Response to Comment on “Unique in the shopping mall: On the reidentifiability of credit card metadata”

Yves-Alexandre de Montjoye and Alex “Sandy” Pentland

Sánchez et al.’s textbook k-anonymization example does not prove, or even suggest, that location and other big-data sets can be anonymized and of general use. The synthetic data set that they “successfully anonymize” bears no resemblance to modern high-dimensional data sets on which their methods fail. Moving forward, deidentification should not be considered a useful basis for policy.

We believe that Sánchez et al. (1) fundamentally misunderstand the size and dimensionality of modern big-data data sets and how they are being used in industry and research. Making data available for socially beneficial applications is vitally important. We are therefore highly concerned by the failure of some in the “statistical disclosure control” community to reassess the limits of data anonymization (or deidentification) in the face of fast technological evolution. This insistence on sticking with “how we’ve done it for 40 years” risks (i) forcing us, as a society, to suffer either a dramatic loss of privacy or a sharp reduction in the availability and use of data in the coming decade, and (ii) preventing the development and adoption of modern privacy-through-security solutions for big data (2).

The textbook analysis presented by Sánchez et al. does not prove or even suggest that high-dimensional data sets, such as the one generated from credit cards, mobile phones, browsers, or the Internet of things, where hundreds or thousands of points are known for each individual across years, can be effectively anonymized. Specifically, the synthetic medical data set that Sánchez et al. successfully anonymize bears no resemblance to the high-dimensional data sets to which we refer in our studies (3, 4) and which their textbook method would utterly fail to anonymize. Their data set contains a total of nine points, or quasi-identifiers (information that could be used to reidentify someone), per individual and cannot be used to track an individual across time.

Mobile phones typically record a person’s location—also a quasi-identifier (5)—from every couple of hours to every 5 minutes, and payment systems up to a couple of times a day, often generating data sets that contain hundreds to thousands of points per individual across time. The intrinsic anonymizability of a data set is substantially driven by basic combinatorics. Showing that a method can anonymize a small (0.027 GB compressed) and low-dimensional data set of nine points does not prove anything about its ability to anonymize modern high-dimensional data sets containing hundreds or thousands of points per individual, such as a person’s location in a country every 5 minutes for a year. In fact, the data set studied by Sánchez et al. is trivially anonymized by any method, including the one we used, which, when applying the least aggressive generalization, already decreases unicity from 0.7467 to 0.081 by Sánchez et al.’s calculations. In contrast, we showed that unicity only decreases very slowly with both spatial and temporal generalization in mobile phone and credit card data sets (3, 4). We furthermore showed that an attacker can easily compensate for this decrease by collecting a few more external points: knowing that the person we are searching for was at a given place at a given time.

In addition, showing that a data set can be anonymized and still be useful for a specific application—which Sánchez et al. do not show—is not sufficient. The privacy guarantees offered by anonymization methods such as the one used by Sánchez et al. only hold if the data are anonymized one time (i.e., one anonymization method with one set of parameters per data set ever). The same mobile phone data are, however, already being used to study human mobility and behavior in cities (6), the geographical partitioning of countries (7), and the spread of information in social networks (8). To argue that sound anonymization methods are sufficient to protect people’s privacy in mobile phone or credit card data, one would need to show that a single anonymization method can anonymize the data and yet allow for most present—and hopefully future—data uses. We currently have no reason to believe that such a method will ever exist for modern high-dimensional data sets.

Furthermore, Sánchez et al. claim that the anonymization method that we used is suboptimal. The choice of a specific anonymization method and set of parameters depends on how the data will be used. The one we picked—lowering the spatial and temporal resolution of the data—is both a general and a natural one. Although, by definition, one can never rule out the existence of a substantially more efficient anonymization method, the authors do not present evidence of one. Our analysis furthermore shows that, even if a new method were to be twice as effective as ours, one would still have to decrease the spatial and temporal resolution of the data by a factor of 15 to approach a reasonably low unicity knowing 10 points (9). This means that the location of an individual would only be known every 15 hours with an accuracy of roughly 15 km², raising doubts on the general utility of this data. In fact, one study that actually attempted to (k-)anonymize high-dimensional location data through trajectory-based clustering (nonorthogonal generalization) concluded that their results are in agreement with ours and “provide insights behind the poor anonymizability” of mobile phone data sets (9).

Finally, Sánchez et al. claim that our “reidentification figures are probably overestimated” because our data set contains “only a fraction of the population of an undisclosed country.” This means that Sánchez et al. decide, when estimating the risk of an individual to be reidentified, that an attacker could never know whether the person he is searching for is in the data set—e.g., is a client of a specific carrier—or not. As we pointed out before, this arbitrary assumption artificially lowers the estimated, and thus perceived and potentially legally sanctioned, risk of reidentification without changing at all the actual risk for people in the released data set (5). Its reliance on such unscientific assumptions when protecting individuals’ privacy is precisely why we, and others (5, 10, 11), have argued that deidentification is not a useful basis for policy.

To conclude, Sánchez et al.’s Comment arises from a fundamental misunderstanding of the size and dimensionality of modern big-data data sets and how they are being used. The textbook analysis they present does not prove, or even suggest, that high-dimensional data sets, such as the ones generated from credit cards, mobile phones, browsers, or the Internet of things, can be effectively anonymized and of general use. We currently have no reason to believe that an efficient enough, yet general, anonymization method will ever exist for high-dimensional data, as all the evidence so far points to the contrary.

REFERENCES

25 December 2015; accepted 26 January 2016
10.1126/science.aaf1578
Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"
Yves-Alexandre de Montjoye and Alex "Sandy" Pentland

Science 351 (6279), 1274.
DOI: 10.1126/science.aaf1578