essentially, generalization protects against linking attacks by weakening the association between QI-values and SA-values. The protection is sufficient if the weakened associations are not informative enough for the adversary to infer individuals' SA-values with high confidence. Aiming at providing adequate protection, a number of anonymization principles have been proposed, including (i) k-anonymity [18], l-diversity [13] and its variants [19], [22], and (c, k)-safety [14] for publishing categorical sensitive data, and (ii) (k, e)-anonymity [23], variance control [10], t-closeness [11] and (ϵ, m)-anonymity [12] for publishing one-dimensional numeric sensitive data.

		Age	Zip	Disease				GID
				flu	asthma	bronchitis	none	
Kevin	1	**18-30**	**12-17k**	**0.5**	**0.3**	**0.1**	**0.1**	1
	2	18-30	12-17k	0.4	0.3	0.2	0.1	1
	3	18-30	12-17k	0.4	0.2	0.2	0.2	1
	4	18-30	12-17k	0.3	0.4	0.2	0.1	1
	5	18-30	12-17k	0.2	0.7	0.1	0	1
	6	32-40	22-30k	0.2	0.6	0.2	0	2
	7	32-40	22-30k	0.8	0.1	0	0.1	2
	8	32-40	22-30k	0.3	0.1	0.5	0.1	2

TABLE I

ILLUSTRATION OF PRIVACY-PRESERVING PUBLICATION.

However, designed with the assumption of specific data types, these principles and their associated anonymization algorithms fail to address the privacy risks for a much wider range of data models, where the proximity of sensitive values is defined by arbitrarily complicated or even customized defined functions, as illustrated in the following example.

Example 2. Recall the example in Table I. If one measures the pair-wise semantic proximity[1] of the `Disease` values in the first QI-group, it is noticed that the first four tuples form a tight "neighborhood" structure, where the value of #2 is semantically close to that of #1, #3 and #4, as shown in Fig. 1.

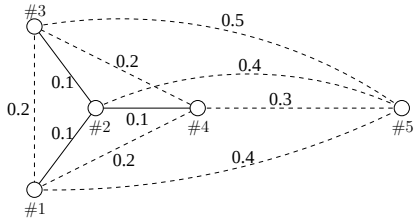

Fig. 1. Illustration of the proximity breach in the generalized data. Dashed lines indicate that the pairwise distances could not be embedded in a two-dimensional space.

Clearly, assuming that each tuple in the group belongs to

[1]Here we use *variational distance* as the distance metric. For two discrete distributions $\mathbf{P} = (p_1, \ldots, p_m)$ and $\mathbf{Q} = (q_1, \ldots, q_m)$, their distance is defined as $\mathsf{D}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|$.

Kevin with identical probability, the adversary can infer that *Kevin*'s `Disease` value falls in the neighborhood structure with probability 80%. Furthermore, by picking the `Disease` value of the center node as the representative, she arrives at a privacy intruding claim that "*Kevin*'s `Disease` value is fairly close to $[0.4, 0.3, 0.2, 0.1]$".

The example above illustrates the general proximity breaches essentially existing in most data models, given a semantic proximity metric is defined over the domain of the sensitive attribute. In this paper, we aim at developing effective privacy principle to tackle such general proximity breaches.

Concretely, we propose a novel principle (ϵ, δ)-*dissimilarity*. It intuitively requires that in each QI-group G, every SA-value, is "dissimilar" to at least $\delta \cdot (|G| - 1)$ other ones, where $|G|$ denotes the cardinality of G and two SA-values are considered "dissimilar" if their semantic distance is above ϵ. We provide theoretical proof that (ϵ, δ)-dissimilarity, used in conjunction with k-anonymity [18], provides effective protection against linking attacks. We analytically discuss the satisfiability problem of both (ϵ, δ)-dissimilarity and k-anonymity. Finally, we point to promising solutions to fulfilling these two principles.

II. Problem Formalization

In this section, we introduce the fundamental concepts, and formalize the problem of *general proximity privacy*.

A. Models and Assumptions

Let T be a microdata table to be published, which contains a set of QI attributes, and a sensitive attribute A^s. We make the following assumptions regarding these attributes:

- All QI-attributes are either categorical or numeric, i.e., an ordering can be juxtaposed;
- A^s can be of arbitrary data type, e.g., categorical, quantitative, customized defined type, etc;
- A distance metric $\Delta(\cdot, \cdot)$ is defined over the domain of A^s, and $\Delta(x, y)$ denotes the semantic distance between two SA-values x and y.

The generalization operation partitions T into a set of QI-groups $\mathcal{G}_T = \{G_1, \ldots, G_m\}$, which satisfies $G_1 \cup \ldots \cup G_m = T$, and $G_i \cap G_j = \emptyset$ for $i \neq j$, i.e., $\mathcal{G}_T$ is a disjoint and complete partition of T. The QI-values in each group G are then transformed to a uniform format. One possible strategy could be: if the QI-attribute A^q is quantitative, the generalized value could be the minimum bounding interval of all the A^q values in G; if A^q is categorical, it could be the lowest common ancestor (LCA) of all the A^q values in G on the domain generalization hierarchy of A^q.

Given the generalized table, the adversary attempts to exploit it to infer the SA-value $o.A^s$ of an individual o. We assume that the adversary possesses full identification information [14], which includes (1) the identifier of o, (2) the exact QI-values of o, and (3) the QI-group G in the generalized table that contains o. Note that by assuming the background knowledge (3), we are dealing with the worst-case scenario that there is only one QI-group with QI-values matching o.

After identifying the QI-group G that contains o, the adversary attempts to estimate $o.A^s$ with a probabilistic approach. We assume that from the perspective of the adversary, every tuple in G belongs to o with identical possibility, therefore $o.A^s$ is a random variable with a discrete distribution over the SA-values appearing in G. Let X denote this random variable, which has the following probability mass function:

$$\mathsf{prob}[X = v] = \frac{\mathsf{num}_G(v)}{|G|} \tag{1}$$

where $\mathsf{num}_G(v)$ is the number of tuples with SA-value as v in G, and $|G|$ is the cardinality of G.

Example 3. In our running example of TABLE I, the adversary identifies that *Kevin* is associated with each specific `Disease` value in the first QI-group with equal probability 20%.

B. General Proximity Privacy

It is noticed that in Equation 1, if for every pair of SA-values v_i, v_j in G, $\mathsf{num}_G(v_i) = \mathsf{num}_G(v_j)$, then all the SA-values in G are indistinguishable in terms of probability, which provides sufficient privacy protection in terms of l-diversity [13], if G contains more than l different SA-values.

Now we introduce the semantic proximity into our privacy concern. For a QI-group G with SA-values as a multi-set $\mathcal{SV}_G = \{v_1, v_2, \ldots, v_n\}$, and the semantic distance metric $\Delta(\cdot, \cdot)$ over the sensitive attribute, the "*ϵ-neighborhood*" of a value $v \in \mathcal{SV}_G$, $\Phi_G(v, \epsilon)$ is defined as the subset of $\mathcal{SV}_G$ with their semantic distance to v at most ϵ, formally

$$\Phi_G(v, \epsilon) = \{v' |\ v' \in \mathcal{SV}_G \text{ and } \Delta(v, v') \leq \epsilon\}$$

Example 4. In the running example shown in Fig. 1, given ϵ = 0.1, the ϵ-neighborhood of v_2 consists of $\{v_1, v_2, v_3, v_4\}$.

The probability that a tuple has SA-value belonging to the private neighborhood of v is defined as

$$\mathsf{prob}[X \in \Phi_G(v, \epsilon)] = \frac{|\Phi_G(v, \epsilon)|}{|G|} \tag{2}$$

where $|\Phi_G(v, \epsilon)|$ denotes the cardinality of the ϵ-neighborhood of v. Specifically, within this framework, the definition in Equation 1 can be re-formalized as

$$\mathsf{prob}[X = v] = \frac{|\Phi_G(v, 0)|}{|G|}$$

Clearly, for any v and $\epsilon > 0$, $\mathsf{prob}[X = v] \leq \mathsf{prob}[X \in \Phi_G(v, \epsilon)]$.

It is noticed that given a reasonable ϵ, if the ϵ-neighborhood of v, $\Phi_G(v, \epsilon)$ contains a considerable portion of the SA-values in G, the adversary can conclude that the victim individual o is associated with the SA-values appearing in $\Phi_G(v, \epsilon)$ with high probability, though she may not be sure about the exact value. Furthermore, using v as the representative value of $\Phi_G(v, \epsilon)$, she can obtain fairly precise estimation about $o.A^s$, the SA-value of o.

To measure the severeness of the breaches, and particularly, to capture the effect of proximate SA-values on enhancing the

estimation of the adversary, we introduce the following risk metric: given the neighborhood width ϵ and a QI-group G, its risk of general proximity breach, $\mathsf{risk}(G, \epsilon)$ is defined as:

$$\mathsf{risk}(G, \epsilon) = \max_{v \in \mathcal{SV}_G} \frac{|\Phi_G(v, \epsilon)| - 1}{|G| - 1} \quad (3)$$

Intuitively, $\mathsf{risk}(G, \epsilon)$ measures the relative size of the largest ϵ-neighborhood in G, and by excluding one from the neighborhood, it highlights the impact of proximate SA-values on improving the belief of the adversary, who priorly associates the victim with each SA-value[2] with probability $1/|G|$.

It is noted that G is free of proximity breach ($\mathsf{risk}(G, \epsilon) = 0$) if all the SA-values are dissimilar, and reaches its maximum ($\mathsf{risk}(G, \epsilon) = 1$) if a specific SA-value v is similar to all other SA-values. In particular, we define that $\frac{|\Phi_G(v,\epsilon)|-1}{|G|-1} = 1$, for the extreme case that $\mathcal{SV}_G = \{v\}$.

Further, we define the risk of general proximity breach for a partition of the microdata table T, $\mathcal{G}_T$, $\mathsf{risk}(\mathcal{G}_T, \epsilon)$ as the maximum risk of the QI-groups in $\mathcal{G}_T$, formally

$$\mathsf{risk}(\mathcal{G}_T, \epsilon) = \max_{G \in \mathcal{G}_T} \mathsf{risk}(G, \epsilon) \quad (4)$$

III. A General Principle

In this section, we present (ϵ, δ)-dissimilarity, a remedy against general proximity breaches, with theoretical proof of effectiveness against linking attacks. We further discuss the relevance of (ϵ, δ)-dissimilarity to other generalization principles in literatures.

A. (ϵ, δ)-Dissimilarity

To remedy the general proximity breaches, we propose a novel privacy principle, *(ϵ, δ)-dissimilarity*. A partition $\mathcal{G}_T$ satisfies (ϵ, δ)-dissimilarity if in every $G \in \mathcal{G}_T$, every SA-value v in G is "dissimilar" to at least $\delta \cdot (|G| - 1)$ other SA-values, while two SA-values are considered dissimilar if their semantic distance is above ϵ. Note that δ essentially controls the risk of the possible proximity breaches.

Now we prove the effectiveness of this principle against general proximity attacks. Concretely we show that a partition $\mathcal{G}_T$ is free of general proximity breaches, if and only if it satisfies (ϵ, δ)-dissimilarity. We have the following theorem:

Theorem 1. *Given a microdata table T and ϵ, for a partition $\mathcal{G}_T$, $\mathsf{risk}(\mathcal{G}_T, \epsilon) \leq 1 - \delta$, if and only if $\mathcal{G}_T$ satisfies (ϵ, δ)-dissimilarity.*

Proof: ($\Rightarrow$) If the partition $\mathcal{G}_T$ violates (ϵ, δ)-dissimilarity, i.e., $\exists G \in \mathcal{G}_T$, $\exists v \in \mathcal{SV}_G$, $|\Phi_G(v, \epsilon)| - 1 > (1 - \delta) \cdot (|G| - 1)$, then trivially $\mathsf{risk}(\mathcal{G}_T, \epsilon) \geq \frac{|\Phi_G(v,\epsilon)|-1}{|G|-1} > 1 - \delta$, which indicates a proximity breach.

($\Leftarrow$) If $\mathcal{G}^T$ contains a proximity breach with risk at least $1 - \delta$, then there must exist certain $G \in \mathcal{G}_T$, and certain $v \in \mathcal{SV}_G$, which violates (ϵ, δ)-dissimilarity.

Essentially, (ϵ, δ)-dissimilarity counters general proximity attacks by specifying constraints on the number of ϵ-neighbors that each SA-value can have, relative to the QI-group size. It captures the impact of proximate SA-values on improving the estimation of the adversary, who has prior belief $1/|G|$ for each SA-value in G.

However, it does not prevent the trivial case of small size QI-group with pair-wise dissimilar SA-values. To remedy this, we introduce k-anonymity [18] into our framework: by requiring that every QI-group contains at least k tuples, the prior belief for each SA-value is at most $1/k$.

Clearly, by applying (ϵ, δ)-dissimilarity, in conjunction of k-anonymity, one can effectively counter linking attacks in terms of both exact and proximate QI-SA associations. We term the combination of (ϵ, δ)-dissimilarity and k-anonymity as $(\epsilon, \delta)^k$-dissimilarity.

B. Relevance to Previous Principles

$(\epsilon, \delta)^k$-dissimilarity makes no specific assumption about the underlying data models, hence is general enough to tackle the proximity breaches within most existing data models. Following we show that most anonymity principles on proximity breaches in literatures are either in-adequate against general proximity breaches, or essentially the special cases of $(\epsilon, \delta)^k$-dissimilarity within the data models which they are designed for.

1) Principles for categorical data: Motivated by the homogeneity breaches, l-diversity [13] and its variants, (α, k)-anonymity [19], m-invariance [22] are designed to ensure sufficient diversity of the SA-values in every QI-group, and are all essentially special forms of $(\epsilon, \delta)^k$-dissimilarity for data models where different values have no sense of semantic proximity, e.g., categorical data.

Take (α, k)-anonymity [19] as an example, which essentially combines k-anonymity and l-diversity, and demands that (1) every QI-group must contain at least k tuples, (2) at most α-percent of the tuples carry an identical SA-value. Trivially, it is equivalent to $(\epsilon, \delta)^k$-dissimilarity in the sense that $\epsilon = 0$ and $1 - \delta \approx \alpha$.

2) Principles for numeric data: For data models within which different values can be strictly ordered, e.g., numeric data, it qualifies as several threats if the adversary can identify the victim individual's SA-value within a short interval, even not the exact value. Attempting to capture such privacy requirement, a set of privacy principles have been proposed for publishing numerical sensitive data, e.g., variance control [10], (k, e)-anonymity [23], t-closeness [11]. However, it is proved in [12] that all these principles provide insufficient protection against proximity attacks.

The principle most relevant to $(\epsilon, \delta)^k$-dissimilarity is (ϵ, m)-anonymity [12], which requires that in a given QI-group G, for each SA-value x, at most $1/m$ of the tuples of G fall in the interval of $[x - \epsilon, x + \epsilon]$. Clearly, (ϵ, m)-anonymity is a special form of $(\epsilon, \delta)^k$-dissimilarity for numeric sensitive data, with $1/m \approx 1 - \delta$. However, targeting one-dimensional numeric data, the theoretical analysis and the corresponding generalization algorithms in [12] are not applicable for general proximity privacy protection. Moreover, since m is an integer,

[2]Note that here we consider the collection of SA-values in a group G, $\mathcal{SV}_G$ as a multi-set, and regard all SA-values as unique.

the users can only specify their proximity privacy requirement in a harmonic sequence manner, i.e., $\frac{1}{2}, \frac{1}{3}, \ldots$, instead of "stepless" continuous adjustment as provided by $(\epsilon, \delta)^k$-dissimilarity.

IV. Fulfillment of $(\epsilon, \delta)^k$-Dissimilarity

In this section, we present a theoretical study of the satisfiability of $(\epsilon, \delta)^k$-dissimilarity. Specifically, and point to promising approximate solutions to fulfilling this principle.

A. Satisfiability of $(\epsilon, \delta)^k$-Dissimilarity

Given the microdata table T, and the privacy requirements as specified by k (k-anonymity), ϵ and δ ((ϵ, δ)-dissimilarity), the first question arises as "does there exist a partition $\mathcal{G}_T$ for T that satisfies both k-anonymity (ϵ, δ)-dissimilarity?", i.e., the *satisfiability* of $(\epsilon, \delta)^k$-dissimilarity for T. Unfortunately, in general, no efficient solution exists to answer this question, as shown in the next theorem.

Theorem 2. *In general, the $(\epsilon, \delta)^k$-dissimilarity satisfiability problem is NP-hard.*

Proof: It suffices to prove that this problem for a specific setting is NP-hard. Consider a stringent version of $(\epsilon, \delta)^k$-dissimilarity: δ = 1, i.e., in every QI-group, all the SA-values are required to be dissimilar.

For given parameter ϵ and microdata T (of cardinality n), we construct an abstract graph $\Psi_T(\epsilon) = (\mathcal{V}_T(\epsilon), \mathcal{E}_T(\epsilon))$: $\mathcal{V}_T(\epsilon)$ is the set of vertices, with each vertex v representing a SA-value in T; $\mathcal{E}_T(\epsilon)$ represents the set of edges over $\mathcal{V}_T(\epsilon)$, and two vertices v and v' are adjacent, if and only if their corresponding SA-values are similar.

Without loss of generality, consider a partition $\mathcal{G}_T$ of T consisting of m ($m \leq \lfloor n/k \rfloor$) QI-groups: $\mathcal{G}_T = \{G_1, G_2, \ldots, G_m\}$, and the vertices in $\Psi_T(\epsilon)$ corresponding to each G_i are labeled with one distinct color. Clearly, in this setting, $\mathcal{G}_T$ satisfies $(\epsilon, \delta)^k$-dissimilarity, only if no two adjacent vertices in $\Psi_T(\epsilon)$ share identical color, called a proper coloring.

However, it is known that for a general graph, determining if a proper coloring using at most m colors exists is NP-complete [5], which implies that the $(\epsilon, \delta)^k$-dissimilarity satisfiability problem is NP-hard.

B. Possible Solution

Instead of attempting to seek an exact answer to whether a partition exists for given privacy requirements, one is more interested in developing approximate algorithms that can (1) provide explicit and intuitive guidance for the setting of privacy parameters, and (2) efficiently find high-quality approximate solution to partitioning the microdata.

One such approximate solution could be: one first reformulates the problem of finding an $(\epsilon, \delta)^k$-dissimilarity satisfied partition in the framework of graph coloring, and projects it to certain relaxed coloring problem, e.g., [2], [7], which can desirably embeds the privacy parameters (ϵ, δ and k). One can then study sufficient and necessary conditions for the existence of a proper coloring. A constructive proof of the conditions can naturally lead to an efficient solution.

V. Conclusion

This work presents a systematic study of the problem of protecting general proximity privacy, with findings applicable to most existing data models. Our ontributions are multi-folded: we highlighted and formulated proximity privacy breaches in a data-model-neutral manner; we proposed a new privacy principle $(\epsilon, \delta)^k$-dissimilarity, with theoretically guaranteed protection against linking attacks in terms of both exact and proximate QI-SA associations; we provided a theoretical analysis regarding the satisfiability of $(\epsilon, \delta)^k$-dissimilarity, and pointed to promising solutions to fulfilling this principle.

Acknowledgement

This work is partially sponsored by grants from NSF CyberTrust, a grant from AFOSR, and an IBM SUR grant.

References

[1] P. Agrawal, O. Benjelloun, A. Das Sarma, C. Hayworth, S. Nabar, T. Sugihara and J. Widom. "Trio: A system for data, uncertainty, and lineage". In *VLDB*, 2006.

[2] L. Cowen, W. Goddard and C. Jesurum. "Coloring with defect". In *SODA*, 1997.

[3] B. Chen, R. Ramakrishnan and K. LeFevre. "Privacy skyline: privacy with multidimensional adversial knowledge". In *VLDB*, 2007.

[4] B. Fung, K. Wang and P. Yu. "Top-down specialization for informaiton and privacy preservation". In *ICDE*, 2005.

[5] M. Garey and D. Johnson. "Computers and intractability: a guide to the theory of NP-completeness". Freeman, San Francisco, CA, 1981.

[6] D. Kifer and J. Gehrke. "Injecting utility into anonymization databases". In *SIGMOD*, 2006.

[7] L. Lovász. "On decompositions of graphs". *Studia Sci. Math. Hungar.*, 1:237-238, 1966.

[8] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Incognito: efficient full-domain k-anonymity". In *SIGMOD*, 2005.

[9] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Mondrian multidimensional k-anonymity". In *ICDE*, 2006.

[10] K. LeFevre, D. DeWitt and R. Ramakrishnan. "Workload-aware anonymization". In *SIGKDD*, 2006.

[11] N. Li, T. Li and S. Venkatasubramanian. "t-closeness: privacy beyond k-anonymity and l-diversity". In *ICDE*, 2007.

[12] J. Li, Y. Tao and X. Xiao. "Preservation of proximity privacy in publishing numerical sensitive data". In *SIGMOD*, 2008.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam. "l-diversity: privacy beyond k-anonymity". In *ICDE*, 2006.

[14] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke and J. Halpern. "Worst-case background knowledge in privacy". In *ICDE*, 2007.

[15] A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity". In *PODS*, 2004.

[16] M. Nergiz, M. Atzori and C. Clifton. "Hiding the presence of individuals from shared databases". In *SIGMOD*, 2007.

[17] H. Park and K. Shim. "Approximate algorithm for k-anonymity". In *SIGMOD*, 2007.

[18] L. Sweeney. "k-anonymity: a model for protecting privacy". In *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5), 2002.

[19] R. Wong, J. Li, A. Fu and K. Wang. "(alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing". In *SIGKDD*, 2006.

[20] K. Wang, P. Yu and S. Chakraborty. "Bottom-up generalization: a data mining solution to privacy protection". In *ICDM*, 2004.

[21] X. Xiao and Y. Tao. "Anatomy: Simple and effective privacy preservation". In *VLDB*, 2006.

[22] X. Xiao and Y. Tao. "m-invariance: towards privacy preserving re-publication of dynamic datasets". In *SIGMOD*, 2007.

[23] Q. Zhang, N. Koudas, D. Srivastava and T. Yu. "Aggregate query answering on anonymized tables". In *ICDE*, 2007.