Robust de-anonymization of large sparse datasets: a decade later

Arvind Narayanan Vitaly Shmatikov

May 21, 2019

We are grateful to be honored with a Test of Time award for our 2008 paper *Robust De-anonymization of Large Sparse Datasets* [17]. Here we reflect on some lessons from the last decade of de-anonymization research.

An old idea

If our paper has stood the test of time, it is because the core technical insight goes back at least 60 years: a small number of data points about an individual, none of which are uniquely identifying, are collectively equivalent to an identifier. Some of the early papers that explored this idea are by Newcombe et al. [20], Fellegi & Sunter [10], and Schlörer [22].

But the focus of this line of research, up to and including Sweeney's seminal work [25], had been on demographic attributes, because that's what was collected in databases at the time. Perhaps the main lesson of our paper is that data collection has grown so comprehensive that de-anonymization need no longer rely on demographic attributes. Techniques for protecting against de-anonymization such as making a few attributes more coarse-grained break down for datasets of watched movies or browsing histories or visited locations when these datasets contain hundreds or thousands of observations per individual.

We and other researchers have since demonstrated robust de-anonymization techniques in many other domains: social networks [18, 16], genetic data [12, 9, 8], location data [11, 27, 5], credit card data [6], browsing histories [23], writing style [15], source code [3], and compiled binaries [2]. This line of research has firmly established that high-dimensional data is inherently vulnerable to de-anonymization. This is also supported by theoretical evidence [4]. When we consider the fact that 33 bits of entropy are sufficient to identify an individual uniquely among the world's population, these research findings should be no surprise.

Impact on research

Attacks only get better with time. The flood of de-anonymization demonstrations in the last decade makes for a strong argument that database privacy should rest on provable guarantees rather than the absence of known attacks. The flourishing research on differential privacy is thus a welcome development.

The truism that attacks get better applies not only to de-identification, but also to other data protection mechanisms such as searchable database encryption and privacy-preserving machine learning. Schemes without provable guarantees in these areas were broken, too, by drawing on the auxiliary information from public databases [19, 1, 14]. The awesome power of auxiliary data

— especially public databases that yield large numbers of individual records, as well as statistical distributions of the values for just about any attribute of interest — is often underestimated by the designers of privacy and confidentiality mechanisms.

Impact on policy

Traditionally, removal of "personally identifiable information" was seen as the dividing line between sensitive and non-sensitive data. De-anonymization research eroded the value of this distinction, necessitating new regulatory paradigms [21].

This insight has gradually been gaining ground in the tech policy realm, but has often faced fierce opposition. One reason is the wish to preserve the simplicity of the de-identification paradigm: the (false) comfort of applying syntactic modifications to data and not having to think about privacy afterward, regardless of how the data is used.

If there is value in de-identification, it is to keep honest people honest. For example, it can be useful as an internal control mechanism — in combination with proper access controls, auditing and other defenses — to minimize the temptation of employees to peek at the records of specific individuals.

Today's privacy regulations, including the GDPR, continue to put substantial weight on de-identification. Our key recommendation is that the burden of proof be on the data controller to affirmatively show that anonymized data cannot be linked to individuals, rather than on privacy advocates to show that linkage is possible.

Impact on companies

There has also been resistance from companies to the idea that de-identification is insufficient to protect privacy. To understand why, it is helpful to consider what happened when companies did adopt differential privacy. Apple's implementation was found to use an "epsilon" of 16 per day [26], resulting in a privacy risk over a million times worse than recommended by scholars[1]. This is arguably "differential privacy theater": a use of differential privacy that gains Apple public-relations benefits by leading to press articles touting Apple's leadership and innovation on privacy, but without actually providing meaningful guarantees.

The processes by which privacy features lead to user trust in products — which is the commodity that companies should ultimately care about — are only very loosely connected to the technical efficacy of those features. Consider that after decades of user education, we have not had much success in teaching mainstream users the difference between the HTTPS lock icon and similar-looking lock icons in the contents of web pages. For better or (mostly) for worse, the claim "Your data is anonymized" makes users *feel* safe. Thus, one major benefit of de-anonymization is that it, too, is relatively easily understood by the end users, helping disabuse them of the illusion that their data is protected.

If we want sophisticated privacy technologies to be adopted, we need to work on the socio-technical infrastructures that minimize the gap between privacy guarantees and perception of privacy. Those infrastructures are sorely lacking today.

[1] The privacy risk is proportional to e^{ϵ}, and $\epsilon \approx 1$ is typically considered acceptable.

Looking forward: new privacy debates

We live in a world of massive aggregations of personal data. This leads to many privacy risks, of which de-anonymization is just one. Paul Ohm warned of the "database of ruin", a single, massive database containing secrets about every individual, formed by linking different companies' data stores [21]. Today there is a booming market for these linkages between different companies' data stores. For the most part, this is not based on sophisticated de-anonymization techniques. Linkages are possible despite assurances of anonymization because sensitive data being collected and sold is often tied to oxymoronic "anonymous identifiers" such as hashed phone numbers or email addresses.

Furthermore, over the last decade, we have come to realize the privacy implications of the ability to infer sensitive facts about people from seemingly innocuous data, such as pregnancy from shopping records [7] or psychometric traits from Facebook likes [13]. We have also come to better recognize that these aggregations of data may result in harms to society and democracy, rather than just to individuals, as illustrated by Cambridge Analytica's activities. In contexts such as behavioral advertising, scholars argue that the power to influence behavior is deeply problematic even if the data is never linked to a real-world identity [24].

These are great motivations and provocations for research. Computer science research on privacy is important — from developing privacy-preserving data analysis techniques to reverse-engineering surreptitious data collection in our browsers and devices — but real impact will require technologists to engage deeply with policy makers and privacy advocates.

References

[1] V. Bindschaedler, P. Grubbs, D. Cash, T. Ristenpart, and V. Shmatikov, "The tao of inference in privacy-protected databases," *Proceedings of the VLDB Endowment*, vol. 11, no. 11, pp. 1715–1728, 2018.

[2] A. Caliskan, F. Yamaguchi, E. Dauber, R. Harang, K. Rieck, R. Greenstadt, and A. Narayanan, "When coding style survives compilation: De-anonymizing programmers from executable binaries," *arXiv:1512.08546*, 2015.

[3] A. Caliskan-Islam, R. Harang, A. Liu, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt, "De-anonymizing programmers via code stylometry," in *24th USENIX Security Symposium*, 2015, pp. 255–270.

[4] A. Datta, D. Sharma, and A. Sinha, "Provable de-anonymization of large datasets with sparse dimensions," in *International Conference on Principles of Security and Trust.* Springer, 2012, pp. 229–248.

[5] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, 2013.

[6] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, *et al.*, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[7] C. Duhigg, "How companies learn your secrets," The New York Times, Feb 16, 2012.

[8] P. M. Ellenbogen and A. Narayanan, "Identification of anonymous DNA using genealogical triangulation," *bioRxiv 531269*, 2019.

[9] Y. Erlich, T. Shor, I. Peer, and S. Carmi, "Identity inference of genomic data using long-range familial searches," *Science*, vol. 362, no. 6415, pp. 690–694, 2018.

[10] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[11] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *International Conference on Pervasive Computing.* Springer, 2009, pp. 390–397.

[12] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[13] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[14] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *40th IEEE Symposium on Security and Privacy*, 2019.

[15] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the feasibility of Internet-scale author identification," in *33rd IEEE Symposium on Security and Privacy*, 2012, pp. 300–314.

[16] A. Narayanan, E. Shi, and B. I. Rubinstein, "Link prediction by de-anonymization: How we won the Kaggle social network challenge," in *The 2011 International Joint Conference on Neural Networks.* IEEE, 2011, pp. 1825–1834.

[17] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *29th IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[18] ——, "De-anonymizing social networks," in *30th IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[19] M. Naveed, S. Kamara, and C. V. Wright, "Inference attacks on property-preserving encrypted databases," in *22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 644–655.

[20] H. B. Newcombe, J. M. Kennedy, S. Axford, and A. P. James, "Automatic linkage of vital records," *Science*, vol. 130, no. 3381, pp. 954–959, 1959.

[21] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Review*, vol. 57, p. 1701, 2009.

[22] J. Schlörer, "Identification and retrieval of personal records from a statistical data bank," *Methods of Information in Medicine*, vol. 14, no. 01, pp. 7–13, 1975.

[23] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *26th International Conference on World Wide Web*, 2017, pp. 1261–1269.

[24] D. Susser, B. Roessler, and H. Nissenbaum, “Online manipulation: Hidden influences in a digital world,” *SSRN 3306006*, 2018.

[25] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[26] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, “Privacy loss in Apple’s implementation of differential privacy on MacOS 10.12,” *arXiv:1709.02753*, 2017.

[27] H. Zang and J. Bolot, “Anonymization of location data does not work: A large-scale measurement study,” in *17th Annual International Conference on Mobile Computing and Networking.* ACM, 2011, pp. 145–156.