66 STAN. L. REV. ONLINE 103
September 3, 2013

PUBLIC VS. NONPUBLIC DATA: THE BENEFITS OF ADMINISTRATIVE CONTROLS

Yianni Lagos & Jules Polonetsky*

This Essay attempts to frame the conversation around de-identification. De-identification is a process used to prevent a person's identity from being connected with information. Organizations de-identify data for a range of reasons. Companies may have promised "anonymity" to individuals before collecting their personal information, data protection laws may restrict the sharing of personal data, and, perhaps most importantly, companies de-identify data to mitigate privacy threats from improper internal access or from an external data breach. Hackers and dishonest employees occasionally uncover and publicly disclose the confidential information of individuals. Such disclosures could prove disastrous, as public dissemination of stigmatizing or embarrassing information, such as a medical condition, could negatively affect an individual's employment, family life, and general reputation. Given these negative consequences, industries and regulators often rely on de-identification to reduce the occurrence and harm of data breaches.

Regulators have justifiably concluded that strong de-identification techniques are needed to protect privacy before publicly releasing sensitive information. With publicly released datasets, experts agree that weak technical de-identification creates an unacceptably high risk to privacy.[1] For example, statisticians have re-identified some individuals in publicly released datasets.

* Yianni Lagos is Legal and Policy Fellow, Future of Privacy Forum. Jules Polonetsky is Director and Co-Chair, Future of Privacy Forum. Contributions also made by Joe Jerome, Legal and Policy Fellow at the Future of Privacy Forum and Julian Flamant, Policy Fellow at the Future of Privacy Forum.

1. *See* Daniel C. Barth-Jones, The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risk and Privacy Protections, Then and Now 5 (July 24, 2012) (unpublished working paper), *available at* https://www.privacyassociation.org/media/pdf/knowledge_center/Re-Identification_of_Welds_Medical_Information.pdf (finding that 29% of individuals examined had a plausible

None of these publicized attacks, however, have occurred using nonpublic databases. Experts also agree that organizations reduce privacy risk by restricting access to a de-identified dataset to only trusted parties.[2] This Essay builds on this consensus to conclude that de-identification standards should vary depending on whether the dataset is released publicly or kept confidential.

This Essay first describes only technical de-identification (DeID-T) and how policymakers have recognized the benefits of de-identifying data before publicly disclosing a dataset. Second, this Essay discusses how administrative safeguards provide an additional layer of protection to DeID-T that reduces the risk of a data breach. Third, this Essay analyzes the use of de-identification in conjunction with administrative safeguards (DeID-AT). DeID-AT minimizes privacy risks to individuals when compared to using DeID-T or administrative safeguards in isolation. Fourth, this Essay discusses how the different privacy risk profiles between DeID-AT and DeID-T may justify using a reasonably good de-identification standard—as opposed to extremely strict de-identification measures—for non-publicly disclosed databases.

I. TECHNICAL DE-IDENTIFICATION (DEID-T)

DeID-T is a process through which organizations remove or obscure links between an individual's identity and the individual's personal information. This process involves deleting or masking personal identifiers, such as names and social security numbers, and suppressing or generalizing quasi-identifiers, such as dates of birth and zip codes. By using technical de-identification, organizations can transform sensitive information from being fully individually identifiable to being unconnected to any particular person. With publicly disclosed datasets, DeID-T provides the sole line of defense protecting individual privacy.

Policymakers have recognized the benefits of DeID-T by providing regulatory inducements to companies that de-identify publicly disclosed databases. For example, if a company adequately anonymizes a dataset under the 1995 E.U. Data Protection Directive (E.U. Directive), de-identification allows for public disclosure of data without violating individual privacy.[3] Following the E.U. Directive and the U.K. Data Protection Act, the United Kingdom's Information Commissioner's Office (ICO) expressed support for de-identification: "[T]he effective anonymization of personal data is possible, desirable and can help society to make rich data resources available whilst protecting individuals' privacy."[4] The U.S. Department of Health and Human Services (HHS) similar-

---

risk of re-identification with full data of birth, gender, and five-digit ZIP code, though actual risk was much lower given incomplete data).

2. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1771 (2010).

3. Council Directive 95/46, 1995 O.J. (L 281) 26 (EC).

4. INFO. COMM'R'S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 7 (2012).

ly recognized the benefits of de-identifying health data: "The process of de-identification, by which identifiers are removed from health information, mitigates privacy risks to individuals and thereby supports the secondary use of data . . . ."[5]

There are, however, limits to the protections provided by DeID-T. Two different threat models create a risk of re-identification—i.e., reconnecting an individual with what is usually called "personal data" in the European Union and "personally identifiable information" (PII) in the United States.[6] First, outsiders can potentially re-identify an individual by comparing quasi-identifiers in a de-identified database with an identified database, such as a voter registration list. Outsider attacks can come from bad actors or academics, attempting to exploit or show weaknesses in DeID-T protections. In fact, the highest profile re-identification attacks have come from academics attempting to re-identify individuals in publicly disclosed databases.[7] Second, insiders can potentially re-identify an individual by using knowledge that is not generally known. For instance, a Facebook friend, acquaintance, or "skillful Googler" might exploit information that only a limited set of people know, such as a Facebook post mentioning a hospital visit.[8] Similarly, an employee might be able to search through other information held by the organization to re-identify a person.

The threats posed by outsiders, and insiders with restricted access to information, vary significantly depending on whether the de-identified data is publicly disclosed or kept confidential within an organization. When organizations publicly disclose a dataset, every academic, bad actor, and friend can attempt to re-identify the data with DeID-T providing the sole protection. When organizations keep datasets confidential, in contrast, the risk of potential attackers having access to the de-identified data is minimized due to the additional defense of administrative safeguards.

---

5. OFFICE OF CIVIL RIGHTS, U.S. DEP'T OF HEALTH AND HUMAN SERVS., GUIDANCE ON DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION 5 (2012), *available at* http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. HHS is referring to the de-identification provisions found in the Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified at scattered section of the U.S. Code): "*Standard: de-identification of protected health information*. Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information." 45 C.F.R. § 164.514 (2012).

6. Felix T. Wu, *Privacy and Utility in the Data Set*, 84 U. COLO. L. REV. 1117 (2013).

7. C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, SHIDLER J.L. COM. & TECH., Sept. 23, 2008, at 1, *available at* http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/417/vol5_no1_art3.pdf (referring to AOL and Netflix as examples of re-identification attacks).

8. Wu, *supra* note 6, at 28 (quoting Nw. Mem'l Hosp. v. Ashcroft, 362 F.3d 923, 929 (7th Cir. 2004)).

II. ADMINISTRATIVE SAFEGUARDS

This Essay uses the term administrative safeguards to mean all non-technical data protection tools that help prevent confidential data from becoming publicly released or improperly used. In the E.U. Directive, these safeguards are referred to as organizational measures. Non-technical protections include two broad categories: 1) internal administrative and physical controls (internal controls) and 2) external contractual and legal protections (external controls).[9] Internal controls encompass security policies, access limits, employee training, data segregation guidelines, and data deletion practices that aim to stop confidential information from being exploited or leaked to the public. External controls involve contractual terms that restrict how partners use and share information, and the corresponding remedies and auditing rights to ensure compliance.

By implementing administrative safeguards, organizations provide important privacy protections independent of DeID-T. A dentist's office, for instance, does not routinely de-identify patient records to protect a person's privacy, which could negatively impact patient care. Instead, privacy law recognizes that a dental office can hold fully identifiable information if it uses appropriate administrative safeguards, such as performing pre-hire background checks on employees, physically locking drawers with patient records, limiting the information on forms to only needed data, and training employees regarding appropriate access, handling, and disposal of patient files. No privacy breach occurs as long as the confidential patient records do not become disclosed.

The use of administrative safeguards as an additional data protection tool along with DeID-T is consistent with both E.U. and U.S. privacy law. Article 17 of the E.U. Directive requires organizations to "implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access . . . ."[10] The General Data Protection Regulation extends the Directive's existing support for using both technical and organizational measures by incorporating those safeguards into a variety of data protection processes, and by granting the European Commission the power to specify "the criteria and conditions for the technical and organizational measures."[11]

U.S. law similarly requires the use of administrative and technical safeguards. The U.S. Privacy Act of 1974 requires federal agencies to "establish appropriate administrative, technical and physical safeguards to insure the

---

9. This Essay combines the administrative and physical safeguards referred to in the Privacy Act of 1974 into one category: administrative safeguards. Privacy Act of 1974, 5 U.S.C. § 552a(e)(10) (2011).

10. Council Directive 95/46, *supra* note 3, at art. 17(1).

11. *Commission Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, at 56-60, COM (2012) 11 final (Jan. 1, 2012).

security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity."[12] The Gramm-Leach-Bliley Act mandates that financial agencies establish "administrative, technical, and physical safeguards" for financial institutions.[13] Policymakers have thus given value to administrative (or organizational) safeguards as a privacy tool separate from DeID-T that organizations can use to enhance data protection. Similar appreciation for administrative safeguards may be appropriate when applied in the de-identification sphere.

III. ADMINISTRATIVE AND TECHNICAL DE-IDENTIFICATION (DEID-AT)

Organizations who use DeID-AT build a two-tiered barrier that significantly enhances individual privacy protections compared with a single layer. One layer, administrative safeguards, reduces the likelihood of personal data being accessed without authorization. If an insider or outsider does get unauthorized access, another layer, technical de-identification, acts as an additional fortification to minimize potential privacy harms. The two-layered defense provided by DeID-AT means that potential bad actors must not only circumvent administrative measures to gain access to data, but also must re-identify that data before getting any value from their malfeasance. Both are low probability events that together greatly reduce privacy risks. Hence, organizations that implement DeID-AT improve individual privacy.

Policymakers have recognized the distinction between DeID-AT and DeID-T. The ICO drew a line of demarcation between public and nonpublic databases: "We also draw a distinction between publication to the world at large and the disclosure on a more limited basis—for example to a particular research establishment with conditions attached."[14] The Canadian De-Identification Working Group also voiced its belief: "Mitigating controls work in conjunction with de-ID techniques to minimize the re-ID risk."[15] These statements appear to support the proposition that DeID-AT provides a different level of privacy protection than when DeID-T is the sole defensive tool used in publicly disclosed databases.

The heightened privacy protection provided by adding de-identification to administrative safeguards is best demonstrated by using simple statistics. Suppose, for example, the probability of a technical attack on a database gives a one percent chance of re-identification. Suppose as well that the probability of a breach of administrative safeguards is also one percent. (In practice, the likelihood of each is generally much lower.) With both technical and administrative

12. 5 U.S.C. § 552a(e)(10).

13. 15 U.S.C. § 6801(b).

14. INFO. COMM'R'S OFFICE, *supra* note 4, at 7.

15. HEALTH SYS. USE TECHNICAL ADVISORY COMM. DATA DE-IDENTIFICATION WORKING GRP., 'BEST PRACTICE' GUIDELINES FOR MANAGING THE DISCLOSURE OF DE-IDENTIFIED HEALTH INFORMATION 19 (2010).

protections, the probability of re-identifying data is thus one percent of one percent, or one in 10,000.[16] This simple statistical example shows that the risk of re-identification with DeID-AT may well be orders of magnitude lower than using only technical safeguards in isolation.

IV. POLICY IMPLICATIONS

The additional protections provided by DeID-AT compared with DeID-T suggest a different risk profile that may justify the use of fairly strong technical measures, combined with effective administrative safeguards. The Federal Trade Commission recognized this fact when it called in its 2012 report for technical measures that made a dataset "not reasonably identifiable."[17] The combination of reasonably good technical measures, as well as good administrative measures, likely leads to a lower risk of re-identification than stronger technical measures acting alone. The HIPAA de-identification standard that requires a "very small" risk of re-identification before publicly releasing health data is an example of a relatively strict standard for re-identification, designed for datasets that can be made fully public.[18] A less strict standard, however, achieves a similar or stronger level of protection for non-publicly available databases.

Giving credit to the use of administrative controls also helps prevent an illogical outcome: greater data restrictions for the original collector of the data than downstream recipients or the public. The original collector commonly has more access to data on an individual than it would disclose to another party. A separate nonpublic database containing an individual's name or email address, for example, would normally not be disclosed. That separate database could potentially be used to re-identify an individual, giving the original collector a re-identification advantage over any other party.[19] Thus, if administrative controls do not receive regulatory recognition, the original data collector would be subject to a steeper regulatory burden than potential downstream recipients.

Relying on the data protection benefits of using DeID-AT to justify allowing reasonably strict de-identification comes with a caveat that it can be diffi-

16. The one in 10,000 statistic is based on the assumption that the probability of the technical and administrative attacks are independent of each other. In practice, under a particular attack scenario, this assumption may not hold. By arguing for a different de-identification standard for public and nonpublic data, we do not claim that pseudonymization is sufficient to constitute pretty good de-identification. Other factors, such as whether companies maintain the cryptographic key when transforming identifiers, will determine the effectiveness of pseudonymization. It is clear, however, that if a company can easily re-identify every individual from a pseudonymous database, the statistical benefits of combining administrative measures with technical measures are lost.

17. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 22 (2012).

18. OFFICE OF CIVIL RIGHTS, *supra* note 5, at 6.

19. KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION 142 (2013).

cult to assess the efficacy of administrative safeguards. Privacy advocates and academics can test DeID-T used in public data releases. In fact, improvements in DeID-T can result from privacy advocates and academics testing claims of anonymization. Companies, however, keep administrative safeguards proprietary for security purposes, and privacy advocates cannot audit non-transparent privacy protections. The use of third-party auditors is one approach for ensuring that administrative safeguards effectively prevent privacy attacks, but without a certain level of public transparency of such measures, regulators and privacy advocates may find it difficult to assess the exact benefits of administrative safeguards.

CONCLUSION

Non-publicly disclosed datasets have a lessened risk of re-identification than publicly disclosed datasets due to the added protection of administrative controls. The different risk profiles suggest requiring different measures of de-identification for publicly disclosed datasets compared with confidential datasets. This Essay urges regulators to recognize the heightened individual privacy protections provided by DeID-AT compared with DeID-T when developing privacy regulations.