μ-ANT: Semantic Microaggregation-based Anonymization Tool

David Sánchez[1*], Sergio Martínez[1], Josep Domingo-Ferrer[1], Jordi Soria-Comas[1] and Montserrat Batet[2]

[1]Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, E-43007 Tarragona, Catalonia

[2] CYBERCAT-Center for Cybersecurity Research of Catalonia, Internet Interdisciplinary Institute, Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, E-08860 Castelldefels, Catalonia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Detailed patient data are crucial for medical research. Yet, these healthcare data can only be released for secondary use if they have undergone *anonymization*.

Results: We present and describe *μ-ANT*, a practical and easily configurable anonymization tool for (healthcare) data. It implements several state-of-the-art methods to offer robust privacy guarantees and preserve the utility of the anonymized data as much as possible. *μ-ANT* also supports the heterogenous attribute types commonly found in electronic healthcare records and targets both practitioners and software developers interested in data anonymization.

Availability (source code, documentation, executable, sample datasets and use case examples): https://github.com/CrisesUrv/microaggregation-based_anonymization_tool

Contact: david.sanchez@urv.cat

1 Introduction

The availability of detailed patient data is crucial for medical research (Arandjelovic, 2015); yet, legal frameworks such as the recent EU GDPR state that personally identifiable information (and healthcare information in particular) cannot be released unless it is subjected to *anonymization*. Data anonymization prevents the identities of subjects from being associated with the released data, while preserving analytical utility as much as possible. In particular, data anonymization should limit two privacy risks: *identity disclosure*, by which external entities may identify the record of a known individual in the released dataset, and *attribute disclosure*, by which the confidential attributes of a known individual may be associated with her.

To properly anonymize data, it is not enough to suppress personal identifiers. It is widely acknowledged (Samarati, 2001; Sánchez, et al., 2016) that combinations of non-identifying attributes (such as gender+age+zipcode) may unequivocally identify a subject. The reason is that these attributes, known as *quasi-identifiers* (QIs), may be present in public non-confidential databases (such as electoral rolls) together with some identifiers. Hence, they could be used to reidentify subjects and thereby discover their confidential attributes. Thus, QIs should be masked to prevent reidentification.

During the last few years, we have developed state-of-the-art methods to anonymize data releases that aim at i) offering robust privacy guarantees against identity and attribute disclosures and ii) preserving the utility of the anonymized data as much as possible, regardless of the data types. These methods have been now implemented in the *μ-ANT* software, which is conceived as a practical and easily configurable anonymization tool for (healthcare) data.

2 Software highlights

To mask QIs, *μ-ANT* relies on *k*-anonymity (Samarati, 2001). *k*-Anonymity requires records to be indistinguishable from at least *k*-1 other records regarding QI values. This ensures that the reidentification probability of any individual is at most 1/*k*, thereby preventing unequivocal reidentifications when *k*>1.

Existing anonymization software enforcing *k*-anonymity, such as the *UTD anonymization toolbox* (Kantarcioglu, et al., 2012) and *ARX* (Prasser, et al., 2019), masks quasi-identifiers by suppressing attribute values or generalizing them. Whereas this prevents reidentification, it significantly reduces the *detail* of the data, which decreases the utility of the anonymized dataset (Soria-Comas, et al., 2015). In contrast, *μ-ANT* uses *microaggregation* as a more utility-preserving alternative. With microaggregation, clusters each containing at least *k* similar records are created, and they are made indistinguishable by replacing attribute values by cluster averages. Microaggregation offers several advantages w.r.t. generalization (and *a fortiori* w.r.t. data suppression) because i) the granularity of the input data is not decreased, ii) continuous numerical attributes are not discretized, and iii) outliers cause less distortion (whereas they may force data generalization to very coarse values).

To preserve the analytical utility of the anonymized data, microaggregated records should be clustered according to their *similarity*.

This increases the homogeneity of the clusters and reduces the information loss incurred when replacing QI values by cluster averages. To this end, we need methods to i) compare attribute values and ii) compute averages. Both operations are trivial for numerical attributes but not for the *nominal* categorical attributes that commonly appear in healthcare datasets (e.g., diagnosis). State-of-the-art anonymization tools (*ARGUS* (Hundepool, et al., 2018) *ARX* (Prasser, et al., 2019), *UTD anonymization toolbox* (Kantarciouglu, et al., 2012) and *scdmicro* (Templ, et al., 2019)) treat nominal attributes as *plain* categorical attributes whose values i) can only be compared for equality/inequality and ii) can only be aggregated with distributional operators such as the mode. As a result, the semantics underlying nominal attributes are severely coarsened (Martínez, et al., 2013). In contrast, *μ-ANT* is the first anonymization software that implements semantic mechanisms to manage nominal values. Specifically, it supports or implements i) *OWL ontologies* modeling the domains of nominal attributes, ii) a state-of-the-art semantic *similarity measure* (Sánchez, et al., 2012) to compare nominal values according to their meanings, and iii) a semantic averaging operator that finds the cluster centroid that best represents the meaning of the cluster members (Martínez, et al., 2012).

μ-ANT also offers protection against *attribute disclosure*. In a *k*-anonymous dataset it may happen that a set of *k* indistinguishable records shares the same confidential value (e.g., the same diagnosis). In such a case, even if an attacker cannot reidentify a target subject within a cluster, he will unequivocally infer the confidential attribute of the target subject because it is the same for all cluster subjects. To protect against attribute disclosure, the microaggregation algorithm implemented in *μ-ANT* also considers, for the first time, the *t*-closeness model (Li and Li, 2007). *t*-Closeness requires the distribution of confidential attribute values within each *k*-anonymous cluster to be similar to the distribution of the attributes in the entire dataset. To satisfy this, our microaggregation algorithm uniformly samples input records prior to creating the groups, thereby ensuring that the distribution of confidential values within each group is similar (up to *t*) to that of the dataset (Soria-Comas, et al., 2015). The combination of *k*-anonymity and *t*-closeness provides more robust *ex ante* privacy guarantees than the rules defined by the HIPAA to protect medical data (Sánchez, et al., 2016).

On a more technical side, *μ-ANT* is offered as a standalone open source Java anonymization tool supporting standard CSV datasets and the heterogenous attribute types typically found in medical datasets. It uses a well-documented XML configuration file that allows specifying which attributes should be protected and how. The user just needs to list the attributes in the dataset to be anonymized, their types (i.e., *continuous numerical*, *discrete numerical*, *date*, *plain categorical* or *semantic nominal* -with an associated OWL ontology-), their sensitivity (i.e., *identifier*, *quasi-identifier*, *confidential* and *non-confidential*) and the desired anonymization parameters to be employed during microaggregation (i.e., *k* and *t* values for *k*-anonymity on QIs and *t*-closeness on confidential attributes, respectively). For medical nominal attributes (e.g., diagnoses, medical procedures, etc.), *μ-ANT* incorporates a tool that transforms the latest SNOMED-CT data files to an OWL ontology that will be used to semantically manage attribute values. QIs are normalized by their variance in the dataset to prevent attributes with wide ranges from dominating attributes with narrower ranges. In addition to the anonymized dataset, *μ-ANT* also calculates several utility metrics quantifying the information loss incurred by the anonymization process.

μ-ANT algorithms scale as $O(n \log n)$ w.r.t. the number of records (n) and as $O(m)$ w.r.t. the number of attributes (m) for the most complex case (i.e., *t*-closeness on top of *k*-anonymity). Thus, the algorithms are highly scalable, thereby making our software suitable for large datasets. Available memory should allow loading the dataset into RAM.

μ-ANT's data anonymization can be easily executed by non-developers from the command line, by indicating the CSV dataset to be anonymized and the configuration file. *μ-ANT* also offers a well-documented Java API for developers that want to embed the anonymization algorithm in their own applications.

The technical soundness of the methods implemented in *μ-ANT* is endorsed by a variety of scientific publications (Martínez, et al., 2013; Martínez, et al., 2012; Sánchez, et al., 2012; Soria-Comas, et al., 2015). The implementation has been empirically evaluated with real and heterogenous medical data, namely patient discharge data containing millions of records of Californian hospitals (Sánchez, et al., 2016). Finally, the anonymization software has been validated in a real healthcare use case at Barcelona's Hospital Clinic, consisting in outsourcing anonymized patients' records to public clouds (see http://clarussecure.eu).

Funding

This work was partly supported by the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705) and the Spanish Government (projects RTI2018-095094-B-C2 "CONSENT" and TIN2016-80250-R "Sec-MCloud").

References

Arandjelovic, O. (2015) Discovering hospital admission patterns using models learnt from electronic hospital records, *Bioinformatics*, **31**, 3970–3976.

Hundepool, A., *et al.* (2018) ARGUS.

Kantarciouglu, M., Inan, A. and Kuzu, M. (2012) UTD anonymization toolbox.

Li, N. and Li, T. (2007) t-Closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd. International Conference on Data Engineering*. Istanbul, pp. 106-115.

Martínez, S., Sánchez, D. and Valls, A. (2013) A semantic framework to protect the privacy of electronic health records with non-numerical attributes, *Journal of Biomedical Informatics*, **46**, 294-303.

Martínez, S., Valls, A. and Sánchez, D. (2012) Semantically-grounded construction of centroids for datasets with textual attributes, *Knowledge-Based Systems*, **35**, 160-172.

Prasser, F., *et al.* (2019) ARX - data anonymization tool.

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

Sánchez, D., *et al.* (2012) Ontology-based semantic similarity: a new feature-based approach, *Expert Systems with Applications*, **39**, 7718-7728.

Sánchez, D., Martínez, S. and Domingo-Ferrer, J. (2016) Comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata', *Science,* **351**, 1274.

Soria-Comas, J., *et al.* (2015) t-Closeness through microaggregation: strict privacy with enhanced utility preservation, *IEEE Transactions on Knowledge and Data Engineering*, **27**, 3098-3110.

Templ, M., Meindl, B. and Kowarik, A. (2019) sdcmicro: Statistical disclosure control methods for anonymization of microdata.