t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation (Extended abstract)

Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez and Sergio Martínez
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Universitat Rovira i Virgili, Tarragona, Catalonia
Email: {jordi.soria, josep.domingo, david.sanchez, sergio.martinezl}@urv.cat

***Abstract*—This paper proposes and shows how to use microaggregation to attain t-closeness on top of k-anonymity to protect data releases. The advantages in terms of data utility preservation of microaggregation over classic approaches based on generalizing values are analyzed. Then several microaggregation algorithms for k-anonymous t-closeness are presented and empirically evaluated.**

I. Introduction

Generating an anonymized data set that is suitable for public release is essentially a matter of finding a good equilibrium between disclosure risk and data utility preservation.

To address this task, a plethora of data protection methods for data set releases have been developed in both the statistical and computer science communities. We focus here on privacy models (which provide *ex ante* privacy guarantees) and, more specifically, on the t-closeness model, which offers protection against confidential *attribute disclosure*.

Algorithms for attaining t-closeness are mostly based on generalizing and suppressing values. In fact, they are slightly modified versions of the algorithms used to attain k-anonymity that, additionally, take into account the restrictions introduced by t-closeness on the distribution of the confidential attributes.

In contrast, in [1] we propose and evaluate three algorithms based on data microaggregation to attain t-closeness on top of k-anonymity. Unlike generalization-based methods, microaggregation perturbs data, which allows improving data utility in several ways, such as increasing data granularity, reducing the impact of outliers and avoiding discretization of numerical data.

II. Background

A data set can be modeled as a table where each row contains data on a different subject and each column corresponds to a specific attribute.

Statistical disclosure control seeks to limit the risk of disclosing confidential information about any specific individual. To this end, a masked version T' of the original data set T is released.

A. k-Anonymity

The goal of k-anonymity [2] is to limit the capability of the intruder to perform successful re-identifications of records. It is assumed that the attributes that can be used for record re-identification are known. They are referred to as *quasi-identifiers*. To prevent re-identification, k-anonymity requires any combination of quasi-identifier values present in the data set to be shared by k or more records.

B. t-Closeness

While k-anonymity prevents attackers from performing accurate re-identifications of records (thus protecting against identity disclosure), it does not protect against attribute disclosure. For example, when the variability of the confidential attribute in the k-anonymous groups is too small, being able to tell the group of an individual may already reveal too much information about the individual (e.g., if all the members of the group share a high salary or the same type of sensitive disease).

t-Closeness [3] seeks to prevent attribute disclosure from happening. This is done by making the distribution of the confidential attributes in each group similar to their distribution in the entire data set. A distance (usually the Earth Mover's Distance, EMD) is used to determine the similarity between distributions.

C. Microaggregation

Microaggregation is a family of methods for disclosure control [4]. It consists of two steps: first, the records in the original data set are partitioned into clusters, each of them containing k or more *similar* records; then, an aggregation operation (*e.g.* the mean) is used to summarize the data in each cluster and the original records are replaced by the aggregated output. By applying this procedure on the quasi-identifier attributes, we obtain an output that straightforwardly satisfies k-anonymity.

The partition and aggregation steps produce some information loss that may hamper data utility. Microaggregation algorithms aim at minimizing the information loss according to some metric (*e.g.* the sum of squared errors).

III. t-Closeness via Microaggregation

A. Standard microaggregation and merging

In our first algorithm, microaggregation is run on the quasi-identifier attributes of the original data; this step produces a k-anonymous data set. Then, clusters of microaggregated values are merged until t-closeness is satisfied. We iteratively improve the level of t-closeness by (i) selecting the cluster whose confidential attribute distribution is most different from the confidential attribute distribution in the entire data set (that is, the cluster farthest from satisfying t-closeness), and (ii) merging it with the cluster closest to it in terms of quasi-identifiers.

B. t-Closeness aware microaggregation

In our second algorithm, microaggregation is modified so that each cluster is generated in a way that satisfies t-closeness. Initially, a cluster of size k is generated based on the quasi-identifier attributes. Then the cluster is iteratively refined until t-closeness is satisfied by (i) adding the record that, based on the quasi-identifiers, is closest to the cluster, and (ii) removing a record so that the t-closeness distance (e.g., EMD) is minimized.

C. t-Closeness-first microaggregation

In our third algorithm, rather than first performing the microaggregation on the quasi-identifiers and then dealing with t-closeness, we take the opposite approach. First, the records in the original data set (sorted by their confidential attribute value(s)) are partitioned into k subsets so that t-closeness is satisfied for the groups of records formed by taking one record from each subset. Then microaggregation based on the quasi-identifiers is performed with the constraint that clusters must contain one record from each subset.

In [1] we formally show that, by construction, the generated clusters satisfy t-closeness, so there is no need to compute any distance. This provides a great advantage in terms of computational cost over the two previous algorithms.

IV. Experimental Evaluation

In [1], we evaluated and compared our three algorithms for several data sets according to the actual cluster size they yield, their speed and scalability, and the information loss incurred. As an example, Figure 1 illustrates the information loss metric (measured as the normalized SSE) with a data set of patient discharges.

Algorithm 1 improves on Algorithm 2 and, in turn, Algorithm 3 improves on Algorithm 2. Thus, the earlier we satisfy t-closeness in the microaggregation, the less information is lost and the more utility is preserved. In practice, the fact that the k-anonymous microaggregation is aware of the level of t-closeness that should be satisfied (Algorithm 3) produces smaller clusters (of size closer to the desired k), which is beneficial to keep the SSE low. In contrast, the other algorithms prioritize quasi-identifier values in the k-anonymous microaggregation and, hence, they require a lot of cluster merging and/or manipulation to attain t-closeness; this tends to produce larger clusters whose aggregation incurs a greater loss of information. To sum up, the increase of information loss that the lower cluster homogeneity of t-closeness aware microaggregation might cause is more than compensated by the information loss reduction resulting from smaller clusters.

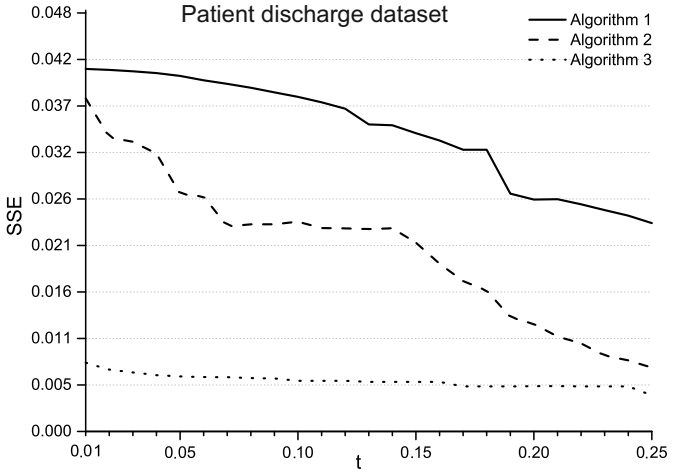


Fig. 1. Normalized SSE for the three algorithms for 2-anonymity with varying t-closeness levels

V. Conclusion

We have proposed several utility-preserving microaggregation-based algorithms to attain t-closeness on top of k-anonymity. Experimental evaluation has shown that it is better to take into account the requirements of t-closeness in the microaggregation process. In particular, t-closeness-first microaggregation produced the best results, both in data utility preservation and in speed.

Acknowledgment and disclaimer

This work was supported by the European Commission through project H2020 "CLARUS", by the Spanish Government (through projects "ICWT" TIN2012-32757, "CO-PRIVACY" TIN2011-27076-C03-01 and TIN2014-57364-C2-R "SmartGlacis") and by the Government of Catalonia under grant 2014 SGR 537. Josep Domingo-Ferrer is partially supported as an ICREA-Acadèmia researcher by the Government of Catalonia. The opinions expressed in this paper are the authors' own and do not necessarily reflect the views of UNESCO.

References

[1] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "t-closeness through microaggregation: strict privacy with enhanced utility preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098–3110, Nov 2015.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, Tech. Rep., 1998.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," in *In Proc. of the 23rd IEEE International Conference on Data Engineering*, 2007, pp. 106–115.

[4] J. Domingo-Ferrer and J. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 189–201, 2002.