On the Connection between t-Closeness and Differential Privacy for Data Releases

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, Spain

Keywords: Differential Privacy, t-Closeness, k-Anonymity, Microaggregation.

Abstract: t-Closeness was introduced as an improvement of the well-known k-anonymity privacy model for data release. On the other hand, ε-differential privacy was originally proposed as a privacy property for answers to on-line database queries and it has been very welcome in academic circles. In spite of their quite diverse origins and motivations, we show in this paper that t-closeness and ε-differential privacy actually provide related privacy guarantees when applied to off-line data release. Specifically, k-anonymity for the quasi-identifiers combined with differential privacy for the confidential attributes yields t-closeness in expectation.

1 INTRODUCTION

There are several privacy models that have been proposed in the literature. k-Anonymity (Samarati and Sweeney, 1998; Samarati, 2001) and, more recently, ε-differential privacy (Dwork, 2006) stand out as probably the two best-known ones. The former was proposed to anonymize data sets for off-line release, whereas the latter was proposed to anonymize answers to interactive queries to on-line databases (Dwork, 2011). Yet, ε-differential privacy can also be extended to anonymize data sets.

Assume a data set X from which direct identifiers have been suppressed, but which contains so-called *quasi-identifier* attributes, that is, attributes (*e.g.* age, gender, nationality, etc.) which can be used by an intruder to link records in X with records in some external database containing direct identifiers. The intruder's goal is to determine the identity of the individuals to whom the values of confidential attributes (*e.g.* health condition, salary, etc.) in records in X correspond (*identity disclosure*).

A data set X is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in it is shared by at least k records. k-Anonymity protects against identity disclosure: given an anonymized record in X, an intruder cannot determine the identity of the individual to whom the record (and hence the confidential attribute values in it) corresponds. The reason is that there are at least k records in X sharing any combination of quasi-identifier attribute values. The most usual computational procedure to attain k-anonymity is generalization of the quasi-identifier attributes (Samarati, 2001), but an alternative approach is based on microaggregation of the quasi-identifier attributes (Domingo-Ferrer and Torra, 2005).

While k-anonymity protects against identity disclosure as mentioned above, in general it does not protect against *attribute disclosure* (Domingo-Ferrer, 2008), that is, disclosure of the value of a confidential attribute corresponding to an external identified individual. Let us assume a target individual T for whom the intruder knows the identity and the values of the confidential attributes. Let G_T be a group of at least k anonymized records sharing a combination of quasi-identifier attribute values that is the only one compatible with T's quasi-identifier attribute values. Then the intruder knows that the anonymized record corresponding to T belongs to G_T. Now, if the values for one (or several) confidential attribute(s) in all records of G_T are the same, the intruder learns the values of that (those) attribute(s) for the target individual T.

The property of l-diversity (Machanavajjhala *et al.*, 2006) 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, 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. Yet, l-diversity may fail to protect against attribute disclosure if the l values of a confidential attribute are very similar or are strongly skewed. p-Sensitive k-

Domingo-Ferrer J..
On the Connection between t-Closeness and Differential Privacy for Data Releases.
DOI: 10.5220/0004500904780481
In *Proceedings of the 10th International Conference on Security and Cryptography* (SECRYPT-2013), pages 478-481
ISBN: 978-989-8565-73-0

anonymity (Truta and Vinay, 2006) is a property similar to l-diversity, which shares similar shortcomings. See (Domingo-Ferrer, 2008) for a summary of criticisms to l-diversity and p-sensitive k-anonymity.

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 group of records sharing a combination of quasi-identifier attribute values, the distance between the empirical distribution of each confidential attribute within the group and the empirical distribution of the same confidential attribute in 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 and did not mention the large utility loss that this property is likely to inflict on the original data.

Differential privacy, as originally proposed for interactive databases, assumes that an anonymization mechanism mediates between the user submitting queries and the database. In this way, instead of getting responses to a query function f computed on the database, the user gets responses to a randomized query function κ. This randomized κ is said to satisfy 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 subsets S of the range of κ, it holds that

$$\Pr(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(\kappa(D_2) \in S). \qquad (1)$$

In plain words, Expression (1) means that the influence of any single record on the returned value of κ is negligible. The computational procedure originally proposed to reach ε-differential privacy is to obtain κ by adding Laplace noise to the query function f (Dwork, 2006).

We have recently shown in (Soria-Comas *et al.*, 2013) that microaggregation-based k-anonymity can be used as a prior step towards achieving ε-differential privacy of a data set. The advantage of doing so is that much less Laplace noise addition is thereafter needed to attain ε-differential privacy, in such a way that the utility of the resulting differentially private data is substantially higher.

1.1 Contribution and Plan of this Paper

In the same spirit of (Soria-Comas *et al.*, 2013) about finding connections between models based on k-anonymity and differential privacy, we explore here how t-closeness and ε-differential privacy are related to each other regarding anonymization of data sets.

We highlight the formal similarities between t-closeness and ε-differential privacy in Section 2. In the same section, we give a lemma showing that k-anonymity for the quasi-identifiers combined with differential privacy for the confidential attributes yields t-closeness in expectation. Section 3 is a conclusion.

2 FROM DIFFERENTIAL PRIVACY TO (EXPECTED) T-CLOSENESS

Let X be a data set with quasi-identifier attributes $Q_1, \cdots, Q_m$ and confidential attributes $C_1, \cdots, C_n$. Let N be the number of records of X. Further, let $I_r(\cdot)$ be the function that returns all the attribute values contained in record $r \in X$; let $IC_r(\cdot)$ be the function that returns the values of the confidential attributes in record $r \in X$.

Consider the multivariate query $(I_1(X), \cdots, I_N(X))$; the answer to that query returns the entire data set X. Further, let $(Y_1(X), \cdots, Y_N(X))$ be the noise that needs to be added to the answer to that query to achieve ε-differential privacy. A differentially private version of the data set X can be obtained as:

$$(I_1(X), \cdots, I_N(X)) + (Y_1(X), \cdots, Y_N(X)).$$

From the definition of ε-differential privacy (Expression (1)), it holds that

$$\Pr((I_1(X_1), \cdots, I_N(X_1)) + (Y_1(X_1), \cdots, Y_N(X_1)) \in S^N)$$
$$\leq \exp(\varepsilon) \times$$
$$\Pr((I_1(X_2), \cdots, I_N(X_2)) + (Y_1(X_2), \cdots, Y_N(X_2)) \in S^N) \qquad (2)$$

for any pair of data sets X_1, X_2 such one can be obtained from the other by suppressing/modifying a single record, and all $S \subset Range(I_i() + Y_i())$, where we assume this range to be the same for all $i = 1, \cdots, N$.

Let us now introduce expected t-closeness. This means t-closeness in expectation, that is at the level of the distributions of the noise used to generate the anonymized confidential attributes, respectively, within each group of records sharing a combination of quasi-identifier attributes and in the overall data set. Actual t-closeness (Li *et al.*, 2007), however, is defined in terms of the actual values obtained for the confidential attributes.

Definition 1 (Expected t-closeness)**.** *Let X' be an anonymized data set with N records obtained from an original data set X by k-anonymizing quasi-identifiers and adding random noise to the projection of X on its confidential attributes. Call the latter projection C*

and the corresponding noise-added projection C'. We say that X' satisfies expected t-closeness if

$$\Pr((IC_{i_1}(Z'),\cdots,IC_{i_{|Z'|}}(Z'))\in S^{|Z'|})$$

$$\le g(t)\times\Pr((IC_1(C'),\cdots,IC_N(C'))\in S^N) \tag{3}$$

for any subset $Z'\subseteq C'$ of records $i_1,\cdots,i_{|Z'|}$ sharing the same combination of quasi-identifier attribute values and all $S\subset Range(IC_i())$, where we assume this range to be the same for all $i=1,\cdots,N$, and where $g(\cdot)$ is a non-decreasing function such that the expected values of X' satisfy t-closeness in the sense of (Li et al.*, 2007).*

Note that expected t-closeness is defined in terms of the sampling distribution of the noise added to obtain C' from C. In other words, Definition 1 states that the noise added to X is expected to produce a data set X' for which actual t-closeness holds. It may occur, however, that the actual X' obtained does not satisfy t-closeness. Thus, in this respect, expected t-closeness is weaker than t-closeness.

The following lemma connects k-anonymity, ε-differential privacy and expected t-closeness. It says that if we k-anonymize the quasi-identifiers of an original data set and we make its confidential attributes ε-differentially private, then the resulting anonymized data set is expected to satisfy t-closeness for t a function of k and ε.

Lemma 1. *Let X be an original data set and X' be a corresponding anonymized data set such that its quasi-identifiers are k-anonymous and the projection of X' on the confidential attributes is ε-differentially private. Then X' satisfies expected t-closeness with $t=g^{-1}(\exp((N-k)\times\varepsilon))$.*

Proof. The projection C' of X' on its confidential attributes is derived from the corresponding projection C of X as:

$$C'=(I_1(C),\cdots,I_N(C))+(Y_1(C),\cdots,Y_N(C))$$

Let $Z\subset C$ be a group of k records with indices $i_1,\cdots,i_k$ sharing the same combination of quasi-identifier attribute values. Note that Z can be obtained from C by suppressing $N-k$ records from C. Now, if we iterate Expression (2) $N-k$ times, we get

$$\Pr((I_{i_1}(Z),\cdots,I_{i_k}(Z))+(Y_{i_1}(Z),\cdots,Y_{i_k}(Z))\in S^k)$$

$$\le\exp((N-k)\times\varepsilon)\times$$

$$\Pr((I_1(C),\cdots,I_N(C))+(Y_1(C),\cdots,Y_N(C))\in S^N) \tag{4}$$

By comparing with Expression (3), it can be seen that Expression (4) guarantees that X' satisfies expected t-closeness with $t=g^{-1}(\exp((N-k)\times\varepsilon)$. □

Note 1. The previous lemma gives a computational procedure to obtain t-closeness, albeit a greedy one: just keep generating differentially private versions of C by random noise addition until a version C' is obtained which satisfies actual t-closeness in the sense of (Li *et al.*, 2007). Of course, the larger the number of records, the larger the number of attributes in C and the larger the variance of the noise distribution used, the longer it will take to terminate this procedure.

3 CONCLUSIONS AND FUTURE WORK

In previous work, we showed how k-anonymity could be used as a prior step to obtain differentially private data releases with higher utility. In the same line of finding synergies between privacy models, in this paper we have highlighted the formal similarity between ε-differential privacy and t-closeness for anonymization of data sets. Furthermore, we have shown how expected t-closeness can be obtained from ε-differential privacy.

In future work we plan to build on the ideas in this paper and leverage differential privacy to achieve actual t-closeness in a way less greedy that the one sketched in Note 1 above. This will address one of the weak points of the original t-closeness proposal, namely the lack of a computational procedure to reach that property.

ACKNOWLEDGMENTS AND DISCLAIMER

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 “E-AEGIS”, 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 author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. He is with the UNESCO Chair in Data Privacy, but he is solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

REFERENCES

Domingo-Ferrer, J. (2008). A critique of *k*-anonymity and some of its enhancements. In *Proceedings of ARES/PSAI 2008*, IEEE Computer Society, pp. 990-993.

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 *Proc. 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, LNCS 4052, Springer, pp. 1-12.

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86-95.

Li, N., Li, T., and Venkatasubramanian, S. (2007). *t*-Closeness: privacy beyond *k*-anonymity and *l*-diversity. In *Proceedings of IEEE ICDE 2007*.

Machanavajjhala, A., Gehrke, J., Kiefer, D., and Venkitasubramanian, M. (2006) *l*-Diversity: privacy beyond *k*-anonymity. In *Proceedings of IEEE ICDE 2006*.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

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

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S. (2013). Improving the utility of differentially private data releases via *k*-anonymity. In *12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications -IEEE TrustCom 2013*, Melbourne, Australia, July 16-18, 2013 (to appear).

Truta, T.M. and Vinay, B. (2006). Privacy protection: *p*-sensitive *k*-anonymity property. In *2nd International Workshop on Privacy Data Management PDM 2006*, IEEE Computer Society, p. 94.