WIKIPEDIA

Data re-identification

Data **re-identification** or **de-anonymization** is the practice of matching anonymous data (also known as de-identified data) with publicly available information, or auxiliary data, in order to discover the person to whom the data belongs.[1] This is a concern because companies with privacy policies, health care providers, and financial institutions may release the data they collect after the data has gone through the de-identification process.

The de-identification process involves masking, generalizing or deleting both direct and indirect identifiers; the definition of this process is not universal. Information in the public domain, even seemingly anonymized, may thus be re-identified in combination with other pieces of available data and basic computer science techniques. The Protection of Human Subjects ('Common Rule'), a collection of multiple U.S. federal agencies and departments including the U.S. Department of Health and Human Services, warn that re-identification is becoming gradually easier because of "big data"—the abundance and constant collection and analysis of information along with the evolution of technologies and the advances of algorithms. However, others have claimed that de-identification is a safe and effective data liberation tool and do not view re-identification as a concern.[2]

More and more data are becoming publicly available over the Internet. These data are released after applying some anonymization techniques like removing personally identifiable information (PII) such as names, addresses and social security numbers to ensure the sources' privacy. This assurance of privacy allows the government to legally share limited data sets with third parties without requiring written permission. Such data has proved to be very valuable for researchers, particularly in health care.

GDPR-compliant pseudonymization seeks to reduce the risk of re-identification through the use of separately kept "additional information". The approach is based on an expert evaluation of a dataset to designate some identifiers as "direct" and some as "indirect." Proponents of this approach argue that re-identification can be avoided by limiting access to "additional information" that is kept separately by the controller. The theory is that access to separately kept "additional information" is required for re-identification, attribution of data to a specific data subject can be limited by the controller to support lawful purposes only. This approach is controversial, as it fails if there are additional datasets that can be used for re-identification. Such additional datasets may be unknown to those certifying the GDPR-compliant pseudonymization, or may not at exist at the time of the pseudonymization but may come into existence at some point in the future.

Legal protections of data in the United States

Existing privacy regulations typically protect information that has been modified, so that the data is deemed anonymized, or de-identified. For financial information, the Federal Trade Commission permits its circulation if it is de-identified and aggregated.[3] The Gramm Leach Bliley Act (GLBA), which mandates financial institutions give consumers the opportunity to opt out of having their information shared with third parties, does not cover de-identified data if the information is aggregate and does not contain personal identifiers, since this data is not treated as personally identifiable information.[3]

Educational records

In terms of university records, authorities both on the state and federal level have shown an awareness

about issues of privacy in education and a distaste for institutions' disclosure of information. The U.S. Department of Education has provided guidance about data discourse and identification, instructing educational institutions to be sensitive to the risk of re-identification of anonymous data by cross-referencing with auxiliary data, to minimize the amount of data in the public domain by decreasing publication of directory information about students and institutional personnel, and to be consistent in the processes of de-identification.[4]

Medical records

Medical information of patients are becoming increasingly available on the Internet, on free and publicly accessing platforms such as HealthData.gov and PatientsLikeMe, encouraged by government open data policies and data sharing initiatives spearheaded by the private sector. While this level of accessibility yields many benefits, concerns regarding discrimination and privacy have been raised.[5] Protections on medical records and consumer data from pharmacies are stronger compared to those for other kinds of consumer data. The Health Insurance Portability and Accountability Act (HIPAA) protects the privacy of identifiable data about health, but authorize information release to third parties if de-identified. In addition, it mandates that patients receive breach notifications should there be more than a low probability that the patient's information was inappropriately disclosed or utilized without sufficient mitigation of the harm to him or her.[6] The likelihood of re-identification is a factor in determining the probability that the patient's information has been compromised. Commonly, pharmacies sell de-identified information to data mining companies that sell to pharmaceutical companies in turn.[3]

There have been state laws enacted to ban data mining of medical information, but they were struck down by federal courts in Maine and New Hampshire on First Amendment grounds. Another federal court on another case used "illusive" to describe concerns about privacy of patients and did not recognize the risks of re-identification.[3]

Biospecimen

The Notice of Proposed Rule Making, published by the Common Rule Agencies in September 2015, expanded the umbrella term of "human subject" in research to include biospecimens, or materials taken from the human body - blood, urine, tissue etc. This mandates that researchers using biospecimens must follow the stricter requirements of doing research with human subjects. The rationale for this is the increased risk of re-identification of biospecimen.[7] The final revisions affirmed this regulation.[8]

Re-identification efforts

There have been a sizable amount of successful attempts of re-identification in different fields. Even if it is not easy for a lay person to break anonymity, once the steps to do so are disclosed and learnt, there is no need for higher level knowledge to access information in a database. Sometimes, technical expertise is not even needed if a population has a unique combination of identifiers.[3]

Health records

In the mid-1990s, a government agency in Massachusetts called Group Insurance Commission (GIC), which purchased health insurance for employees of the state, decided to release records of hospital visits to any researcher who requested the data, at no cost. GIC assured that the patient's privacy was not a concern since it had removed identifiers such as name, addresses, social security numbers. However, information such as zip codes, birth date and sex remained untouched. The GIC assurance was reinforced by the then governor of Massachusetts, William Weld. Latanya Sweeney, a graduate student at the time, put her mind to picking out the governor's records in the GIC data. By combining the GIC data with the voter database of

the city Cambridge, which she purchased for 20 dollars, Governor Weld's record was discovered with ease.[9]

In 1997, a researcher successfully de-anonymized medical records using voter databases.[3]

In 2011, Professor Latanya Sweeney again used anonymized hospital visit records and voting records in the state of Washington and successfully matched individual persons 43% of the time.[10]

There are existing algorithms used to re-identify patient with prescription drug information.[3]

Consumer habits and practices

Two researchers at the University of Texas, Arvind Narayanan and Professor Vitaly Shmatikov, were able to re-identify some portion of anonymized Netflix movie-ranking data with individual consumers on the streaming website.[11][12][13] The data was released by Netflix 2006 after de-identification, which consisted of replacing individual names with random numbers and moving around personal details. The two researchers de-anonymized some of the data by comparing it with non-anonymous IMDb (Internet Movie Database) users' movie ratings. Very little information from the database, it was found, was needed to identify the subscriber.[3] In the resulting research paper, there were startling revelations of how easy it is to re-identify Netflix users. For example, simply knowing data about only two movies a user has reviewed, including the precise rating and the date of rating give or take three days allows for 68% re-identification success.[9]

In 2006, after AOL published its users' search queries, data that was anonymized prior to the public release, *The New York Times* reporters successfully carried out re-identification of individuals by taking groups of searches made by anonymized users.[3] AOL had attempted to suppress identifying information, including usernames and IP addresses, but had replaced these with unique identification numbers to preserve the utility of this data for researchers. Bloggers, after the release, pored over the data, either trying to identify specific users with this content, or to point out entertaining, depressing, or shocking search queries, examples of which include "how to kill you wife", "depression and medical leave", "car crash photos." Two reporters, Michael Barbaro and Tom Zeller, were able to track down a 62 year old widow named Thelma Arnold from recognizing clues to the identity of User 417729 search histories. Arnold acknowledged that she was the author of the searches, confirming that re-identification is possible.[9]

Location data

Location data - series of geographical positions in time that describe a person's whereabouts and movements - is a class of personal data that is specifically hard to keep anonymous. Location shows recurring visits to frequently attended places of everyday life such as home, workplace, shopping, healthcare or specific sparetime patterns.[14] Only removing a person's identity from location data will not remove identifiable patterns such as commuting rhythms, sleeping places, or work places. By mapping coordinates onto addresses, location data is easily re-identified[15] or correlated with a person's private life contexts. Streams of location information play an important role in the reconstruction of personal identifiers from smartphone data accessed by apps.[16]

Court decisions

In 2019, Professor Kerstin Noëlle Vokinger (https://www.ius.uzh.ch/en/staff/professorships/alphabetical/vokinger/vokinger.html) and Dr. Urs Jakob Mühlematter, two researchers at the University of Zurich, analyzed cases of the Federal Supreme Court of Switzerland to assess which pharmaceutical companies and which medical drugs were involved in legal actions against the Federal Office of Public Health (FOPH) regarding pricing decisions of medical drugs. In general, involved private parties (such as pharmaceutical

companies) and information that would reveal the private party (for example, drug names) are anonymized in Swiss judgments. The researchers were able to re-identify 84% of the relevant anonymized cases of the Federal Supreme Court of Switzerland by linking information from publicly accessible databases.[17][18] This achievement was covered by the media and started a debate if and how court cases should be anonymized.[19][20]

Concern and consequences

In 1997, Latanya Sweeney found from a study of Census records that up to 87 percent of the U.S. population can be identified using a combination of their 5-digit zip code, gender, and date of birth.[21][22]

Unauthorized re-identification on the basis of such combinations does not require access to separately kept "additional information" that is under the control of the data controller, as is now required for GDPR-compliant pseudonymization.

Individuals whose data is re-identified are also at risk of having their information, with their identity attached to it, sold to organizations they do not want possessing private information about their finances, health or preferences. The release of this data may cause anxiety, shame or embarrassment. Once an individual's privacy has been breached as a result of re-identification, future breaches become much easier: once a link is made between one piece of data and a person's real identity, any association between the data and an anonymous identity breaks the anonymity of the person.[3]

Re-identification may expose companies and institutions which have pledged to assure anonymity to increased tort liability and cause them to violate their internal policies, public privacy policies, and state and federal laws, such as laws concerning financial confidentiality or medical privacy, by having released information to third parties that can identify users after re-identification.[3]

Remedies

To address the risks of re-identification, several proposals have been suggested:

- Higher standards and uniform definition of de-identification while retaining data utility: the definition of de-identification should balance privacy protections to reduce re-identification risk with the refusal of companies to delete data[23]
- Heightened privacy protections of anonymized information[3]
- Tighter security for databases that store anonymized information[3]
- Strong ban on malicious re-identification, the passing of broader anti-discrimination and privacy legislation that ensures privacy protections as well as encourage participation in data sharing projects and endeavors, as well as establishment of uniform data protection standards in academic communities, such as in the scientific community, in order to minimize privacy violations[24]
- Creation of data-release policies: making sure de-identification rhetoric is accurate, drawing up contracts that prohibit re-identification attempts and dissemination of sensitive information, establishing data enclaves, and utilizing data-based strategies to match required protection standards to the level of risk.[25]
- Implementation of Differential Privacy on requested data sets
- Generation of Synthetic Data that exhibits the statistical properties of the raw data, without allowing real individuals to be identified

While a complete ban on re-identification has been urged, enforcement would be difficult. There are, however, ways for lawmakers to combat and punish re-identification efforts, if and when they are exposed: pair a ban with harsher penalties and stronger enforcement by the Federal Trade Commission and the Federal Bureau of Investigation; grant victims of re-identification a right of action against those who re-identify them; and mandate software audit trails for people who utilize and analyze anonymized data. A small-scale re-identification ban may also be imposed on trusted recipients of particular databases, such as government data miners or researchers. This ban would be much easier to enforce and may discourage re-identification.[9]

Examples of de-anonymization

- "Researchers at MIT and the Université catholique de Louvain, in Belgium, analyzed data on 1.5 million cellphone users in a small European country over a span of 15 months and found that just four points of reference, with fairly low spatial and temporal resolution, was enough to uniquely identify 95 percent of them. In other words, to extract the complete location information for a single person from an "anonymized" data set of more than a million people, all you would need to do is place him or her within a couple of hundred yards of a cellphone transmitter, sometime over the course of an hour, four times in one year. A few Twitter posts would probably provide all the information you needed, if they contained specific information about the person's whereabouts."[26]
- "Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target."[27]

See also

- De-identification – Preventing personal identity from being revealed
- doxing – Publication of the private details of individuals, often on the Internet
- K-anonymity – Property of certain anonymized data
- Protected health information – Information about healthcare status of individual
- Statistical disclosure control – Technique used in data-driven research

References

1. Pedersen, Torben (2005). "HTTPS, Secure HTTPS". *Encyclopedia of Cryptography and Security*. pp. 268–269. doi:10.1007/0-387-23483-7_189 (https://doi.org/10.1007%2F0-387-23483-7_189). ISBN 978-0-387-23473-1.
2. Richardson, Victor; Milam, Sallie; Chrysler, Denise (April 2015). "Is Sharing De-Identified Data Legal? The State of Public Health Confidentiality Laws and Their Interplay with Statistical Disclosure Limitation Techniques". *The Journal of Law, Medicine & Ethics*. **43** (1_suppl): 83–86. doi:10.1111/jlme.12224 (https://doi.org/10.1111%2Fjlme.12224). hdl:2027.42/111074AA (https://hdl.handle.net/2027.42%2F111074AA). ISSN 1073-1105 (https://search.worldcat.org/issn/1073-1105). PMID 25846173 (https://pubmed.ncbi.nlm.nih.gov/25846173). S2CID 9384220 (https://api.semanticscholar.org/CorpusID:9384220).

3. Porter, Christine (2008). "Constitutional and Regulatory: De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information" (https://digitalcommons.law.uw.edu/wjlta/vol5/iss1/3/). *Shidler Journal of Law, Commerce & Technology*. **5** (1).
4. Peltz, Richard (2009). "From the Ivory Tower to the Glass House: Access to "De-Identified" Public University Admission Records to Study Affirmative Action" (https://harvardblackletter.org/wp-content/uploads/sites/8/2012/11/181-198.pdf) (PDF). *Harvard BlackLetter Law Journal*. **25**: 181–197. SSRN 1495788 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495788).
5. Hoffman, Sharona (2015). "Citizen Science: The Law and Ethics of Public Access to Medical Big Data". *Berkeley Technology Law Journal*. doi:10.15779/Z385Z78 (https://doi.org/10.15779%2FZ385Z78).
6. Greenberg, Yelena (2016). "Recent Case Developments: Increasing Recognition of "Risk of Harm" as an Injury Sufficient to Warrant Standing in Class Action Medical Data Breach Cases". *American Journal of Law & Medicine*. **42** (1): 210–4. doi:10.1177/0098858816644723 (https://doi.org/10.1177%2F0098858816644723). PMID 27263268 (https://pubmed.ncbi.nlm.nih.gov/27263268). S2CID 77790820 (https://api.semanticscholar.org/CorpusID:77790820).
7. Groden, Samantha; Martin, Summer; Merrill, Rebecca (2016). "Proposed Changes to the Common Rule: A Standoff Between Patient Rights and Scientific Advances?" (https://www.americanhealthlaw.org/content-library/journal-health-law/article/cba33b33-8c05-4e73-b6cf-e3238f532176/Proposed-Changes-to-the-Common-Rule-A-Standoff-Bet). *Journal of Health & Life Sciences Law*. **9** (3).
8. 24 C.F.R. § .104 2017.
9. Ohm, Paul (August 2010). "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=53753152&site=eds-live&scope=site). *UCLA Law Review*. **57** (6): 1701–1777. ISSN 0041-5650 (https://search.worldcat.org/issn/0041-5650). OCLC 670569859 (https://search.worldcat.org/oclc/670569859) – via EBSCO.
10. Sweeney, Latanya (28 September 2015). "Only You, Your Doctor, and Many Others May Know" (https://techscience.org/a/2015092903/). *Technology Science*. 2015092903. Retrieved 12 July 2024.
11. Rouse, Margaret. "de-anonymization (deanonymization)" (http://whatis.techtarget.com/definition/de-anonymization-deanonymization). WhatIs.com. Retrieved 19 January 2014.
12. Narayanan, Arvind; Shmatikov, Vitaly. "Robust De-anonymization of Large Sparse Datasets" (http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf) (PDF). Retrieved 19 January 2014.
13. Narayanan, Arvind; Shmatikov, Vitaly (22 November 2007). "How To Break Anonymity of the Netflix Prize Dataset". arXiv:cs/0610105 (https://arxiv.org/abs/cs/0610105).
14. Fritsch, Lothar (2008), "Profiling and Location-Based Services (LBS)", *Profiling the European Citizen*, Springer Netherlands, pp. 147–168, doi:10.1007/978-1-4020-6914-7_8 (https://doi.org/10.1007%2F978-1-4020-6914-7_8), ISBN 978-1-4020-6913-0
15. Rocher, Luc; Hendrickx, Julien M.; de Montjoye, Yves-Alexandre (23 July 2019). "Estimating the success of re-identifications in incomplete datasets using generative models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6650473). *Nature Communications*. **10** (1): 3069. Bibcode:2019NatCo..10.3069R (https://ui.adsabs.harvard.edu/abs/2019NatCo..10.3069R). doi:10.1038/s41467-019-10933-3 (https://doi.org/10.1038%2Fs41467-019-10933-3). ISSN 2041-1723 (https://search.worldcat.org/issn/2041-1723). PMC 6650473 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6650473). PMID 31337762 (https://pubmed.ncbi.nlm.nih.gov/31337762).

16. Fritsch, Lothar; Momen, Nurul (2017). *Derived Partial Identities Generated from App Permissions* (http://dl.gi.de/handle/20.500.12116/3585). Gesellschaft für Informatik, Bonn. ISBN 978-3-88579-671-8.
17. Vokinger / Mühlematter, Kerstin Noëlle / Urs Jakob (2 September 2019). "Identifikation von Gerichtsurteilen durch "Linkage" von Daten(banken)" (https://jusletter.weblaw.ch/juslissues/2019/990/re-identifikation-vo_21cb82c096.html__ONCE&login=false). *Jusletter* (990).
18. Vokinger / Mühlematter, Kerstin Noëlle / Urs Jacob. "Re-Identifikation von Gerichtsurteilen durch "Linkage" von Daten(banken)" (https://www.researchgate.net/publication/335543645).
19. Chandler, Simon (4 September 2019). "Researchers Use Big Data And AI To Remove Legal Confidentiality" (https://www.forbes.com/sites/simonchandler/2019/09/04/researchers-use-big-data-and-ai-to-remove-legal-confidentiality/). *Forbes*. Retrieved 10 December 2019.
20. "SRF Tagesschau" (https://www.youtube.com/watch?v=2FYbF1-VQUQ). SRF Swiss Radio and Television. 2 September 2019. Retrieved 10 December 2019.
21. "How Unique am I?" (https://aboutmyinfo.org/identity/about). Data Privacy Lab, Harvard University. Retrieved 22 July 2021.
22. Sweeney, Latanya. "Simple Demographics Often Identify People Uniquely" (https://dataprivacylab.org/projects/identifiability/paper1.pdf) (PDF). *Carnegie Mellon University, Data Privacy Working Paper 3*. Retrieved 22 July 2021.
23. Lagos, Yianni (2014). "Taking the Personal Out of Data: Making Sense of De-identification" (https://mckinneylaw.iu.edu/practice/law-reviews/ilr/pdf/vol48p187.pdf) (PDF). *Indiana Law Review*. **48**: 187–203. ISSN 2169-320X (https://search.worldcat.org/issn/2169-320X). OCLC 56050778 (https://search.worldcat.org/oclc/56050778).
24. Sejin, Ahn (Summer 2015). "Whose Genome Is It Anyway?: Re-Identification and Privacy Protection in Public and Participatory Genomics" (https://digital.sandiego.edu/sdlr/vol52/iss3/7/). *San Diego Law Review*. **52** (3): 751–806. ISSN 2994-9599 (https://search.worldcat.org/issn/2994-9599). OCLC 47865544 (https://search.worldcat.org/oclc/47865544).
25. Rubinstein, Ira S.; Hartzog, Woodrow (June 2016). "Anonymization and Risk" (https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=116583482&site=eds-live&scope=site). *Washington Law Review*. **91** (2): 703–760. ISSN 0043-0617 (https://search.worldcat.org/issn/0043-0617). OCLC 3899779 (https://search.worldcat.org/oclc/3899779) – via EBSCO.
26. Hardesty, Larry (27 March 2013). "How hard is it to 'de-anonymize' cellphone data?" (https://newsoffice.mit.edu/2013/how-hard-it-de-anonymize-cellphone-data). MIT news. Retrieved 14 January 2015.
27. Melissa Gymrek; Amy L. McGuire; David Golan; Eran Halperin; Yaniv Erlich (18 January 2013). "Identifying personal genomes by surname inference". *Science*. **339** (6117): 321–4. Bibcode:2013Sci...339..321G (https://ui.adsabs.harvard.edu/abs/2013Sci...339..321G). doi:10.1126/SCIENCE.1229566 (https://doi.org/10.1126%2FSCIENCE.1229566). ISSN 0036-8075 (https://search.worldcat.org/issn/0036-8075). PMID 23329047 (https://pubmed.ncbi.nlm.nih.gov/23329047). Wikidata Q29619963.
