DRAFT NISTIR 8053

De-Identification of Personally Identifiable Information

Simson L. Garfinkel

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8053
DRAFT

De-Identification of Personally Identifiable Information

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

April 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

National Institute of Standards and Technology Internal Report 8053
vi + 28 pages (April 2015)

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by Federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, Federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST Computer Security Division publications, other than the ones noted above, are available at http://csrc.nist.gov/publications.

Comments on this publication may be submitted to: draft-nistir-deidentify@nist.gov

Public comment period: *April 15, 2015* through *May 15, 2015*

National Institute of Standards and Technology
Attn: Computer Security Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930
Email: draft-nistir-deidentify@nist.gov

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation’s measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL’s responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in Federal information systems.

Abstract

De-identification is the removal of identifying information from data. Several US laws, regulations and policies specify that data should be de-identified prior to sharing as a control to protect the privacy of the data subjects. In recent years researchers have shown that some de-identified data can sometimes be re-identified. This document summarizes roughly two decades of de-identification research, discusses current practices, and presents opportunities for future research.

Keywords

De-identification; HIPAA Privacy Rule; k-anonymity; re-identification; privacy

Acknowledgements

We wish to thank Khaled El Emam, Bradley Malin, Latanya Sweeney and Christine M. Task for answering questions and reviewing earlier versions of this document.

Audience

This document is intended for use by officials, advocacy groups and other members of the community that are concerned with the policy issues involving the creation, use and sharing of data sets containing personally identifiable information. It is also designed to provide technologists and researchers with an overview of the technical issues in the de-identification of data sets. While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience. For this reason, this document minimizes the use of mathematical notation.

Note to Reviewers

NIST requests comments especially on the following:

- Is the terminology that is provided consistent with current usage?
- To what extent should this document’s subject be broadened to discuss differential privacy and statistical disclosure limitation techniques?
- Should the glossary be expanded? If so, please suggest words, definitions, and appropriate citations.

Table of Contents

List of Appendices

1 Introduction

Government agencies, businesses and other organizations are increasingly under pressure to make raw data available to outsiders. When collected data contain personally identifiable information (PII) such as names or Social Security numbers (SSNs), there can be a conflict between the goals of sharing data and protecting privacy. *De-identification* is one way that organizations can balance these competing goals.

De-identification is a process by which a data custodian alters or removes identifying information from a data set, making it harder for users of the data to determine the identities of the data subjects. Once de-identified, data can be shared with trusted parties that are bound by data use agreements that only allow specific uses. In this case, de-identification makes it easier for trusted parties to comply with privacy requirements. Alternatively, the de-identified data can be distributed with fewer controls to a broader audience. In this case, de-identification is a tool designed to assist privacy-preserving data publishing (PPDP).

De-identification is not without risk. There are many de-identification techniques with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting data set. As long as any utility remains in the data, there exists the possibility that some information might be linked back to the original identities, a process called *re-identification*. The use of de-identified data can also result in other harms to the data subjects, even without having the data first re-identified.

1.1 Document Purpose and Scope

This document provides an overview of de-identification issues and terminology. It summarizes significant publications to date involving de-identification and re-identification.

1.2 Intended Audience

This document is intended for use by officials, advocacy groups and other members of the community that are concerned with the policy issues involving the creation, use and sharing of data sets containing personally identifiable information. It is also designed to provide technologists and researchers with an overview of the technical issues in the de-identification of data sets. While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience. For this reason, this document minimizes the use of mathematical notation.

1.3 Organization

The remainder of this report is organized as follows: Section 2 introduces the concepts of de-identification, re-identification and data sharing models. Section 3 discusses *syntactic de-identification*, a class of de-identification techniques that rely on the masking or altering of fields in tabular data. Section 4 discusses current challenges of de-identification information that are not tabular data, such as free-format text, images, and genomic information. Section 5 concludes. Appendix A is a glossary, and Appendix B provides a list of additional resources.

2 De-identification, Re-Identification, and Data Sharing Models

This section explains the motivation for de-identification, discusses the use of re-identification attacks to gauge the effectiveness of de-identification, and describes models for sharing de-identified data. It also introduces the terminology used in this report.

2.1 Motivation

Increasingly organizations that are collecting data and maintaining databases are under challenged to protect the data while using and sharing as widely as possible. For government databases, data sharing can increase transparency, provide new resources to private industry, and lead to more efficient government as a whole. Private firms can also benefit from data sharing in the form of increased publicity, civic engagement, and potentially increased revenue if the data are sold.

When datasets contains personally identifiable information such as names, email addresses, geolocation information, or photographs, there can be a conflict between the goals of effective data use and privacy protection. Many data sharing exercises appear to violate the Fair Information Practice Principles[1] of ***Purpose Specification***[2] and ***Use Limitation***[3]. Retaining a database of personal information after it is no longer needed, because it was expensive to create and the data might be useful in the future, may be a violation of the ***Data Minimization***[4] principle.

De-identification represents an attempt to uphold the privacy promise of the FIPPs while allowing for data re-use, with the justification that the individuals' will not suffer a harm from the use of their data because their identifying information has been removed from the dataset.

Several US laws and regulations specifically recognize the importance and utility of data de-identification:

- The Department of Education has held that the Family and Educational Records Privacy Act does not apply to de-identified student records. "Educational agencies and institutions are permitted to release, without consent, educational records, or information from educational records that have been de-identified through the removal of all personally identifiable information."[5]

[1] National Strategy for Trusted Identities in Cyberspace, Appendix A—Fair Information Practice Principles. April 15, 2011. http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

[2] "**Purpose Specification:** Organizations should specifically articulate the authority that permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used." Ibid.

[3] "**Use Limitation:** Organizations should use PII solely for the purpose(s) specified in the notice. Sharing PII should be for a purpose compatible with the purpose for which the PII was collected." Ibid.

[4] "**Data Minimization:** Organizations should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s)."

[5] Dear Colleague Letter about Family Educational Rights and Privacy Act (FERPA) Final Regulations, US Department of Education, December 17, 2008. http://www2.ed.gov/policy/gen/guid/fpco/hottopics/ht12-17-08.html

- The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule allows de-identified medical records to be used without any restriction, provided that organizations distributing the records have no direct knowledge that the records can be re-identified.[6]
- The Health Information Technology for Economic and Clinical Health Act (HITECH Act) requirements for security and privacy explicitly do not apply to the "use, disclosure, or request of protected health information that has been de-identified."[7]
- The Foodborne illness surveillance system is required to allow "timely public access to aggregated, de-identified surveillance data."[8]
- Entities contracted by Health and Human Services to provide drug safety data must have the ability to provide that data in de-identified form.[9]
- Voluntary safety reports submitted to the Federal Aviation Submission are not protected from public disclosure if the data that they contain is de-identified.[10]

Each of these laws and regulations implicitly assume that it is possible to remove personally identifiable information from a data set in a way that protects privacy but still leaves useful information. They also assume that de-identified information will not be re-identified at a later point in time.

In practice many de-identification techniques are not able to provide such strong privacy guarantees. Section 3.2 and Section 3.5 discuss some of the well-publicized cases in which data that were thought to be properly de-identified were published and then later re-identified by researchers or journalists. The results of these re-identifications violated the privacy of the data subjects, who were not previously identified as being in the dataset. Additional privacy harms can result from the disclosure of specific attributes that the data set linked to the identities.

2.2 Models for Privacy-Preserving use of Private Information

Academics have identified two distinct models for making use of personally identifiable information in a database while protecting the privacy of the data subjects:

- ***Privacy Preserving Data Mining.*** In this model, data are not released, but are used instead for statistical processing or machine learning. The results of the calculations may be released in the form of statistical tables, classifiers, or other kinds of results.

[6] 45 CFR 160, 45 CFR 162, and 45 CFR 164. See also "Combined Regulation Text of All Rules," US Department of Health and Human Services, Office for Civil Rights, Health Information Privacy. http://www.hhs.gov/ocr/privacy/hipaa/administrative/combined/index.html

[7] 42 USC 17935

[8] 21 USC 2224

[9] 21 USC 355

[10] 49 USC 44735

- ***Privacy Preserving Data Publishing.*** In this model, data are processed to produce a new data product that is distributed to users.

Privacy Preserving Data Mining (PPDM) is a broad term for any use of sensitive information to publish public statistics. Statistical reports that summarize confidential survey data are an example of PPDM.

Statistical Disclosure Limitation[11] is a set of principles and techniques that have been developed by researchers concerned with the generation and publication of official statistics. The goal of disclosure limitation is to prevent published statistics from impacting the privacy of those surveyed. Techniques developed for disclosure limitation include generalization of reported information to broader categories, swapping data between similar entities, and the addition of noise in reports.

Differential Privacy is a set of techniques based on a mathematical definition of privacy and information leakage from operations on a data set by the introduction of non-deterministic noise.[12] Differential privacy holds that the results of a data analysis should be roughly the same before and after the addition or removal of a single data record (which is usually taken to be the data from a single individual). In its basic form differential privacy is applied to online query systems, but differential privacy can also be used to produce machine-learning statistical classifiers and synthetic data sets.[13]

Differential privacy is an active research area, but to date there have been few applications of differential privacy techniques to actual running systems. Two notable exceptions are the Census Bureau's "OnTheMap" website, which uses differential privacy to create reasonably accurate block-level synthetic census data;[14] and Fredrikson et al.'s study to determine the impact of applying differential privacy to a clinical trial that created a statistical model for correlating genomic information and warfarin dosage.[15] The Fredrikson study concluded that the models constructed using differential privacy gains came at the cost of would result negative clinical outcomes for a significant number of patients.

Privacy Preserving Data Publishing (PPDP) allows for information based on private data to be published, allowing other researchers to perform novel analyses. The goal of PPDP is to provide

---

[11] Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.

[12] Cynthia Dwork, Differential Privacy, in ICALP, Springer, 2006

[13] Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron, Zhiwei Steven Wu, Dual Query: Practical Private Query Release for High Dimensional Data, Proceedings of the 31st Intenrational Conference on Machine Learning, Beijing, China. 2014. JMLR: W&CP volume 32.

[14] Abowd et al., "Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data," Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, Suitland, MD, July 31, 2009.

[15] Fredrikson et al., Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Wafrin Dosing, 23rd Usenix Security Symposium, August 20-22, 2014, San Diego, CA.

data that have high utility without compromising the privacy of the data subjects.

De-identification is the "general term for any process of removing the association between a set of identifying data and the data subject." (ISO/TS 25237-2008) De-identification is designed to protect individual privacy while preserving some of the dataset's utility for other purposes. De-identification protects the privacy of individuals, making it hard or impossible to learn if an individual's data is in a data set, or to determine any attributes about an individual known to be in the data set. De-identification is one of the primary tools for achieving PPDP.

Synthetic data generation uses some PPDM techniques to create a dataset that is similar to the original data, but where some or all of the resulting data elements are generated and do not map to actual individuals. As such synthetic data generation can be seen as a fusion of PPDM and PPDP.

2.3 De-Identification Data Flow Model

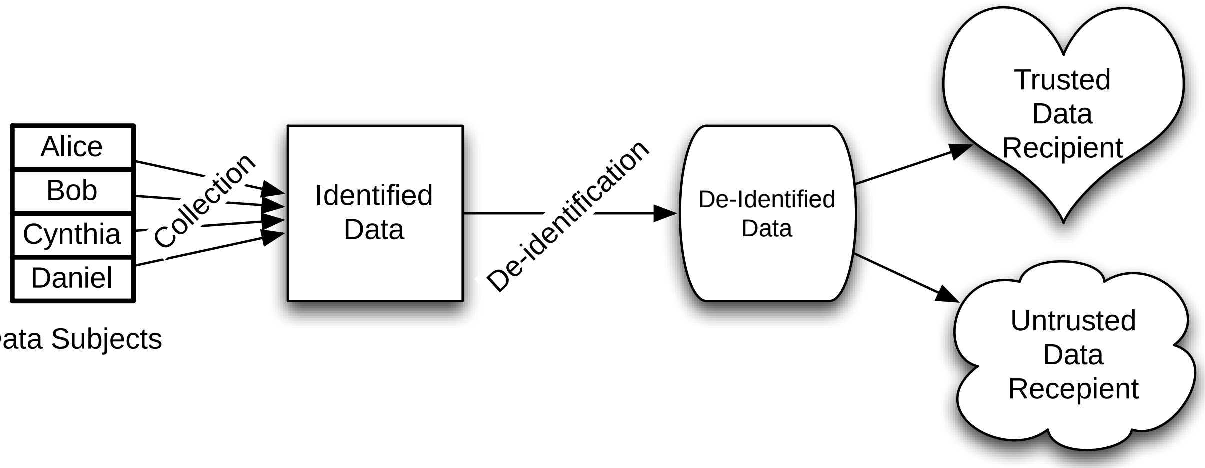

Figure 1: Data Collection, De-Identification and Use

Figure 1 provides an overview of the de-identification process. Data are collected from ***Data Subjects,*** the "persons to whom data refer." (ISO/TS 25237-2008) These data are combined into a ***data set*** containing *personally identifiable information* (PII). De-identification creates a new data set of *de-identified data.* This data set may eventually be used by a small number of trusted data recipients. Alternatively, the data might be made broadly available to a larger (potentially limitless) number of untrusted data recipients.

Pseudonymization is a specific kind of de-identification in which the direct identifiers are replaced with pseudonyms (ISO/TS 25237:2008). If the pseudonymization follows a repeatable algorithm, different practitioners can match records belonging to the same individual from different data sets. However, the same practitioners will have the ability to re-identify the pseudonymized data as part of the matching process. Pseudonymization can also be reversed if the entity that performed the pseudonymization retains a table linking the original identities to the pseudonyms, a technique called *unmasking*.

2.4 Re-identification Risk and Data Utility

Those receiving de-identified data may attempt to learn the identities of the data subjects that have been removed. This process is called *re-identification.* Because an important goal of de-identification is to prevent unauthorized re-identification, such attempts are sometimes called *re-identification attacks.*

The term “attack” is borrowed from the literature of computer security, in which the security of a computer system or encryption algorithm is analyzed through the use of a hypothetical “attacker” in possession of specific skills, knowledge, and access. A risk assessment involves cataloging the range of potential attackers and, for each, the likelihood of success.

There are many reasons that an individual or organization might attempt a re-identification attack:

- **To test the quality of the de-identification.** For example, a researcher might conduct the re-identification attack at the request of the data custodian performing the de-identification
- **To gain publicity or professional standing for performing the de-identification.** Several successful re-identification efforts have been newsworthy and professionally rewarding for the researchers conducting them.
- **To embarrass or harm the organization that performed the de-identification.** Organizations that perform de-identification generally have an obligation to protect the personal information contained in the original data. As such, demonstrating that their privacy protecting measures were inadequate can embarrass or harm these organizations.
- **To gain direct benefit from the de-identified data.** For example, a marketing company might purchase de-identified medical data and attempt to match up medical records with identities, so that the re-individuals could be sent targeted coupons.

In the literature, re-identification attacks sometimes described as being performed by a hypothetical *data intruder* who is in possession of the de-identified dataset and some additional *background information*.

Re-identification risk is the measure of the risk that the identities and other information about individuals in the data set will be learned from the de-identified data. It is hard to quantify this risk, as the ability to re-identify depends on the original data set, the de-identification technique, the technical skill of the data intruder, the intruder’s available resources, and the availability of additional data that can be linked with the de-identified data. In many cases the risk of re-identification will increase over time as techniques improve and more background information become available.

Researchers have taken various approaches for computing and reporting the re-identification risk including:

- The risk that a specific person in the database can be re-identified. (The “prosecutor scenario.”)
- The risk that any person in the database can be re-identified. (The “journalist scenario.”)
- The percentage of identities in the database that is actually re-identified.

- The distinguishability between an analysis performed on a database containing an individual and on a database that does not contain the individual. (The "differential identifiability" scenario.[16])

Likewise, different standards that have been used to describe the abilities of the "attacker" including:

- A member of general public who has access to public information on the web
- A computer scientist skilled in re-identification ("expert")
- A member of the organization that produced the dataset ("insider")
- A friend or family member of the data subject
- The data subject ("self re-identification")

The purpose of de-identifying data is to allow some uses of the de-identified data while providing for some privacy protection. These two goals are generally antagonistic, in that there is a trade off between the amount of de-identification and the utility of the resulting data. The more securely the data are de-identified, the less utility remains. In general, privacy protection increases as more information is removed or modified from the original data set, but the remaining data are less useful as a result. It is the responsibility of those de-identifying to determine an acceptable trade-off.

A variety of harms that can result from the use or distribution of de-identified data, including:

- ***Incomplete de-identification.*** Identifiable private information may inadvertently remain in the de-identified data set. This was the case in search query data released by AOL in 2006, in which journalists re-identified and contacted an AOL user through identifying information that the user had typed as search queries.[17]
- ***Identity disclosure (***also called ***attribute disclosure*** and ***re-identification by linking).*** It may be possible to re-identify specific records by linking some of the remaining data with similar attributes in another, identifying data set. De-identification is supposed to protect against this harm.
- ***Inferential disclosure.*** If a data set reveals that all individuals who share a characteristic have a particular attribute, and if the adversary knows of an individual in the sample who has that characteristic, than that individual's attribute is exposed. For example, if a hospital releases information showing that all 20-year-old female patients treated had a specific diagnosis, and if Alice Smith is a 20-year-old female that is known to have been treated at the hospital, then Alice Smith's diagnosis can be inferred, even though her

[16] Jaewoo Lee and Chris Clifton. 2012. Differential identifiability. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 1041-1049. DOI=10.1145/2339530.2339695 http://doi.acm.org/10.1145/2339530.2339695

[17] Barbaro M, Zeller Jr. T. A Face Is Exposed for AOL Searcher No. 4417749 New York Times. 9 August, 2006.

individual de-identified medical records cannot be distinguished from the others.[18] In general, de-identification is not designed to protect against inference-based attacks.

- ***Association harms.*** Even though it may not be possible to match a specific data record with an individual, it may be possible to associate an individual with the dataset as a whole or with a group of records within the dataset. That association may result in some kind of stigma for the data subject.
- ***Group harms.*** Even if it is not possible to match up specific data records with individuals, the data may be used to infer a characteristic and associate it with a group represented in the data.
- ***Unmasking***. If the data were pseudonymized, it may be possible reverse the pseudonymization process. This might be done by using a table that shows the mapping of the original identities to the pseudonyms, by reversing the pseudonymization algorithm, or by performing a brute-force search in which the pseudonymization algorithm is applied to every possible identity until the matching pseudonym is discovered.

Organizations considering de-identification must therefore balance:

- The effort that the organization can spend performing and testing the de-identification process.
- The utility desired for the de-identified data.
- The harms that might arise from the use of the de-identified data.
- The ability to use other controls that can minimize the risk.
- The likelihood that an attacker will attempt to re-identify the data, and the amount of effort that the attacker might be willing to spend.

Privacy laws in the US tend to be concerned with regulating and thereby preventing the first two categories of harms—the release of incompletely de-identified data, and assigning of an identity to a specific record in the de-identified set. The other harms tend to be regulated by organizations themselves, typically through the use of Institutional Review Boards or other kinds of internal controls.

2.5 Release models and data controls

One way to limit the chance of re-identification is to place controls on the way that the data may be obtained and used. These controls can be classified according to different release models. Several named models have been proposed in the literature, ranging from no restrictions to tightly restricted. They are:

- **The Release and Forget model**[19]**:** The de-identified data may be released to the public, typically by being published on the Internet. It can be difficult or impossible for an organization to recall the data once released in this fashion.

[18] El Emam Methods for the de-identification of electronic health records for genomic research. Genome Medicine 2011, 3:25 http://genomemedicine.com/content/3/4/25

- **The Click-Through model**[20]: The de-identified data can are made available on the Internet, but the user must agree in advance to some kind of "click-through" data use agreement. In this event, an entity that performed and publicized a successful re-identification attack might be subject to shaming or sanctions.
- **The Qualified Investigator model**[21]**:** The de-identified data may be made available to qualified researchers under data use agreements. Typically these agreements prohibit attempted re-identifying, redistribution, or contacting the data subjects.
- **The Enclave model**[22]**:** The de-identified data may be kept in some kind of segregated enclave that accepts queries from qualified researchers, runs the queries on the de-identified data, and responds with results. (This is an example of PPDM, rather than PPDP.)

Gellman has proposed model legislation that would strengthen data use agreements.[23] Gellman's proposal would recognize a new category of information *potentially identifiable personal information (PI^2)*. Consenting parties could add to their data-use agreement a promise from the data provider that the data had been stripped of personal identifiers but still might be re-identifiable. Recipients would then face civil and criminal penalties if they attempted to re-identify. Thus, the proposed legislation would add to the confidence that de-identified data would remain so. "Because it cannot be known at any time whether information is re-identifiable, virtually all personal information that is not overtly identifiable is PI^2," Gellman notes.

3 Syntactic De-Identification Approaches and Their Criticism

Syntactic de-identification techniques[24] are techniques that attempt to de-identify by removing specific data elements from a data set based on element type. This section introduces the terminology used by such schemes, discusses the de-identification standard of the Health Insurance Portability and Privacy Act (HIPAA) Privacy Rule, and discusses critiques of the syntactic techniques and efforts that have appeared in the academic literature.

---

[19] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010

[20] K El Emam and B Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

[21] Ibid.

[22] Ibid.

[23] Gellman, Robert; "The Deidentification Dilemma: A Legislative and Contractual Proposal," July 12, 2010.

[24] Chris Clifton and Tamir Tassa, 2013. On Syntactic Anonymity and Differential Privacy. Trans. Data Privacy 6, 2 (August 2013), 161-183.

3.1 Removal of Direct Identifiers

Syntactic de-identification approaches are easiest to understand when applied to a database containing a single table of data. Each row contains data for a different individual.

Direct identifiers, also called *directly identifying variables* and *direct identifying data*, are "data that directly identifies a single individual." (ISO/TS 25237:2008) Examples of direct identifiers include names, social security numbers and any "data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."[25] Many practitioners treat information such as medical record numbers and phone numbers as direct identifiers, even though additional information is required to link them to an identity.

Direct identifiers must be removed or otherwise transformed during de-identification. This processes is sometimes called *data masking*. There are at least three approaches for masking:

1) The direct identifiers can be removed.
2) The direct identifiers can be replaced with random values. If the same identify appears twice, it receives two different values. This preserves the form of the original data, allowing for some kinds of testing, but makes it harder to re-associate the data with individuals.
3) The direct identifiers can be systematically replaced with pseudonyms, allowing records referencing the same individual to be matched. Pseudonymization may also allow for the identities to be unmasked at some time in the future if the mapping between the direct identifiers and the pseudonyms is preserved or re-generated.

Direct Identifiers								
Name	Address	Birthday	ZIP	Sex	Weight	Diagnosis	…	…

Table 1: A hypothetical data table showing direct identifiers

Early efforts to de-identify databases stopped with the removal of direct identifiers.

3.2 Re-identification through Linkage

The *linkage attack* is the primary technique for re-identifying data that have been syntactically de-identified. In this attack, each record in the de-identified dataset is linked with similar records in a second dataset that contains both the linking information and the identity of the data subject.

Linkage attacks of this type were developed by Sweeney, who re-identified the medical records of Massachusetts governor William Weld as part of her graduate work at MIT. At the time Massachusetts was distributing a research database containing de-identified insurance

[25] ISO/TS 25237:2008(E), p.3

reimbursement records of Massachusetts state employees that had been hospitalized. To protect the employees' privacy, their names were stripped from the database, but the employees' date of birth, zip code, and sex was preserved to allow for statistical analysis.

Knowing that Weld had recently been treated at a Massachusetts hospital, Sweeney was able to re-identify the governor's records by searching for the "de-identified" record that matched the Governor's date of birth, zip code, and sex. She learned this information from the Cambridge voter registration list, which she purchased for $20. Sweeney then generalized her findings, arguing that up to 87% of the US population was uniquely identified by 5-digit ZIP code, date of birth, and sex.[26]

Sweeney's linkage attack can be demonstrated graphically:

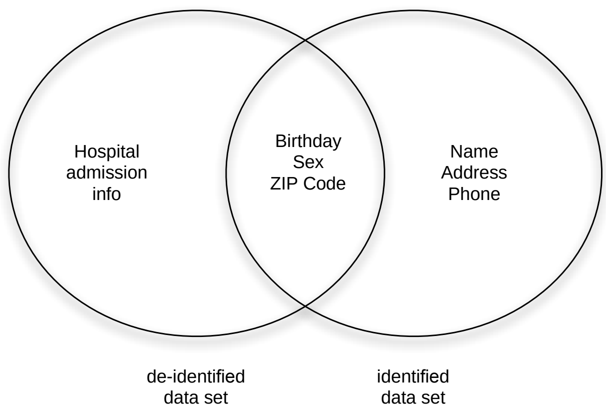

Figure 2: Linkage attacks combine information from two or more data sets to re-identify records

Many factors complicate such linkage attacks, however;

- In order to be linkable, a person needs to be in both data sets. Sweeney knew that Weld was in both data sets.
- Only records that are uniquely distinguished by the linking variables in both sets can be linked. In this case, a person's records can only be linked if no one else shares their same birthday, sex and ZIP in either data set. As it turned out, no other person in Cambridge shared Weld's date of birth.
- If the variables are not the same in both data sets, then the data must be normalized so that they can be linked. This normalization can introduce errors. This was not an issue in the Weld case, but it could be an issue if one dataset reported "age" and another reported "birthday."
- Verifying whether or not a link is correct requires using information that was not used as part of the linkage operation. In this case, Weld's medical records were verified using newspaper accounts of what had happened.

---

[26] Sweeney L., Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000. http://dataprivacylab.org/projects/identifiability/paper1.pdf

3.3 De-identification of Quasi-Identifiers

Quasi-identifiers, also called *indirect identifiers* or *indirectly identifying variables,* are identifiers that by themselves do not identify a specific individual but can be aggregated and "linked" with information in other data sets to identify data subjects. The re-identification of William Weld's medical records demonstrated that birthday, ZIP, and Sex are quasi-identifiers.

Direct Identifiers		Quasi-Identifiers						
Name	Address	Birthday	ZIP	Sex	Weight	Diagnosis	…	…

Table 2: A hypothetical data table showing direct identifiers and quasi-identifiers

Quasi-identifiers pose a significant challenge for de-identification. Whereas direct identifiers can be removed from the data set, quasi-identifiers generally convey some sort of information that might be important for a later analysis. As such, they cannot be simply masked.

Several approaches have been proposed for de-identifying quasi-identifiers:

1) **Suppression:** The quasi-identifier can be suppressed or removed. Removing the data maximizes privacy protection, but decreases the utility of the dataset.
2) **Generalization:** The quasi-identifier can be reported as being within a specific range or as a member of a set. For example, the ZIP code 12345 could be generalized to a ZIP code between 12000 and 12999. Generalization can be applied to the entire data set or to specific records.
3) **Swapping:** Quasi-identifiers can be exchanged between records. Swapping must be handled with care if it is necessary to preserve statistical properties.
4) **Sub-sampling.** Instead of releasing an entire data set, the de-identifying organization can release a sample. If only subsample is released, the probability of re-identification decreases.[27]

K-anonymity[28] is a framework developed by Sweeney for quantifying the amount of manipulation required of the quasi-identifiers to achieve a given desired level of privacy. The technique is based on the concept of an ***equivalence class,*** the set of records that have the same quasi-identifiers. A dataset is said to be *k-anonymous* if, for every combination of quasi-identifiers, there are at least *k* matching records. For example, if a dataset that has the quasi-identifiers birth year and state has k=4 anonymity, then there are at least four records for every combination of (birth year, state) combination. Successive work has refined *k-anonymity* by

[27] El Emam, Methods for the de-identification of electronic health records for genomic research, Genome Medicine 2011, 3:25 http://genomemedicine.com/content/3/4/25

[28] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570. DOI=10.1142/S0218488502001648 http://dx.doi.org/10.1142/S0218488502001648

adding requirements for diversity of the sensitive attributes within each equivalence class[29], and requiring that the resulting data are statistically close to the original data[30].

El Emam and Malin[31] have developed an 11-step process for de-identifying data based on the identification of identifiers and quasi-identifiers:

- **Step 1: Determine direct identifiers in the data set.** An expert determines the elements in the data set that serve only to identify the data subjects.
- **Step 2: Mask (transform) direct identifiers.** The direct identifiers are either removed or replaced with pseudonyms.
- **Step 3: Perform threat modeling**. The organization determines "plausible adversaries," the additional information they might be able to use for re-identification, and the quasi-identifiers that an adversary might use for re-identification.
- **Step 4: Determine minimal acceptable data utility.** In this step the organization determines what uses can or will be made with the de-identified data, to determine the maximal amount of de-identification that could take place.
- **Step 5: Determine the re-identification risk threshold.** The organization determines acceptable risk for working with the data set and possibly mitigating controls.
- **Step 6: Import (sample) data from the source database.** Because the effort to acquire data from the source (identified) database may be substantial, the authors recommend a test data import run to assist in planning.
- **Step 7: Evaluate the actual re-identification risk.** The actual identification risk is mathematically calculated.
- **Step 8: Compare the actual risk with the threshold.** The result of step 5 and step 7 are compared.
- **Step 9: Set parameters and apply data transformations.** If the actual risk is less than the minimal acceptable risk, the de-identification parameters are applied and the data is transformed. If the risk is too high then new parameters or transformations need to be considered.
- **Step 10: Perform diagnostics on the solution.** The de-identified data are tested to make sure that it has sufficient utility and that re-identification is not possible within the allowable parameters.
- **Step 11: Export transformed data to external data set.** Finally, the de-identified data are exported and the de-identification techniques are documented in a written report.

---

[29] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[30] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE (Purdue University).

[31] K. El Emam and B. Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

The chief criticism of de-identification based on direct identifiers and quasi-identifiers is that it is difficult to determine which fields are identifying, and which are non-identifying data about the data subjects. Aggarwal identified this problem in 2005, noting that when the data contains a large number of attributes, "an exponential number of combinations of dimensions can be used to make precise inference attacks… [W]hen a data set contains a large number of attributes which are open to inference attacks, we are faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity."[32]

Work since has demonstrated some of Aggarwal's concerns: many seemingly innocuous data fields can become identifying for an adversary that has the appropriate matching information (see Section 3.5). Furthermore, values that cannot be used as quasi-identifiers today may become quasi-identifiers in the future as additional datasets are developed and released. To accurately assess re-identification risk, it is therefore necessary to accurately model the knowledge, determination, and computational resources of the adversaries that will be attempting the re-identification.

3.4 De-identification of Protected Health Information (PHI) under HIPAA

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule describes two approaches for de-identifying Protected Health Information (PHI): The Expert Determination Method (§164.514(b)(1)) and the Safe Harbor method (§164.514(b)(2)).

The "Expert Determination" method provides for an expert to examine the data and determine an appropriate means for de-identification that would minimize the risk of re-identification. The specific language of the Privacy Rule states:

> "(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
> (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
> (ii) Documents the methods and results of the analysis that justify such determination; or"

The "Safe Harbor" method allows a covered entity to treat data as de-identified if it by removing 18 specific types of data for "the individual or relatives, employers, or household members of the individual." The 18 types are:

> "(A) Names
> (B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau

[32] Charu C. Aggarwal. 2005. On *k*-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases* (VLDB '05). VLDB Endowment 901-909.

of the Census:

(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and

(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000

(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older

(D) Telephone numbers

(E) Fax numbers

(F) Email addresses

(G) Social security numbers

(H) Medical record numbers

(I) Health plan beneficiary numbers

(J) Account numbers

(K) Certificate/license numbers

(L) Vehicle identifiers and serial numbers, including license plate numbers

(M) Device identifiers and serial numbers

(N) Web Universal Resource Locators (URLs)

(O) Internet Protocol (IP) addresses

(P) Biometric identifiers, including finger and voiceprints

(Q) Full-face photographs and any comparable images

(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"];"

In addition to removing these data, the covered entity must not "have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information."

The Privacy Rule is heavily influenced by Sweeny's research, as evidenced by its citation of Sweeny's research the rule's specific attention to the quasi-identifiers identified by Sweeny (ZIP code and birthdate) for generalization. The Privacy Rule strikes a balance between the risk of re-identification and the need to retain some utility in the data set—for example, by allowing the reporting of the first 3 digits of the ZIP code and the year of birth. Researchers have estimated that properly applied, the HIPAA Safe Harbor rule seems to allow the identification probability of approximately 1.5%.[33]

The actual rate of re-identification may be lower in some cases. In 2010 the Office of the National Coordinator for Health Information Technology (ONC HIT) at the US Department of Health and Human Services conducted a test of the HIPAA de-identification standard. As part of

[33] Jaewoo Lee and Chris Clifton, Differential Identifiability, KDD '12, Aug. 12-16, 2012. Bejing, China.

the study, researchers were provided with 15,000 hospital admission records belonging to Hispanic individuals from a hospital system between 2004 and 2009. Researchers then attempted to match the de-identified records to a commercially available data set of 30,000 records from InfoUSA. Based on the Census data the researchers estimated that the 30,000 commercial records covered approximately 5,000 of the hospital patients. When the experimenters matched using Sex, ZIP3 (the first 3 digits of the ZIP code, as allowed by HIPAA), and Age, they found 216 unique records in the hospital data, 84 unique records in the InfoUSA data, and only 20 records that matched on both sides. They then attempted to confirm these matches with the hospital and found that only 2 were actual matches, which were defined as having the same 5-digit ZIP code, the same last name, same street address, and same phone number. This represents a re-identification rate of 0.013%; the researchers also calculate a more conservative re-identification risk of 0.22%.

HIPAA also allows the sharing of *limited data sets* that have been partially de-identified but still include dates, city, state, zip code, and age. Such data sets may only be shared for research, public health, or health care operations, and may only be shared with if a data use agreement is executed between the covered entities to assure for subject privacy.[34] At minimum, the data use agreements must require security safeguards, require that all users of the data be similarly limited, and prohibit contacting of the data subjects.

3.5 Evaluation of Syntactic De-identification

The basic assumption of syntactic de-identification is that some of the columns in a data set might contain useful information without being inherently identifying. In recent years a significant body of academic research has shown that this assumption is not true in some cases.

- ***Netflix Prize:*** Narayanan and Shmatikov showed in 2008 that in many cases the set of movies that a person had watched could be used as an identifier. [35] Netflix had released a de-identified data set of movies that some of its customers had watched and ranked as part of its "Netflix Prize" competition. The researchers showed that a set common movies could be used to link many records in the Netflix dataset with similar records in the Internet Movie Data Base (IMDB), which had not been de-identified. The threat scenario is that by rating a few movies on IMDB, a person might inadvertently reveal *all* of the movies that they had watched, since the IMDB data could be linked with the public data from the Netflix Prize.

- ***Medical Tests:*** Atreya et al. showed in 2013 that 5-7 laboratory results from a patient could be used "as a search key to discover the corresponding record in a de-identified biomedical research database." [36] Using a dataset with 8.5 million laboratory results from

[34] http://privacyruleandresearch.nih.gov/pr_08.asp

[35] Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

[36] Atreya, Ravi V, Joshua C Smith,Allison B McCoy, Bradley Malin and Randolph A Miller, "Reducing patient re-identification risk for laboratory results within research datasets," J Am Med Inform Assoc 2013;20:95–101. doi:10.1136/amiajnl-2012-001026.

61,280 patients, the researchers found that four consecutive laboratory test results uniquely identified between 34% and 100% of the population, depending on the test. The two most common test results, CHEM7 and CBC, respectively identified 98.9% and 98.8% of the test subjects. The threat scenario is that a person who intercepted a single lab identified lab report containing a CHEM7 or CBC result could use report to search the de-identified biomedical research database for other records belonging to the individual.

- ***Mobility Traces:*** Also in 2013, Montjoye et al. showed that people and vehicles could be identified by their "mobility traces" (a record of locations and times that the person or vehicle visited). In their study, trace data for 1.5 million individuals was processed, with time values being generalized to the hour and spatial data generalized to the resolution provided by a cell phone system (typically 10-20 city blocks). The researchers found that four randomly chosen observations of an individual putting them at a specific place and time was sufficient to uniquely identify 95% of the data subjects.[37] Space/time points for individuals can be collected from a variety of sources, including purchases with a credit card, a photograph, or Internet usage. A similar study performed by Ma et al. found that 30%-50% of individuals could be identified with 10 pieces of side information.[38] The threat scenario is that person who revealed 5 place/time pairs (perhaps by sending email from work and home at four times over the course of a month) would make it possible for an attacker to identify their entire mobility trace in a publicly released data set.

- ***Taxi Ride Data:*** In 2014 The New York City Taxi and Limousine Commission released a dataset containing a record of every New York City taxi trip in 2013 (173 million in total). The data did not include the names of the taxi drivers or riders, but it did include a 32-digit alphanumeric code that could be readily converted to each taxi's medallion number. A data scientist intern at the company Neustar discovered that he could find time-stamped photographs on the web of celebrities entering or leaving taxis in which the medallion was clearly visible.[39] With this information the was able to discover the other end-point of the ride, the amount paid, and the amount tipped for two of the 173 million taxi rides. A reporter at the Gawker website was able to identify another nine. [40]

The experience with the Netflix Prize indicates and the laboratory results shows that many sets

---

[37] Yves-Alexandre de Montjoye et al., Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3 (2013), Article 1376.

[38] Ma, C.Y.T.; Yau, D.K.Y.; Yip, N.K.; Rao, N.S.V., "Privacy Vulnerability of Published Anonymous Mobility Traces," Networking, IEEE/ACM Transactions on , vol.21, no.3, pp.720,733, June 2013

[39] "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," Anthony Tockar, September 15, 2014, http://research.neustar.biz/author/atockar/

[40] "Public NYC Taxicab Database Lets you See How Celebrities Tip," J. K. Trotter, GAWKER, October 23, 2014. http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

of sensitive values might also be identifying, provided that there is sufficient range or diversity for the identifiers in the population. The experience with the taxi data shows that there are many unanticipated sources of data that might correlate with other information in the data record.

The taxi and mobility trace studies demonstrate the strong identification power of geospatial information. Since each person can only be at one place at one time, just a few observations of a person's location and time can be highly identifying, even in a data set that generalized and noisy. Furthermore, some locations are highly identifying—either because they are isolated or well photographed.

However, the medical tests and taxi studies also show that relatively small changes to the data may make re-identification difficult or impossible. Atreya et al. demonstrated this directly. In the case of the Taxi data, the celebrities were only identified because the taxi medallion number pseudonymization could be unmasked, and the main privacy impact was the release of the specific geographical locations and tip amounts. If the medallion number had been properly protected and if the GPS location data had be aggregated to a 100 meter square grid, the risk of re-identification would have been considerably reduced. As it was, the taxi data demonstrates that the risk of re-identification under the "journalist scenario" (which sees any failure as a significant shortcoming) may be high, but risk under the "prosecutor scenario" might be very low (11 out of 173 million).

Putting this information into context of real-world de-identification requirements is difficult. For example, the ONC HIT 2010 study only attempted to match using the specific quasi-identifiers anticipated by the HIPAA Privacy Rule—age in years, sex, and ZIP3. Atreya et al. used a different threat model, one in which the attacker was assumed to have the results of a laboratory test. The results of Atreya imply that *if* the ONC HIT study included laboratory test results, and *if* the attacker had a laboratory test report including the patient's name and seven or more test results, then there is an overwhelming probability that there is a specific set of records in the de-identified data that are an exact match. However, this test was never done, and many may feel that it is not a realistic threat model.

El Emam et al[41] reviewed 14 re-identification attempts published between 2001 and 2010. For each the authors determined whether or not health data had been included, the profession of the adversary, the country where the re-identification took place, the percentage of the records that had been re-identified, the standards that were followed for de-identification, and whether or not the re-identification had been verified. The researchers found that the successful re-identification events typically involved small data sets that had not been de-identified according to existing standards. As such, drawing scientific conclusions from these cases is difficult. In many cases the re-identification attackers have re-identified just a few records but stated that many more could be re-identified.

De-identification and PPDP are still possible, but require a more nuanced attention to the potential for re-identification of the data subjects. One approach is to treat all data in the dataset

[41] K El Emam, E Jonker, L Arbuckle, B MalinB (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi:10.1371/journal.pone.0028071

as quasi-identifiers and accordingly manipulate them to protect privacy. This is possible, but may require developing specific technology for each different data type. For example, Atreya et al. developed an "expert" algorithm that could de-identify the data by perturbing the test results with minimal impact on diagnostic accuracy.[42]

3.6 Alternatives to Syntactic De-identification

An alternative to syntactic de-identification is to generate synthetic data or synthetic data sets that are statistically similar to the original data but which cannot be de-identified because they are not based on actual people. Synthetic data elements are widely used in statistical disclosure controls—for example, by aggregating data into categories, suppressing individual cells, adding noise, or swapping data between similar records.

4 Challenges in De-Identifying Contextual Data

Whereas the last chapter was concerned mostly with the de-identification of tabular or structured data, this section concerns itself with the open challenges of de-identifying contextual data.

4.1 De-identifying medical text

Medical records contain significant amounts of unstructured text. In recent years there has been a significant effort to develop and evaluate tools designed to remove the 18 HIPAA data elements from free-format text using natural language processing techniques. The two primary techniques explored have been rule-based systems and statistical systems. Rule-based systems tend to work well for specific kinds of text but do not work well when applied to new domains. Statistical tools generally perform less accurately and require labeled training data, but are easier to repurpose to new domains.

Multiple factors combine to make de-identifying text narratives hard:

1) Direct identifiers such as names and addresses may not be clearly marked.
2) Important medical information may be mistaken for personal information and removed. This is especially a problem for eponyms which are commonly used in medicine to describe diseases (e.g. Addison's Disease, Bell's Palsy, Reiter's Syndrome, etc.)
3) Even after the removal of the 18 HIPAA elements, information may remain that allows identification of the medical subject.
4) Medical information currently being released as "de-identified" frequently does not conform to the HIPAA standard.

In general the best systems seem to exhibit overall accuracy between 95-98% compared to human annotators. A study by Meystre, Shen et. al showed the automatically de-identified records from the Veteran's Administration were not recognized by the patient's treating professional.[43]

---

[42] Atreya, *supra*.

[43] Meystre S et al., Can Physicians Recognize Their Own Patients in De-Identified Notes? In Health – For Continuity of Care C.

Several researchers have performed formal evaluations of de-identification tools:

- In 2012 Deleger et al at Cincinnati Children's Hospital Medical Center tested The MITRE Identification Scrubber Toolkit (MIST)[44] against MCRF, an in-house system developed by CCHMC based on the MALLET machine-learning package. The reference corpora were 3503 clinical notes selected from 5 million notes created at CCHMC in 2010, the 2006 i2b2 de-identification challenge corpus,[45] and the PhisyoNet corpus.[46] [47]

- In 2013 Ferrández *et al* at the University of Utah Department of Biomedical Informatics performed an evaluation of five automated de-identification systems against two reference corpora. The test was conducted with the 2006 i2b2 de-identification challenge corpus, consisting of 889 documents that had been de-identification and then given synthetic data,[48] and a corpus of 800 documents provided by the Veterans Administration that was randomly drawn from documents with more than 500 words dated between 4/01/2008 and 3/31/2009.

- In 2013 The National Library of Medicine issued a report to its Board of Scientific Counselors entitled "Clinical Text De-Identification Research" in which the NLM compared the performance of its internally developed tool, the NLM Scrubber (NLM-S), with the MIT de-identification system (MITdeid) and MIST.[49] The test was conduct with an internal corpus of 1073 Physician Observation Reports and 2020 Patient Study Reports from the NIH Clinical Center.

Both the CCHMC and the University of Utah studies tested the systems "out-of-the-box" and after they were tuned by using part of the corpus as training data. The Utah study found that none of the de-identification tools worked well enough to de-identify the VHA records for public release, and that the rule-based systems exceled for finding certain kinds of information (e.g. SSNs and phone numbers), while the trainable systems worked better for other kinds of data.

Lovis et al. (Eds.) © 2014 European Federation for Medical Informatics and IOS Press.

[44] Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;79:849e59.

[45] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de- identification. J Am Med Inform Assoc 2007;14:550e63.

[46] Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.

[47] Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. Circulation 2000;101:E215e20.

[48] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de- identification. J Am Med Inform Assoc 2007;14:550e63.

[49] Kayaalp M et al, A report to the Board of Scientific Counselors, 2013, The Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Although there are minor variations between the systems, they are all had similar performance. The NLM study found that NLM-S significantly outperformed MIST and MITdeid on the NLM data set, removing 99.2% of the tokens matching the HIPAA Privacy Rule. The authors concluded that the remaining tokens would not pose a significant threat to patient privacy.

It should be noted that none of these systems attempt to de-identify data beyond removal of the 18 HIPAA data elements, leaving the possibility that individuals could be re-identified using other information. For example, regulations in both the US and Canada require reporting of adverse drug interactions. These reports have been re-identified by journalists and researchers by correlating reports of fatalities with other data sources, such as news reports and death registers.

4.2 De-identifying Imagery

Multimedia imagery such as still photographs, consumer videos and surveillance video pose special de-identification challenges because of the wealth of identity information they potentially contain. Similar issues come into play when de-identifying digital still imagery, video, and medical imagery (X-Rays, MRI scans, etc.)

In general there are a three specific identification concerns:

1) The image itself may contain the individual's name on a label that is visible to a human observer but readily difficult to detect programmatically.
2) The file format may contain metadata that specifically identifies the individual. For example, there may be a GPS address of the person's house, or the person's name may be embedded in a header.
3) The image may contain an identifying biometric such as a scar, a hand measurement, or a specific injury.

Early research had the goal of producing images in which the faces could not be reliably identified by face recognition systems. In many cases this is sufficient: blurring is used by Google Street View, one of the largest deployments of photo de-identification technology.[50] Google claims that its completely automatic system is able to blur 89% of faces and 94-96% of license plates. Nevertheless, journalists have criticized Google for leaving many faces unblurred[51] and for blurring the faces of religious effigies[52],[53].

Some researchers have developed systems that can identify and blur bodies,[54] as research has

---

50 Frome, Andrea, et al, "Large-scale Privacy Protection in Google Street View," IEEE International Conference on Computer Vision (2009).

51 Stephen Chapman, "Google Maps, Street View, and privacy: Try harder, Google," ZDNet, January 31, 2013. http://www.zdnet.com/article/google-maps-street-view-and-privacy-try-harder-google/

52 Gonzalez, Robbie. "The Faceless Gods of Google Street View," io9, October 4, 2014. http://io9.com/the-faceless-gods-of-google-street-view-1642462649

53 Brownlee, John, "The Anonymous Gods of Google Street View," Fast Company, October 7, 2014. http://www.fastcodesign.com/3036319/the-anonymous-gods-of-google-street-view#3

54 Prachi Agrawal and P. J. Narayanan. 2009. Person de-identification in videos. In Proceedings of the 9th Asian conference on

shown that bodies are frequently identifiable without faces.[55] An experimental system can locate and remove identifying tattoos from still images.[56]

Blurring and pixilation have the disadvantage of creating a picture that is visually jarring. Care must be taken if pixilation or blurring are used for obscuring video, however, as technology exists for de-pixelating and de-blurring video by combining multiple images. To address this, some researchers have developed systems that can replace faces with a composite face,[57,58] or with a face that is entirely synthetic.[59,60]

Quantifying the effectiveness of these algorithms is difficult. While some researchers may score the algorithms against face recognition software, other factors such as clothing, body pose, or geo-temporal setting might make the person identifiable by associates or themselves. A proper test of image de-identification should therefore include a variety of re-identification scenarios.

4.3 De-identifying Genetic sequences and biological materials

Genetic sequences are not considered to be personally identifying information by HIPAA's de-identification rule. Nevertheless, because genetic information is inherited, genetic sequences have been identified through the use of genetic databanks even if the individual was not previously sequenced and placed in an identification database.

In 2005 a 15-year-old teenager used the DNA-testing service FamilyTreeDNA.com to find his sperm donor father. The service, which cost $289, did not identify the boy's father, but it did identify two men who had matching Y-chromosomes. The two men had the same surname but with different spellings. As the Y-Chromosome is passed directly from father to son with no modification, it tends to be inherited the same way as European surnames. With this information and with the sperm donor's date and place of birth (which had been provided to the boy's

---

Computer Vision - Volume Part III (ACCV'09), Hongbin Zha, Rin-ichiro Taniguchi, and Stephen Maybank (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 266-276. DOI=10.1007/978-3-642-12297-2_26 http://dx.doi.org/10.1007/978-3-642-12297-2_26

[55] Rice, Phillips, et al., Unaware Person Recognition From the Body when Face Identification Fails, Psychological Science, November 2013, vol. 24, no. 11, 2235-2243 http://pss.sagepub.com/content/24/11/2235

[56] Darijan Marčetić et al., An Experimental Tattoo De-identification System for Privacy Protection in Still Images, MIPRO 2014, 26-30 May 2014, Opatija, Croatia

[57] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. In: Protecting Privacy in Video Surveillance, A. Senior, editor. Springer, 2009 Preserving Privacy by De-identifying Facial Images. http://dataprivacylab.org/projects/facedeid/paper.pdf

[58] E. Newton, L. Sweeney, and B. Malin. Preserving Privacy by De-identifying Facial Images, Carnegie Mellon University, School of Computer Science, Technical Report, CMU-CS-03-119. Pittsburgh: March 2003.

[59] Saleh Mosaddegh, Löıc Simon, Frederic Jurie. Photorealistic Face de-Identification by Aggregating Donors' Face Components. Asian Conference on Computer Vision, Nov 2014, Singapore. pp.1-16.

[60] Umar Mohammed, Simon J. D. Prince, and Jan Kautz. 2009. Visio-lization: generating novel facial images. In ACM SIGGRAPH 2009 papers (SIGGRAPH '09), Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 57, 8 pages. DOI=10.1145/1576246.1531363 http://doi.acm.org/10.1145/1576246.1531363

mother), the boy was able to identify his father using an online search service.[61]

In 2013 a group of researchers at MIT extended the experiment, identifying surnames and complete identities of more than 50 individuals who had DNA tests released on the Internet as part of the Study of Human Polymorphisms (CEPH) project and the 1000 Genomes Project.[62]

At the present time there is no scientific consensus on the minimum size of a genetic sequence necessary for re-identification. There is also no consensus on an appropriate mechanism to make de-identified genetic information available to researchers without the need to execute a data use agreement.

4.4 De-identification of geographic and map data

De-identification of geographic data is not well researched. Current methods rely on perturbation and generalization. Perturbation is problematical in some cases, because perturbed locations can become nonsensical (e.g. moving a restaurant into a body of water). Generalization may not be sufficient to hide identity, however, especially if the population is sparse or if multiple observations can be correlated.

However, without some kind of generalization or perturbation there is so much diversity in geographic data that it may be impossible to de-identify locations. For example, measurement of cell phone accelerometers taken over a time period can be used to infer position by fitting movements to a street grid.[63] This is of concern because the Android and iOS operating systems do not consider accelerometers to be sensitive information.

4.5 Estimation of Re-identification Risk

Practitioners are in need of easy-to-use procedures for calculating the risk of re-identification given a specific de-identification protocol. Calculating this risk is complicated, as it depends on many factors, including the distinctiveness of different individuals within the sampled data set, the de-identification algorithm, the availability of linkage data, and the range of individuals that might mount a re-identification attack.

There are also different kinds of re-identification risk. A model might report the average risk of each subject being identified, the risk that *any* subject will be identified, the risk that individual subjects might be identified as being 1 of *k* different individuals, etc.

Danker et al. propose a statistical model and decision rule for estimating the distinctiveness of different kinds of data sources.[64] El Emam et al. developed a technique for modeling the risk of

[61] Sample, Ian. Teenager finds sperm donor dad on internet. The Guardian, November 2, 2005. http://www.theguardian.com/science/2005/nov/03/genetics.news

[62] Gymrek et al, Identifying Personal Genomes by Surname Inference, Science 18 Jan 2013, 339:6117.

[63] Jun Han; Owusu, E.; Nguyen, L.T.; Perrig, A.; Zhang, J., "ACComplice: Location inference using accelerometers on smartphones," *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pp.1,9, 3-7 Jan. 2012

[64] Dankar et al. Estimating the re-identification risk of clinical data sets, BMC Medical Informatics and Decision Making 2012, 12:66.

re-identifying adverse drug event reports based on two attacker models: a "mildly motivated adversary" whose goal is to identify a single record, and a "highly motivated adversary" that wishes to identify and verify all matches, "and is only limited by practical or financial considerations."[65]

Practitioners are also in need of standards for acceptable risk. As previously noted, researchers have estimated that properly applied, the HIPAA Safe Harbor rule seems to allow the identification probability of approximately 1.5%.[66] El Emam and Alvarez are critical of the "Article 29 Working Party Opinion 05/2014 on data anonymization techniques" because the document appears to only endorse de-identification techniques that produce zero risk of re-identification.[67]

5 Conclusion

De-identification techniques can reduce or limit the privacy harms resulting from the release of a data set, while still providing users of the data with some utility.

To date, the two primary harms associated with re-identification appear to be damage to the reputation of the organization that performed the de-identification, and the discovery of private facts of people who were re-identified. Researchers or journalists performed most of the publicized re-identifications, and many of those re-identified were public figures.

Organizations sharing de-identified information should assure that they do not leave quasi-identifiers in the dataset that could readily be used for re-identification. They should also survey for the existence of linkable databases. Finally, organizations may wish to consider controls on the de-identified agreements that prohibit re-identification, including click-through licenses and appropriate data use agreements.

Appendix A Glossary

Selected terms used in the publication are defined below. Where noted, the definition is sourced to another publication.

aggregated information: Information elements collated on a number of individuals, typically used for the purposes of making comparisons or identifying patterns. (SP800-122)

confidentiality: "Preserving authorized restrictions on information access and disclosure,

---

[65] El Emam et al., Evaluating the risk of patient re-identification from adverse drug event reports, BMC Medical Informatics and Decision Making 2013, 13:114 http://www.biomedcentral.com/1472-6947/13/114

[66] Jaewoo Lee and Chris Clifton, Differential Identifiability, KDD '12, Aug. 12-16, 2012. Bejing, China.

[67] Khaled El Emam and Cecelia Álvarez, A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques, International Data Privacy Law, 2015, Vol. 5, No. 1

including means for protecting personal privacy and proprietary information."[68]‖(SP800-122)

Context of Use: The purpose for which PII is collected, stored, used, processed, disclosed, or disseminated. (SP800-122)

data linking: "matching and combining data from multiple databases." (ISO/TS 25237:2008)

De-identification: "General term for any process of removing the association between a set of identifying data and the data subject." (ISO/TS 25237-2008)

De-identified Information: Records that have had enough PII removed or obscured such that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual. (SP800-122)

direct identifying data: "data that directly identifies a single individual." (ISO/TS 25237:2008)

Distinguishable Information: Information that can be used to identify an individual. (SP800-122)

Harm: Any adverse effects that would be experienced by an individual (i.e., that may be socially, physically, or financially damaging) or an organization if the confidentiality of PII were breached. (SP800-122)

Healthcare identifier: "identifier of a person for exclusive use by a healthcare system." (ISO/TS 25237:2008)

HIPAA Privacy Rule: "establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically." (HHS OCR 2014)

identifiable person: "one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." (ISO/TS 25237:2008)

identifier "information used to claim an identity, before a potential corroboration by a corresponding authenticator." (ISO/TS 25237:2008)

Limited data set: A partially de-identified data set containing health information and some identifying information including complete dates, age to the nearest hour, city, state, and complete ZIP code.

Linkable Information: Information about or related to an individual for which there is a possibility of logical association with other information about the individual. (SP800-122)

[68] 44 U.S.C. § 3542, http://uscode.house.gov/download/pls/44C35.txt.

Linked Information: Information about or related to an individual that is logically associated with other information about the individual. (SP800-122)

Obscured Data: Data that has been distorted by cryptographic or other means to hide information. It is also referred to as being masked or obfuscated. (SP800-122)

personal identifier: "information with the purpose of uniquely identifying a person within a given context." (ISO/TS 25237:2008)

personal data: "any information relating to an identified or identifiable natural person ("data subject")" (ISO/TS 25237:2008)

Personally Identifiable Information (PII): **—**"Any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."[69] (SP800-122)

PII Confidentiality Impact Level: The PII confidentiality impact level—low, moderate, or high—indicates the potential harm that could result to the subject individuals and/or the organization if PII were inappropriately accessed, used, or disclosed. (SP800-122)

Privacy: "freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual." [ISO/IEC 2382-8:1998, definition 08-01-23]

Privacy Impact Assessment (PIA): "An analysis of how information is handled that ensures handling conforms to applicable legal, regulatory, and policy requirements regarding privacy; determines the risks and effects of collecting, maintaining and disseminating information in identifiable form in an electronic information system; and examines and evaluates protections and alternative processes for handling information to mitigate potential privacy risks."[70] (SP800-122)‖

Protected Health Information:

Pseudonymization: "particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms." [ISO/TS 25237:2008]

Pseudonym: "personal identifier that is different from the normally used personal identifier." [ISO/TS 25237:2008]

---

[69] GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, http://www.gao.gov/new.items/d08536.pdf

[70] OMB M-03-22.

Recipient: "natural or legal person, public authority, agency or any other body to whom data are disclosed." [ISO/TS 25237:2008]

Appendix B Resources

B.1 Official publications

AU:

- Office of the Australian Information Commissioner, *Privacy business resource 4: De-identification of data and information*, Australian Government, April 2014. http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf

EU:

- Article 29 Data Protection Working Party, 0829/14/EN WP216, Opinion 05/2014 on Anonymisation Techniques, Adopted on 10 April 2014

ISO:

- ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. Geneva, Switzerland. 2008. This ISO Technical Specification describes how privacy sensitive information can be de-identified using a "pseudonymization service" that replaces direct identifiers with pseudonyms. It is provides a set of terms and definitions that are considered authoritative for this document.

UK:

- UK Anonymisation Network, http://ukanon.net/
- Anonymisation: Managing data protection risk, Code of Practice 2012, Information Commissioner's Office. https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. 108 pages

US:

- McCallister, Erika, Tim Grance and Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, Special Publication 800-122, National Institute of Standards and Technology, US Department of Commerce. 2010.
- US Department of Health & Human Services, Office for Civil Rights, Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, 2010.
- Data De-identification: An Overview of Basic Terms, Privacy Technical Assistance Center, US Department of Education. May 2013. http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf

- Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.

B.2 Law Review Articles and White Papers:

- Barth-Jones, Daniel C., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now (June 4, 2012). Available at SSRN: http://ssrn.com/abstract=2076397 or http://dx.doi.org/10.2139/ssrn.2076397
- Cavoukian, Ann, and El Emam, Khaled, De-identification Protocols: Essential for Protecting Privacy, Privacy by Design, June 25, 2014. https://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identifcation_essential.pdf
- Lagos, Yianni, and Jules Polonetsky, Public vs. Nonpublic Data: the Benefits of Administrative Controls, Stanford Law Review Online, 66:103, Sept. 3, 2013
- Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: http://ssrn.com/abstract=1450006
- Wu, Felix T. Defining Privacy and Utility in Data Sets, University of Colorado Law Review 84:1117 (2013).

B.3 Reports and Books:

- Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015.
- Emam, Khaled El and Luk Arbuckle, *Anonymizing Health Data*, O'Reilly, Cambridge, MA. 2013

B.4 Survey Articles

- Chris Clifton and Tamir Tassa. 2013. On Syntactic Anonymity and Differential Privacy. *Trans. Data Privacy* 6, 2 (August 2013), 161-183.
- Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.
- Ebaa Fayyoumi and B. John Oommen. 2010. A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. *Softw. Pract. Exper.* 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12 http://dx.doi.org/10.1002/spe.v40:12
- Fayyoumi, E. and Oommen, B. J. (2010), A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. Softw: Pract. Exper., 40: 1161–1188. doi: 10.1002/spe.992