TECHNICAL COMMENT

IDENTITY AND PRIVACY

Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"

David Sánchez,* Sergio Martínez, Josep Domingo-Ferrer

De Montjoye *et al*. (Reports, 30 January 2015, p. 536) claimed that most individuals can be reidentified from a deidentified transaction database and that anonymization mechanisms are not effective against reidentification. We demonstrate that anonymization can be performed by techniques well established in the literature.

De Montjoye *et al*. (*1*) concluded that, for most customers in a deidentified credit card transaction database, knowing the spatiotemporal features of four purchases by the customer was enough to reidentify her. Reidentification was measured according to "unicity" (*2*) (a neologism for uniqueness), which, given a number of personal features assumed to be known to an attacker, counts the number of individuals for whom these features are unique.

First, de Montjoye *et al*.'s reidentification figures are probably overestimated, because their database of 1.1 million customers seems only a fraction of the population of an undisclosed country (presumably, several millions). Unfortunately, they did not make their data set public, which prevents reproducing their results. As highlighted by Barth-Jones *et al*. (*3*), with a nonexhaustive sample, an individual's sample uniqueness/unicity does not imply population uniqueness and, hence, does not allow unequivocal reidentification; assuming otherwise clearly overestimates the reidentification risk. Moreover, not even population uniqueness automatically yields reidentification: The attacker still needs to link the records with unique features to external identified data sources (e.g., electoral rolls).

To reduce the high unicity in their database, de Montjoye *et al*. implemented some unreferenced "anonymization" strategies to coarsen data (such as clustering locations) that fell short of sufficiently reducing unicity. From this, de Montjoye *et al*. drew conclusions about the ineffectiveness of anonymization methods and highlighted the need for "more research in computational privacy."

We must recall that reidentification risk in data releases has been treated in the statistical disclosure control (*4*, *5*) and privacy-preserving data publishing (*6*) literatures for nearly four decades. As a result, a broad choice of anonymization methods exists. These usually suppress personal identifiers (such as passport numbers) and mask quasi-identifiers (QIs). The latter are attributes such as zip code, job, or birthdate, each of which does not uniquely identify the subject but whose combination may. Because QIs may be present in public nonconfidential data bases (like electoral rolls) together with some identifiers (like passport number), it is crucial to mask them to avoid reidentification. It is easy to see that reidentification through QIs [studied at least since 1988 (*7*) and popularized by the k-anonymity model (*8*)] is equivalent to the unicity idea rediscovered by de Montjoye *et al*. in 2013 (*2*)—that is, a subject whose QI values are unique in a data set risks being reidentified.

Data coarsening is indeed a method often used in anonymization to mask QIs (*8*). However, de Montjoye *et al*. concluded that their coarsening-based anonymization was ineffective. This is unsurprising because they coarsened attributes independently and used value ranges fixed ex ante, which is inappropriate for at least two reasons: (i) to offer true anonymity guarantees, coarsening should be based on the actual distribution of the data set (i.e., a fixed range may contain a single value among those in the data set); and (ii) independently coarsening each QI attribute cannot ensure that unique QI value combinations disappear (coarsening must consider all QIs together).

To illustrate the effectiveness of anonymization, the simple and well-known k-anonymity notion is enough. In a k-anonymous data set, records should not include strict identifiers, and each record should be indistinguishable from, at least, $k - 1$ other ones regarding QI values. Thus, the probability of reidentification of any individual is $1/k$. Hence, for $k > 1$, this probability is less than 1 for all records, thereby ensuring zero unequivocal reidentifications. Moreover, by tuning k, we can also tune the level of exposure of individuals.

We looked for a data set with similar structure and unicity/reidentification risk properties to de Montjoye *et al*.'s data (which were unavailable) to show the effectiveness of k-anonymity. We chose a synthetically generated version of a publicly available patient discharge data set (that we call SPD), which includes the nearly 4 million patients admitted in 2009 to California hospitals [see details at (*9*)]. This data set includes a set of spatiotemporal features of the patients and, unlike de Montjoye *et al*.'s data set, it covers the whole population of 2009 California patients; hence, uniqueness in this data set quantifies the population reidentification risk (*9*). As shown in Fig. 1 the high risk reached for the SPD data set when the attacker knows all the patient's features (75%) is coherent with the high (even though overestimated) unicities reported by de Montjoye *et al*..

We enforced k-anonymity by grouping records with similar QIs (census + spatiotemporal features) in clusters of k or more and generalizing/coarsening their QI values to their common range (*9*).

Figure 2 compares the risk of unequivocal reidentification and correct random reidentification of k-anonymity versus a naïve coarsening similar to de Montjoye *et al*.'s, with "fixed" intervals covering 1/32, 1/16, and 1/8 of the domain ranges of the attributes (*9*). Unlike naïve coarsening, k-anonymity

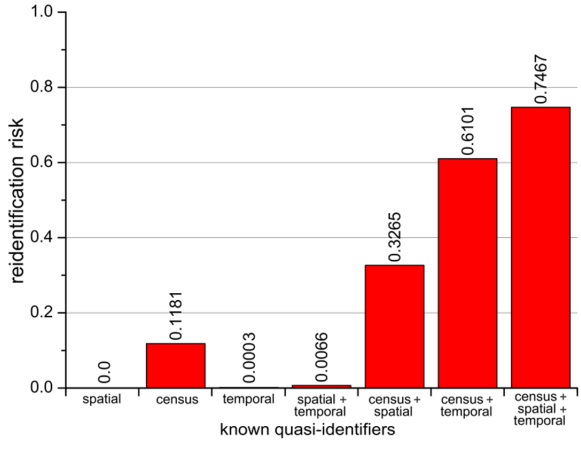

Fig. 1. Reidentification risk in the SPD data set depending on the attributes known by the attacker.

United Nations Educational, Scientific, and Cultural Organization (UNESCO) Chair in Data Privacy, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili (URV), Avenue Països Catalans, 26, E-43007, Tarragona, Catalonia.
*Corresponding author. E-mail: david.sanchez@urv.cat

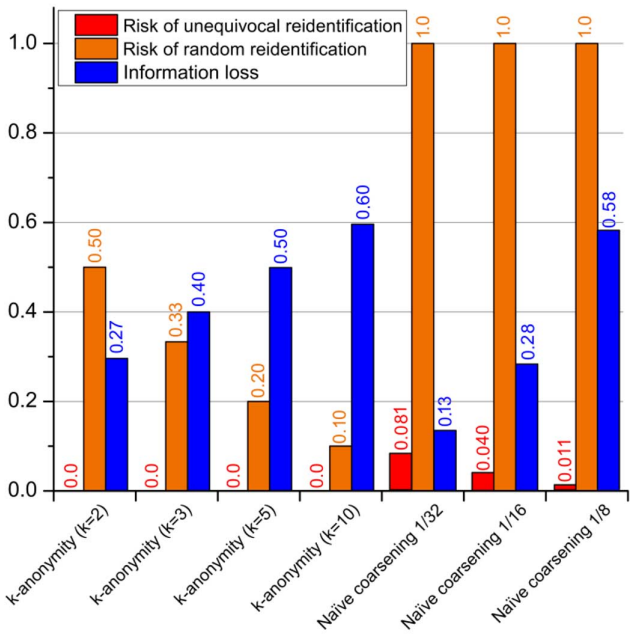

Fig. 2. Reidentification risk and information loss for *k*-anonymity and naïve coarsening.

yielded zero unequivocal reidentifications and a rate $1/k$ of correct random reidentifications when the attacker knows all QIs.

Furthermore, anonymized data should also retain analytical utility, which ultimately justifies data publishing. With *k*-anonymity, data utility is retained by grouping similar records together and by masking only those that do not fulfill the privacy criterion (de Montjoye *et al.*'s naïve coarsening fails to do either). Moreover, the trade-off between privacy and utility can be balanced by adjusting *k*. To illustrate, we have measured the information loss incurred by masking as the average distance between the SPD data set and its anonymized versions (*9*). Figure 2 shows that 2-anonymity not only yields less reidentifications but also less information loss than the safest naïve coarsening.

In addition to *k*-anonymity, there is much more in the anonymization literature. Specifically, extensions of *k*-anonymity [e.g., *t*-closeness (*10*)] also address attribute disclosure, which occurs if the values of the confidential attributes within a group of records sharing all QI values are too close. In (*9*), we report how *t*-closeness mitigates attribute disclosure by using the algorithm we proposed in (*11*). Moreover, the current research agenda includes more challenging scenarios, like big-data anonymization (in which scalability and linkability preservation are crucial) (*12*, *13*), streaming data anonymization (*14*), and local or co-utile collaborative anonymization by the data subjects themselves (*15*).

In conclusion, data owners and subjects can be reassured that sound anonymization methodologies exist to produce useful anonymized data that can be safely shared for research.

REFERENCES AND NOTES

1. Y.-A. de Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland, *Science* **347**, 536–539 (2015).
2. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, *Sci. Rep.* **3**, 1376 (2013).
3. D. Barth-Jones, K. El Emam, J. Bambauer, A. Cavoukian, B. Malin, *Science* **348**, 194–195 (2015).
4. T. Dalenius, *Stat. Tidskr.* **5**, 429–444 (1977).
5. A. Hundepool *et al.*, *Statistical Disclosure Control* (Wiley, 2012).
6. B. Fung, K. Wang, R. Chen, P. S. Yu, *ACM Comput. Surv.* **42**, 14 (2010).
7. G. Paass, *J. Bus. Econ. Stat.* **6**, 487–500 (1988).
8. P. Samarati, L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" (SRI International, 1998).
9. D. Sánchez, S. Martínez, J. Domingo-Ferrer, Supplementary materials for "How to avoid reidentification with proper anonymization." Available at http://arxiv.org/abs/1511.05957 (2015).
10. N. Li, T. Li, t-Closeness, in *IEEE 23rd International Conference on Data Engineering* (IEEE Computer Society, 2007), pp. 106–115.
11. J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, *IEEE Trans. Knowl. Data Eng.* **27**, 3098–3110 (2015).
12. A. Machanavajjhala, J. Reiter, *XRDS: Crossroads* **19**, 20–23 (2012).
13. J. Soria-Comas, J. Domingo-Ferrer. Big data privacy: Challenges to privacy, principles and models. *Data Sci. Eng.* (published online Sep. 15, 2015); http://dx.doi.org/10.1007/s41019-015-0001-x
14. J. Cao, B. Carminati, E. Ferrari, K.-L. Tan, *IEEE T. Depend. Secure* **8**, 337–352 (2011).
15. J. Soria-Comas, J. Domingo-Ferrer, in *MDAI 2015-Modeling Decisions for Artificial Intelligence* (Springer, 2015), pp. 192–206.

ACKNOWLEDGMENTS

Additional materials for this study are available at http://arxiv.org/abs/1511.05957. They detail the structure and synthetic generation of the SPD data set, describe the risk assessment and anonymization algorithms we used to obtain the reported results, and provide extended results and discussions. The SPD data set, with the synthetic quasi-identifiers, and its *k*-anonymous, *k*-anonymous and *t*-close, and coarsened versions can be found at http://crises-deim.urv.cat/opendata/SPD_Science.zip. The source code of the algorithms detailed in the additional materials is also available, together with some usage examples. These materials allow reproducing all results reported here and in the additional materials. Thanks go to J. Bambauer, A. Cavoukian, K. El Emam, K. Muralidhar, and V. Torra for useful reviews and discussions.
We gratefully acknowledge the following funding sources and grants: European Commission (H2020-644024 "CLARUS"), Spanish Government (TIN2012-32757 and TIN2014-57364-C2-1-R), Government of Catalonia (2014 SGR 537 and ICREA-Acadèmia award to J.D.-F.), and Templeton World Charity Foundation (TWCF0095/AB60). The opinions expressed in this paper are the authors' own and do not necessarily reflect the views of any funder or UNESCO.

23 November 2015; accepted 23 December 2015
10.1126/science.aad9295

Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"

David Sánchez *et al.*

Science **351**, 1274 (2016);
DOI: 10.1126/science.aad9295