Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Connecting privacy models: synergies between *k*-anonymity, *t*-closeness and differential privacy

Prepared by Jordi Soria-Comas and Josep Domingo-Ferrer, Universitat Rovira i Virgili, Catalonia, Spain

Connecting privacy models: synergies between k-anonymity, t-closeness and differential privacy

Jordi Soria-Comas, Josep Domingo-Ferrer

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

Abstract. The usual approach to generate k-anonymous data sets, based on generalization of the quasi-identifier attributes, does not provide any control on the variability of the confidential attributes within the k-anonymous groups. If the latter variability is too small, privacy is not sufficiently protected, while, for large variabilities, data utility is substantially damaged. Some refinements to the basic k-anonymity privacy model, like l-diversity and t-closeness, seek to prevent the variability of the confidential attributes within a k-anonymous group from being too small. However, upper-bounding the variability of the confidential attributes to improve utility has not yet been considered. We propose a method to attain k-anonymity, based on microaggregation of the confidential data, that seeks the lowest possible variability for the confidential attributes, thereby maximizing utility. Our proposal can be combined with k-anonymity refinements such as l-diversity and t-closeness, hence yielding simultaneous utility and privacy guarantees.

ε-Differential privacy is another popular privacy model that is often opposed to k-anonymity like models. k-Anonymity is usually presented as a model that preserves data utility to a good extent but offers only limited privacy guarantees. In contrast, ε-differential privacy provides strong privacy guarantees but only limited data utility. We show that for microdata releases, ε-differential privacy can be seen as a kind of t-closeness with a specific distance measure. Hence, our proposal to minimize the variability of the confidential attributes can also be applied for ε-differential privacy.

1 Introduction

The usual approach to generate k-anonymous data sets, based on generalization of the quasi-identifiers, does not guarantee any pre-specified variability level for the confidential attributes within the k-anonymous groups. If the latter variability is too small, privacy is not sufficiently protected, while, for large variabilities, data utility is widely damaged. Some improvements of k-anonymity exist (such as l-diversity and t-closeness) that ensure lower bounds on the variability of the confidential attributes, thereby providing real privacy protection. However, upper-bounding the variability of the confidential data to improve utility has not been considered. We propose a method to attain k-anonymity, based on microaggregation of the confidential data, that seeks the lowest possible variability for the confidential attributes, hence improving the utility of the anonymized data. Our proposal can be combined with

l-diversity and t-closeness, in such a way to offer simultaneous utility and privacy guarantees.

k-Anonymity and ε-differential privacy are two mainstream privacy models originated within the computer science community. Their approaches towards disclosure limitation are quite different: k-anonymity is a model for releases of microdata (*i.e.* individual records) that seeks to prevent record re-identification by hiding each original record within a group of k indistinguishable anonymized records, while ε-differential privacy originated as a model for interactive databases and seeks to limit the knowledge that users obtain from query responses. Both models are often presented as antagonistic: ε-differential privacy supporters view k-anonymity as an old-fashioned privacy notion that offers only poor disclosure limitation guarantees, while ε-differential privacy detractors criticize the limited utility of ε-differentially private outputs and the cumbersomeness of not having access to the data set.

1.1 Contribution and plan of this paper

We show in this paper that k-anonymity like privacy models and differential privacy are not only complementary (as pointed out in Clifton and Tassa, 2013), but they are intimately related. Specifically, the t-closeness extension to k-anonymity turns out to be closely related to ε-differential privacy.

Section 2 reviews background on k-anonymity, t-closeness and ε-differential privacy. Section 3 reviews partitioning strategies used to achieve k-anonymity and its extensions, including t-closeness. Section 4 shows that $\exp(\varepsilon)$-closeness, when reached using a specific distance, behaves as ε-differential privacy versus uninformed intruders. Section 5 concludes the paper.

2 Background

2.1 k-Anonymity

k-anonymity (Samarati and Sweeney, 1998; Samarati, 2001) seeks to prevent identity disclosure in microdata releases based on the quasi-identifier attributes. To this end, k-anonymity requires each combination of quasi-identifier values to be shared by (at least) k records in the released data set.

Definition 1 (k-anonymity) *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.*

The most usual procedure to attain k-anonymity is based on generalization of the quasi-identifier attributes (Samarati and Sweeney, 1998) and supression of records that would lead to increased generalization. Yet alternative approaches exist, such as Domingo-Ferrer and Torra (2005), which is based on microaggregation of the quasi-identifier attributes.

2.2 l-Diversity and t-closeness

While k-anonymity protects against identity disclosure, as mentioned above, it does not protect in general against attribute disclosure: if the values for one (or several) confidential attribute(s) are identical within a group of records sharing the quasi-identifier attribute values, disclosure happens.

The property of l-diversity (Machanavajjhala *et al.*, 2007) has been proposed as an extension of k-anonymity which tries to address the attribute disclosure problem. A data set is said to satisfy l-diversity if, for each group of records sharing a combination of quasi-identifier attribute values (which we call an equivalence class), there are at least l "well-represented" values for each confidential attribute. Achieving l-diversity in general implies more distortion than just achieving k-anonymity.

Definition 2 (l-diversity) *An equivalence class is l-diverse if it contains at least l* "well-represented" *values for each confidential attribute. A data set is l-diverse if every equivalence class in it is l-diverse.*

t-Closeness (Li *et al.*, 2007) is another extension of k-anonymity which also tries to solve the attribute disclosure problem. A data set is said to satisfy t-closeness if, for each equivalence class, the distance between the distribution of each confidential attribute within the class and the distribution of the same confidential attribute over the whole data set is no more than a threshold t. This property clearly solves the attribute disclosure vulnerability, although the original t-closeness paper did not propose a computational procedure to achieve this property nor did it mention the large utility loss that this property is likely to inflict on the original data.

Definition 3 (t-closeness) *An equivalence class is said to satisfy t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute over the whole data set is no more than a threshold t. A data set is said to satisfy t-closeness if all its equivalence classes satisfy t-closeness.*

2.3 ε-Differential privacy

Differential privacy (Dwork, 2006; Dwork *et al.* 2006) aims at turning the probability of any output insensitive to the presence or absence of any individual in the data set.

Definition 4 (ε-Differential privacy) *A randomized function κ gives ε-differential privacy if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$, it holds that*

$$Pr(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times Pr(\kappa(D_2) \in S).$$

Given a query f, the goal is to find a randomized function κ_f that satisfies ε-differential privacy and approximates f as closely as possible. To attain differential privacy we need to introduce uncertainty to the query output. We can think of this uncertainty in terms of noise addition: $\kappa_f(D) = f(D) + N_f(D)$, where $N_f(D)$ is a random noise whose distribution may depend on f and D. A common choice for the random noise is a Laplace distribution with a scale parameter that depends on the L_1-sensitivity of f (the variability of f between data sets that differ in one record).

3 Data utility in k-anonymous data releases

To generate a k-anonymous data set, a partition of the original data set in groups of k (or more) records is performed, and the records within each of these groups are made indistinguishable w.r.t the quasi-identifiers, usually by setting the quasi-identifier attributes to a group-shared value. This creates equivalence classes. Current methods base the partitioning on the quasi-identifier attributes (*e.g.* Domingo-Ferrer and Torra, 2005; Samarati and Sweeney, 1998), that is, they group together records that are closest in terms of quasi-identifier values.

The partitioning strategy used to create equivalence classes is key to data utility. Assume a user who wants to analyze a specific group of individuals. The utility that this user derives from the k-anonymous data depends on how well the target group of individuals can be approximated by the equivalence classes. When the target group of individuals can be expressed as a union of equivalence classes, the best data utility is attained (regardless of the variability of the confidential attributes). For a target group of individuals that includes exactly one record from each of a set of equivalence classes, we are only able to determine a probability distribution for the confidential attribute related to each of the individuals (the one that assigns probability $1/k$ to each of the values of confidential attributes in the corresponding equivalence class). Therefore, the more concentrated is the distribution of the confidential attributes within the equivalence classes, the more utility we can extract from the k-anonymous data.

If the data collector is aware of the kind of analyses that data users are interested in, then the collector can tailor the partitioning to those analyses. However, one of the goals of microdata releases is to allow for arbitrary data analyses and, thus, most of the time the data collector is unaware of the intended use of the data, which renders customized partitioning infeasible. Even if the data collector knew the relevant analyses, different analyses might require different partitions, but releasing several versions of the same data set using a different partition each is not advisable, as it would endanger whatever anonymity is gained by partitioning.

3.1 Partitioning based on confidential attributes

When releasing microdata, the goal is usually to empower analysts to carry out arbitrary analyses on the data. Thus, customizing the k-anonymous partition to match the requirements of a specific analysis is not an option. Hence, the utility of the data depends on the amount of variability of the confidential attributes within each equivalence class.

To limit this variability, we propose to partition the records in the data set by means of a microaggregation algorithm based on the values of the confidential attributes. Microaggregation (Domingo-Ferrer and Torra, 2005) seeks to group the records in a data set into groups of cardinality k (or greater), in such a way that the within-groups homogeneity of the values of certain attributes is maximized.

3.2 Reducing information loss in quasi-identifier attributes

By partitioning based on the confidential attributes, we minimize the variability of these attributes within each of the partition groups. However, to satisfy k-anonymity, we still need to make the records within each partition group indistinguishable w.r.t. the quasi-identifier attributes. The values of the quasi-identifier attributes in each group of the partition may span the entire domains of those attributes (or substantial portions of them). Therefore, replacing all values of each quasi-identifier attribute within a group by a single generalized value would lead to a great utility loss. Moreover, note that the generalized values for the quasi-identifiers might coincide for different groups.

By generalizing the quasi-identifiers we seek to break the relationship between the quasi-identifier and the confidential part of a record. However, there are other ways of breaking that relationship that will allow us to improve data utility, such as the Anatomy approach (Xiao and Tao, 2006), which preserves the original values of the quasi-identifiers. To dissociate (break the relation between) quasi-identifiers and confidential attributes, two tables are generated: the first one assigns a group identifier to the quasi-identifiers, and the second one relates each group identifier to the confidential attributes.

3.3 Dealing with attribute disclosure

When generating the partition based on the quasi-identifiers, small values of the parameter k are typically used. Usually, the variability of the confidential attribute thus obtained is large enough not to lead to attribute disclosure. However, when basing the partition on the confidential attribute, small values of k will almost certainly lead to attribute disclosure, because in this case the within-group variability of the confidential attribute is small. Therefore, either we increase k to a large enough value so that disclosure is prevented, or we rely on additional criteria (*e.g.* l-diversity or t-closeness) to set a lower bound on the variability of the confidential data within each equivalence class.

Note that if partitioning is based on the quasi-identifiers, we cannot control the level of variability of the confidential attribute inside each of the equivalence classes: some of them may exhibit a large variability (which offers protection against attribute disclosure, but poor data utility) and others may not (which offers good data utility, but high risk of attribute disclosure). The problem is the impossibility of enforcing a predetermined amount of variability: variability increases with k, but the relationship between k and the amount of variability of the confidential attribute is not clear. Usually, k must be small (if any utility is to be provided), which results in poor disclosure limitation guarantees. The fact is that using a small value for k may, indeed, provide sufficient protection, but we have no guarantees of it.

4 ε-Differential privacy via t-closeness

t-Closeness and ε-differential privacy take approaches towards disclosure limitation that are essentially different. However, for microdata releases a link between them

can be found with some additional assumptions:

1. In the same way as k-anonymity does, we classify the attributes into quasi-identifiers and confidential attributes. Only the relation between quasi-identifier attributes and confidential attributes should be protected, so that an intruder cannot relate a specific identity to specific confidential attribute values.

2. The intruder's prior knowledge about the confidential attributes is limited to their distribution.

The link between t-closeness and ε-differential privacy relies on the fact that both models seek to limit the knowledge a user gets from accessing the released data. By taking a suitable distance (different from the usual earthmover's distance) and with the previous assumptions, we can have a t-closeness model with the semantics of differential privacy. ε-Differential privacy guarantees that, for any two data sets that differ in one individual, the probability for a query response computed on either data set to belong to an arbitrary set S differs at most by a factor $\exp(\varepsilon)$. The distance function we propose mimics the ε-differential privacy criterion.

Definition 5 *Given two random distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, we define the distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ as*

$$d(\mathcal{D}_1, \mathcal{D}_2) = \max\{\frac{Pr_{\mathcal{D}_1}(S)}{Pr_{\mathcal{D}_2}(S)}, \frac{Pr_{\mathcal{D}_2}(S)}{Pr_{\mathcal{D}_1}(S)}\}$$

where S is an arbitrary (measurable) set, and we take the quotients to be zero, if both $Pr_{\mathcal{D}_1}(S)$ and $Pr_{\mathcal{D}_2}(S)$ are zero, and to be infinity if only the denominator is zero.

In terms of t-closeness, the above distance implies that the distribution of the confidential attributes in the equivalence classes must differ from the global distribution of the confidential data by a factor not greater than t (and not smaller than $1/t$). We want to show that, in the case of a microdata release, $\exp(\varepsilon)$-closeness (for the distance function in Definition 5) implies ε-differential privacy. In other words, we mean that the information that an intruder obtains from accessing the released $\exp(\varepsilon)$-close microdata set satisfies the ε-differential privacy condition.

Let I be a specific individual in the data set. Before accessing the data set, the intruder views the value of the confidential attributes of individual I as being distributed according to the distribution of the confidential attributes over the whole data set. By the assumption that limits the prior knowledge to the distribution of the confidential attributes, that is the most precise information that the intruder possesses on I. ε-Differential privacy guarantees that the knowledge gain obtained from the response to a query that asks for I's confidential data is at most $\exp(\varepsilon)$; that is, the distribution of the response must differ at most by a factor of $\exp(\varepsilon)$ from the assumed prior knowledge. Note that if we did not take into account the intruder's prior knowledge (usual ε-differentially private mechanisms do not assume any prior knowledge), the $\exp(\varepsilon)$-differentially private distribution for the confidential attributes of individual I's would be different. However, the possibility of using the available prior knowledge exists, and thus any distribution that differs from it by a factor of $\exp(\varepsilon)$ satisfies $\exp(\varepsilon)$-differential privacy.

5 Conclusions

We have shown that the usual methods to attain k-anonymity, based on generalization of the quasi-identifiers, are not optimal in terms of the variability of the confidential attributes within each equivalence class. This is not an issue if the kind of analyses that will be performed on the released data can be anticipated (and the anonymization adjusted to it); however, that is hardly the case in real situations. For the released data set to be useful for arbitrary data analyses, the confidential attribute values within each of the equivalence classes should be as homogeneous as possible. To maximize such homogeneity, we have proposed a method to attain k-anonymity where the partitioning of the records is based on microaggregation of the confidential data.

By maximizing the homogeneity of the confidential data within equivalence classes, we improve data utility but we also raise the risk of attribute disclosure. To deal with attribute disclosure, we rely on refinements of the k-anonymity model (such as l-diversity or t-closeness) based on lower-bounding the amount of variability of the confidential attributes within equivalence classes.

We have also shown that the k-anonymity family of models is powerful enough to achieve ε-differential privacy in the context of data publishing. Specifically, we have shown that $\exp(\varepsilon)$-closeness implies ε-differential privacy for uninformed intruders.

Acknowledgments and disclaimer

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

Clifton, C., and Tassa, T. (2013) "On syntactic anonymity and differential privacy", *Transactions on Data Privacy*, **6(2)**, 161–183.

Domingo-Ferrer, J. and Torra, V. (2005) "Ordinal, continuous and heterogeneous k-anonymity through microaggregation", *Data Mining and Knowledge Discovery*, **11(2)**, 195–212.

Dwork, C. (2006) "Differential privacy". In *International Colloquium on Automata, Languages and Programming-ICALP 2006*, LNCS 4052, 1–12. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006) "Calibrating noise to sensitivity in private data analysis". In *Theory of Cryptography Conference-TCC 2006*, LNCS 3876, 265–284. Springer.

Li, N., Li, T., and Venkatasubramanian, S. (2007) "t-Closeness: privacy beyond k-anonymity and l-diversity". In *International Conference on Data Engineering-ICDE 2007*, 106–115. IEEE.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007) "l-Diversity: privacy beyond k-anonymity", *ACM Transactions on Knowledge Discovery from Data-TKDD*, **1(1)**, art. no. 3.

Samarati, P. (2001) "Protecting respondents' identities in microdata release", *IEEE Trans. Knowl. Data Eng.*, **13(6)**, 1010–1027.

Samarati, P., and Sweeney, L. (1998) "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". Research Report, SRI International, Menlo Park CA.

Xiao, X. and Tao, Y. (2006) "Anatomy: simple and effective privacy preservation". In *International Conference on Very Large Data Bases-VLDB 2006*, 139–150. ACM.