De-Identification Guideline

De-Identification Guideline

by

Leah Krehling

Western Information Security and Privacy Research Lab

Department of Electrical and Computer Engineering

Western University

Technical Report: WL-2020-01

Executive Summary

There have recently been reports of de-identified data sets failing at preventing an individual's personal information from being revealed. To discover the cause of these failures a review of 60 cases of re-identified data sets claimed to have been de-identified was performed. From this review, it was found that in majority of the cases the re-identification may have been preventable if more advanced methods from the literature had been implemented by the data custodians. In many cases the methods of de-identification used had previously been shown to be ineffective in preventing an attacker from re-identifying individuals. In some cases the implementations of de-identification methods were such that there was an exploitable vulnerability created.

In response to the re-identification attacks the goal of this paper is to collate the information gained from studying them and lay out the methods and ways in which de-identification must be performed in order to avoid the failures observed.

From the review of these attacks some main points about performing de-identification can be made. De-identification requires more than the removal of names and other information directly linked to someone's identity. When de-identifying data, external information and its ease of access must be considered. This external information gives the attacker an avenue to link with the de-identified information and uncover identities. It is also important consider what information can be inferred about an individual from the information within the data set.

De-identification will change the data and thus alter the data utility. Therefore implementing a measure of how much the chosen de-identification methods will change the data utility is also important. De-identification requires a privacy measure, used to determine the minimum level of privacy afforded to individuals. At the same time a measure of the utility of the data, used to determine the maximum amount of alteration that can be performed before the data is not usable for the intended research, should be determined. This utility measure could assist in determining if the data should be released for analysis.

From the study of these re-identification attacks it was found that if a data set is to be de-identified, more needs to be done than the suppression of directly identifying values. The field of de-identification research has a variety of methods which have yet to be broken when implemented on real world data sets. The apparent cause of failures in de-identification currently do not lay with the theories.

Contents

Publication Note

Cite this report as:

L. Krehling, "De-Identification Guideline." Western Information Security and Privacy Research Laboratory Technical Report WL-2020-01, Western University, Canada, 2020.
Available online: `https://whisperlab.org/technical-reports/de-identification-guideline-WL-2020-01.pdf`

1 Introduction

Data has become a commodity of the modern world. Both companies and research groups want to use the vast amounts of the data that has been collected from people to create innovative solutions, perform ground breaking research, or optimise their designs. A common use of this data is to train machine learning algorithms with it to make them better predictors of customer behaviour. For example, companies like Amazon, Apple, Google, and Microsoft [10, 59, 60] have all fed voice data to their software based assistants (Echo, Siri, Google assistant, and Cortana) to teach them to better understand human speech patterns and accents. Voice data is not the only type of data that can be leveraged and there are many different data types and structures that can be used. Some examples being geolocation data, cell phone records, social media networks, medical data, and browsing data. The data collected is diverse in type and form, making even single data sets complex, as they can contain any combination of this variety.

Much of the data that has been collected has the potential to reveal the identity of the people it is about. The information would initially contain personal details, like data about online shopping habits that enters a database connected to an account and thus contains the name, email address, a phone number, credit card information etc. of the customer. Due to this, to protect people's privacy, many countries have legislation in place to limit the collection and control the distribution of data about its citizens. For example, Canada has the federal Personal Information Protection and Electronic Documents Act (PIPEDA) [74], which places protections on personally identifiable information collected about Canadians.

De-identification refers to the processes used to separate someone's identity from the data collected from them to prevent their identity from being revealed through observation and analysis of the data, or linking data sources together. The idea of data de-identification then is to reduce the risk of connecting the data to the originating individual to a statistically insignificant amount. Under legislation such as that in Canada, data that has been de-identified is no longer considered protected personal information because of the idea of there being a statistically low risk of exposure [74].

There is a growing view that the risk assurances of de-identification are flawed, and that de-identification does not work [54]. Several notable cases have shown that supposedly de-identified data could still be used to re-identify individuals [39]. Cases of this occurring cast doubt on whether de-identification can protect an individual's privacy. Testing the accuracy of this view is a goal of this paper as well as the work discussed in Section 5 on prior work.

1.1. De-identification Considerations

When a data custodian intends to release de-identified data sets considerations about the type of data, its intended use, how to determine the balance between data utility and privacy, how to measure the privacy levels, and other important questions need to be asked in order to systematize and standardize the de-identification. These considerations can be categorized as:

Taxonomy: How to define key concepts (e.g. re-identification risk) in way that is flexible and cuts across application domains?

Methodology: What are the appropriate assessment techniques and controls? What are the business, privacy and security requirements?

Standards: What guidance should exist with respect to development, procurement, testing and compliance? How is the privacy risk of a particular data set managed?

Process: What are the quantitative and qualitative metrics used in risk determination, and privacy impact assessments? What are the governance and ethics considerations?

Oversight: How should the use of a data set be monitored throughout time? Who is accountable?

Knowledge Transfer: How should data scientists, engineers, security consultants, etc. be trained to work with de-identified data to support privacy goals?

A goal of this paper was to enable individuals to make informed decisions regarding these questions by providing information on the key concepts of de-identification and insight into the field through research into methods that have been tested on real data sets.

1.2. De-identification Methods

The field of research into different methods of lowering the re-identification risk is quite broad. Not only are there are many different forms of data but each data set is different in its collected attributes. For example, geolocation data sets, one from a vehicles GPS system and another from a social network site. The data set generated from the vehicle might contain only two pieces of information, the time and position coordinates. The social network might contain the same time stamp and coordinate information, but also the user's name, age, the content of their post on the site, and other information outside of the scope of what one would consider geolocation data. This extra information can help inform researchers that intend to use the data to learn something about the populace. For example, the social media data could be useful to the planning of public transport routes by providing the movement patterns of specific age groups. Due to this variance between the data within sets of the same "type" often data sets need to be considered individually. The unique combination of features in a data set could reveal more information than the features of another data set would.

De-identification literature typically splits data features into two types: direct identifiers and indirect identifiers. Direct identifiers include anything that on its own can be tied to an individual's identity, a name, social insurance number, account number, etc. These are values that are unique to an individual and can act as their identity in some contexts. An indirect identifier is a value that on its own is not enough to identify a person; date of birth, gender, race, etc. However, when used in combination they create indirect identifier sets that can be used to identify a person. These can also be called quasi-identifiers. A third category could be non-identifiers, values that cannot identify a person, though it can be difficult to draw the line between them and indirect identifiers, as generally the larger the indirect identifier set the easier it becomes to identify an individual, even when the information seems Inconsequential. Thus, except for cases where the attribute is not externally available for an adversary to exploit, the line between non-identifier and indirect identifier is difficult to set.

Due to the variance in data types and data sets different actions need to be taken to de-identify data sets. These actions can be generally grouped into six types of manipulations: suppression, masking, generalization, perturbation, aggregation, and access control and monitoring.

Suppression.

Suppression is the removal of some aspect of the data set entirely. This could include completely removing a field or column in the data, such as the individual's name, or removing specific entries in the data set. Outliers in the data set that are too extreme might be suppressed for a variety of reasons, one reason being that the uniqueness of any outlier makes that individual more likely to be identified in the data set. Providing the same privacy assurance to the outlier as all other individuals in the set could require manipulations that reduce the research utility of the data set too far, as the section on generalization explains. Thus removing the outlier can be the more prudent choice.

Generalization.
Generalization is a decrease in the granularity of the information, resulting in the information being less specific. This could be changing the time stamp from the second that it occurred to the minute, to a five minute interval or more as required. It could also be changing the location from the exact address to the city block, or general neighbourhood. The chosen generalization level for a feature in the data set is applied to that feature for every entry in the set. The risk with generalization is that increases in generalization are also decreases in utility. The greatest amount of privacy that could be provided to individuals would be to make all of the entries the same, or not include any at all, which would not provide much information for researchers to use. For example, if every location is generalized to the country but the researchers need to know which provinces or even townships an illness is most prevalent in the data has been too generalized to be useful to them.

Masking.
Masking is when the data is left in the set, but is obscured so that the original values cannot be readily obtained. This is usually used on direct identifiers. Sometimes this is done to make it easier for researchers to identify entries that have been made by the same individual, creating a perpetual identifier used through the data set that allows them to trace a single person. For example, in cell phone records for many users, every individual would require a pseudo-anonymous identifier so that the logs for the same person can be grouped together. Or, depending on the intended usage, the data may need to be able to be linked back to the original individuals at some later time and so a value connecting the entry back to the data set containing their identity is required. This can be done using techniques like encryption or, in limited cases, hashing on the original value.

Perturbation.
Perturbation is to alter the reliable accuracy of the value, typically seen as adding noise to the system. This is not generalization, as the granularity does not need to change, but instead the value itself is slightly altered. An example is shifting GPS coordinates randomly by a small amount so that the exact location cannot be assumed to be reliable. The goal is not to alter the result of any analysis but the reliability of a single value so that it cannot be used to concretely learn someone's identity.

Aggregation.
Aggregation is the process where raw data is collected or grouped together. In some cases only statistics about the data may be released, in others attributes or entries may be combined. This way instead of revealing the entire data set collections of statistics derived from the data set or information about small groups within the data set can be revealed. For example, if analysis of the raw data shows that 50% of men over a certain age living in a specific township have a disease that is being studied, only that statistic would be released. For geolocation data this could also mean showing the popularity of certain locations rather than the actual mobility traces of the individuals, or grouping multiple traces together into average movement patterns.

Access Control and Monitoring.
Access control and monitoring is the limitation of access to the data set. For this report this will be considered anything that actively limits how someone can access the data, or shields the data in some way, as opposed to simply a signed agreement to not misuse the data upon receiving it. Things like the data only being queryable instead of allowing access to the raw data, or limitations on who has access to the data, or parts of the data. The data set might also have monitoring on it that records who is accessing it when, for how long, a record of queries made to the data, etc.

2

The Threat of Re-Identification

The purpose of de-identification is to strip the personal identity of the data source from the data itself. This is done so the data can be passed to researchers or made publicly available with minimal risk to individual privacy. In encryption literature there is the concept of encryption being broken. Essentially if currently available computers could guess the encryption key within a relatively short amount of time, then the encryption is not secure. However if it would require more computing power than is currently conceivably possible the encryption can be considered secure. A similar concept is used in de-identification research. “Factual Anonymity", sometimes referred to as “Practical Anonymity" says that if it would require an excessive amount of time, expertise, manpower, and expense to re-identify individuals of the data set, then the set can be considered “factually anonymous" [33]. From practical anonymity it becomes clear that the study of re-identification attacks should take into account the expertise required of the attacker, as well as the information that they have available to them or the cost of the type of data that might be required to link to someone's identity.

There are some commonalities between re-identification attacks. Typically they require: information outside of the de-identified data set to match with individuals within, an individual that then appears in both data sets, common information about the individual in both data sets, and enough information to be statistically sure that a match is correct [2].

2.1. Identity Disclosure

Identity disclosure is full re-identification of one or more individuals within the data set. The attacker has somehow re-identified these individuals despite the de-identification efforts that were made to prevent this. There are many ways this can occur, and the accuracy of the re-identification can vary depending on the attack method and original data set. Though it should be noted that the miss-identification of an individual as someone in the data set can also be harmful to them as many of the negative impacts of identification could still occur to them in real life despite the inaccuracy of the attack.

2.2. Attribute Disclosure

This can also be referred to as “Homogeneity Attacks" or “Inference Attacks" when discussing k-anonymity, which is described in Section 3.1. The idea of this type of privacy loss is that the attacker learns something about the individual that is not public knowledge without fully re-identifying their entry within the data set. For example, if an attacker is trying to discover an individual's illness and has access to health data they know to contain that information, then they are likely able to narrow down the possibilities of who the individual is. Say they are looking for a male age 40, removing anyone who does not meet this requirement might leave them with 10 possible individuals. If all 10 individuals have the same illness, then the attacker has learned what they wanted to without ever re-identifying the exact entry of the person they sought information on.

2.3. Internal Attacks

These are attacks that originate from within the data set. The attacker starts with the data set and attempts to find unique individuals within the set. This will vary depending on the type of data. In geolocation data,

an attack will look for unique movements or outliers. Once individuals are isolated, the attacker will take the information available and compare to public information to identify someone. With geolocation data this could mean analyzing the mobility traces to find a likely home address that can be searched in public databases, thus providing them the identity of these individuals.

2.4. External Attacks

These are attacks that originate from outside of the data set. An attacker might be someone who knows that a particular individual is within the data set and seeks them within it. It could be someone who knows the individual personally or, in the case of public figures, simply has background information on the individual. In this case the attacker would access the database and use the background information to isolate the individual they are seeking. For geolocation data if an attacker knows where an individual was at a specific time, they could use that to isolate their mobility trace and learn their movements over the entire time coverage of the data set.

3

De-Identification Measures

There are two main measurements of privacy in the literature that can be used to determine how much privacy is being provided for individuals in the set. These are k-anonymity and differential privacy. k-anonymity looks at the uniqueness of every individual in the data set, while differential privacy balances the amount of accurate information revealed with the error created in the analysis of the data.

3.1. k-anonymity

k-anonymity is a de-identification measure that was developed by Sweeney [64]. The general idea is that every individual should be indistinguishable from $k-1$ other individuals in the data set when $k > 1$. To make individuals indistinguishable, information is typically either suppressed or generalized. Any generalization to an attribute must be uniform throughout the data set. It also might require the removal of extreme outliers from the data set as they may be so unique that generalizing the entire data set would remove too much information. By doing this it becomes less likely that an attacker could to isolate a single individual in the data set.

There are also multiple extensions of k-anonymity; ℓ-diversity [42], p-sensitive k-anonymity [73], and t-closeness [36]. These extend the requirements of k-anonymity based on examinations of when the asserted k-anonymity promise of privacy might break. The methods of these papers should be considered when looking at what constitutes best practice.

An issue with k-anonymity is that since alterations to the data set must be made computational resources are required to perform these operations. Though much of the manipulation could be automated, often manual manipulations or checks on the anonymity level are still made, and the alterations could make the data significantly less useful. By generalizing the data the caretaker could make the data too generic to be useful for the purpose of the release and would then have no reason to release the data.

3.2. Differential Privacy

Differential privacy was introduced by Dwork et al. [14]. The basic idea is to balance the level of privacy provided to individuals in the data set with the accuracy of the data. The privacy is created by introducing unreliability to the singular attribute entries by adding noise to the values of the attributes. If the noise is calibrated carefully then the results of analysis will be within acceptable error of the unchanged data so the result of the intended analysis of the data is unharmed. However, anyone trying to identify someone will be prevented from doing so because there will be no way to tell if the values are correct so long as the noise that is introduced is not predictable.

The exact method of introducing noise is up to the data custodian to determine. As well as the level of privacy and error. Differential privacy is often discussed when dealing with query systems, as the noise can be added to the data at the time of the query and adjusted according to the specific query.

An issue with differential privacy is that, similar to k-anonymity, it is difficult to say what value the privacy/accuracy level should be set to. If too much noise is added to the data then the analysis might be unreliable, and the amount of noise that is added depends on the values themselves. If there is little variance in a value across all the data entries, then only small amounts of noise can be added. It requires knowledge of the data, and an understanding of the attributes themselves to tune the noise. This also means that each feature

of the data set may require individual tuning. In a data set that has many features this can be computationally and time intensive. As well, even with the balanced noise, the more queries that are performed the more data that is released, and potentially enough data is released to re-identify someone if the queries are not limited.

4

Tools of Re-Identification

Though this list is not exhaustive it provides a basic overview of the common methods and tools that attackers will use to identify people within a data set.

4.1. Linkage Attacks

These are attacks that are performed by connecting the information from two data sets. It is a very common attack, particularly when public records of names and addresses are available to the attacker. This could also involve linking two de-identified data sets that originated from the same raw data (or not) to gain enough information to identify the individuals that appear in both data sets [63]. These attacks can be done manually or be automated in various manners.

4.2. Machine Learning Models

Common with internal attacks, machine learning models can be trained to analyze the data and determine information that can be used to identify individuals. Geolocation data can be analyzed with clustering models to find an individual's most likely home address or workplace [13]. It could also be taught an individual's patterns and then used to find that individual again within other data sets that are released [20].

4.3. Markov Chains

A Markov Chain is a statistical model. Specifically, it is a model that describes the probability of the next state of the system based on its current state. Markov chains are often used with geolocation data to create statistical models of an individuals movements. It could be used to make predictions of where an individual will be and when for an attacker to identify them in person [68].

4.4. Reversing Masking

This depends on the type of algorithm used to mask values within the data set. In some cases it is possible that the actions performed on the original data can be reversed and so the original value can be discovered by the attacker [8]. If masking is done by performing reversible mathematical operations on the raw data the attacker could discover the steps of alteration and reverse them [35]. Or if hashing was performed it might be possible for an attacker to brute-force match the hash with its originating value [56]. Even encryption could be vulnerable depending on the type and its implementation.

5

Prior Work

A systematic review was performed of the literature demonstrating successful re-identification attacks on data sets that had been claimed to be de-identified. The papers and articles studied came from various fields of research, including statistics, computer science, and health informatics. Research performed by journalists for articles was also examined.

The final review looked at 60 total cases, 18 of de-identified health data, 23 of geolocation data, 6 of demographical data, 3 of social media data, 2 of billing information, 3 of browser/browsing information, 2 of movie ratings/reviews, 1 of educational information, 1 of call records, and 1 of homicide case files.

From these works observations were made of the causes of failure in the de-identification performed on the data sets. In majority of the cases the researchers found that the re-identification was preventable if more advanced methods from the literature had been implemented by the data custodians. In many cases the methods of de-identification that were used had previously been shown to be ineffective in preventing someone from identifying individuals. In some cases the implementations of methods of de-identification were such that there was an exploitable vulnerability created.

In response to these vulnerabilities in de-identification the goal of this paper is to collate the information gained from these findings and lay out the methods and ways in which de-identification must be performed in order to avoid the failures observed.

6

Best Practices for All Data Types

There are some aspects of de-identification that are not dependant on the type of data that is being de-identified. The actions that need to be taken on any data set intended to be de-identified are laid out in this section.

6.1. Suppression

When determining what data should be suppressed it is important to look at what single elements of the data set are unique to the person. There are many identifiers or personal information that custodians know to remove, such as a SIN number, credit card number, account numbers and the like. A name is an obvious identifier that people will remove, however an email address is likely more unique than a name. There might be many John Smith's but there is only one `johnsmith@gmail.com`. Other examples of identifiers that might be overlooked are phone numbers, physical addresses, IP addresses, and unique hardware identifiers such as MAC addresses. In many cases these are unique to the person and so all need to be removed to protect identities [49].

When determining if a value is a direct identifier the uniqueness of the value and the ability of someone to search for that value are considered. In some cases there is no publicly available data to compare to, for example, if a company uses a unique number to reference a user that has been randomly assigned. Releasing that information likely has no way of leading an attacker back to the person's identity. However, how that unique value was created can determine whether it is an identifier as well, see Section 6.3. On the other hand, values like phone numbers can be overlooked though these have been proven to be trivially re-identifiable in a study of telephone metadata [49].

When suppressing data the it is important to note all of the places that the information appears in. When looking at the personal genome project the information was supposed to have direct identifiers such as names removed, however the researchers found that because the data was supplied by the users some profiles still contained names in unexpected areas, and many of the downloadable files associated with the profile had a filename that included the name of the person [67].

A summary of the best practices of suppressing data:

1. Consider if the value is unique to the person to determine whether it is a direct identifier,
2. Ensure that none of the information seen by the attacker contains the direct identifiers you intend to suppress,
3. Consider whether the information that is being suppressed can be found or inferred from any of the values remaining in the data set,
4. Create a limit of utility to use to determine when the alterations required on the data have rendered its utility too low for the intended research.

6.2. Generalization

When determining how to generalize data first a set of indirect identifiers should be recognized within the data set. This can contain all of the attributes of the set. Then the privacy level that the de-identification should provide to the individuals should be set. Finally a utility level should be set that determines how general the attribute can be before it is no longer useful for the intended research should be determined.

Though it can be difficult to know exactly what information is available to an attacker to leverage it is important to be aware of what information is available to be linked to the released data set. It has been noted that the difference in available information between regions will significantly alter the ease with which an attacker can re-identify individuals [61].

It is important to consider the distribution of values within the data set when generalizing. If the data contains individuals mostly between the ages of 30 and 50, then users outside of that range will be more unique than those inside, and thus have less privacy in the data set [26, 63]. Using a k-anonymity measure explained in the 3.1 section can help to remove this difference in privacy levels, and control how much privacy a user has.

Another consideration should be made for data that has been collected over time. When generalizing time data the difference between 1 year and 2 years might be large in terms of privacy loss, but 10 years to 11 years will not have the same impact on privacy loss. Thus when generalizing time frames large time frames need to be reduced more in order to significantly lower the privacy loss [32].

A summary of the best practices of generalizing data:

1. All quasi-identifiers must be considered together as a set when performing risk assessment,
2. Use a k-anonymity measure to control the amount of generalization being done on the data and measure privacy level,
3. Generalizations should be done to a greater degree when the data is from the same population over time,
4. Reducing a data set over a time span should be proportional, the larger the time span the greater the time alteration needs to be for a significant change in risk to occur,
5. Have an awareness of the types of large organized data sets about the populace that have been made available and consider that in the risk assessment.

6.3. Masking

Though the reversing of masking is not a common method of re-identification, as it typically requires more skill or specialized knowledge than other methods, bad masking practices can be a serious privacy risk. When masking is broken the original value is revealed, which could be the full name of the individual, an identifying number [35], or an account number [8].

There are different methods of masking that can be performed and it is important that when performed they are implemented correctly, used on data appropriate for their intended purpose, and any extra requirements of their standard use practice are performed. The two most common methods of masking are hashing and encryption. Though often seen as similar there are important differences in their implementations and best practice. An encryption algorithm takes the plain text and then using a key creates a cipher text value that that can then be returned to the original plain text value using either the same or another key. While a hashing algorithm will take in the plain text and output the hashed value.

There are some principles of thought that both hashing and encryption algorithms share. For one the cipher text or hashed value that is output should not share similarities to one that came from a similar plain text value. That is to say if the plain text is 12345, the cipher text should be no more similar to that received from 12346 than any other. Another is Kerckhoffs' Principle, which states that "A cryptosystem should be secure even if everything about the system, except the key, is public knowledge". That is to say a hashing algorithm should be one in which the entire structure of the algorithm can be revealed and the attacker still cannot reverse the hash to plain text. For encryption this means that the only secret required should be the key. For example, AES, RSA, and TripleDES are all well known encryption algorithms the are explained in thorough detail online and are still secure. As well, a masking algorithm should also be such that observation of the output does not allow an attacker to guess any of the secret values used [35].

Hashing has unique qualities, the first of which is string length uniformity, the output will be the same size no matter the size of the original plain text that was used. For instance SHA-256 is a hashing function that will take in plain text of any length and always output a string of 256-bits long. This can be important as the length of output changing with the length of the input can reveal information to the attacker [35]. Hashing functions have output uniformity, meaning that the output hashed value will be the same every time the same plain text is sent through the algorithm. Hashing functions should also be mathematically infeasible to take the hash value and reverse the operations to return it to the plain text value.

The main aspect of an encryption algorithm is that it is designed to be mathematically reversible. The output cipher text of an encryption algorithm can be returned to the original plain text through the algorithm with the correct key. This key can be either the same as was used to encrypt the plain text (symmetric key), or a different key (public key). Encryption algorithms all have different key lengths, the Advanced Encryption Standard (AES) has typical key lengths of 128, 192, or 256-bits. The key length can affect the computational cost of the algorithm, but it also affects how long it would take an attacker to brute force attack the algorithm.

When using hashing functions best practice is to add a random salt to the value. This makes attacks like dictionaries and rainbow tables more difficult to implement against the hash. Hashing also becomes less secure when the possible input is of a limited structure. This means if the inputs are all structured the same, have same length, are comparatively small, some digits are constant or have a small range, then an attacker will potentially be able to brute force the hash. For example, taxi medallion numbers are all 4-6 characters with specific characters being letters and the rest numbers. Overall there are 22M possible plain text values and outputs of the hashing function or 2 minutes of computation time [56]. Another thing that should be done is looking at the data itself to ensure that there are no errors that may reveal information. If data for a few entries was imported incorrectly and are all 0 for instance once hashed these values can create an anomaly that will reveal information to an attacker [56].

When using encryption one of the most important things to remember is that the key needs to be kept secret. Whether using public or symmetric encryption the key that is required to decrypt the cipher text must not be stored along with the cipher text values, or hard-coded into the system performing the encryption. Doing so would provide attackers the information that they need to reverse the masking.

A common error with masking is not using the appropriate algorithm, or not using the appropriate algorithm according to best practice. For example, if you have an ID number that needs to be encrypted, performing just any mathematical algorithm on it is not the same as performing encryption. Adding, subtracting, multiplying, or dividing the ID number by constants will change the ID value and is reversible for the data holder, but it is also reversible for the attacker. These constant values can be discovered from analysis of the ID numbers and knowledge of the original structure of the number. For instance, the manipulations described will result in identifiers that have a different number of digits depending on the input value, comparing these different length identifiers to the original can give an attacker information they require to reverse engineer the algorithm [35]. Masking algorithms should all follow Kerckhoffs' principle.

Something that should also be considered is whether the information needs to be masked at all. Though it is necessary if the information needs to be tied back to the original individual by the data holder if all that is required is a unique identifier to be carried through the data set that does not contain private information it is more secure to use a value that is not derived from the individual at all. A completely random string or number to replace the attribute is a more secure identifier to carry through the data set than one generated by a true value. As well this would be less computationally intensive for the data holder than implementing encryption. It is important that this number not be derived from a value belonging to the original individual as seeded pseudo-random generators can be reversed depending on their implementations [7, 8].

A summary of the above best practices of masking data:

1. Masking requires mathematical operations whose secret values cannot be reverse engineered,
2. Encryption or hashing algorithms of fixed length output should be used,
3. Weed out anomalous data or errors in the data before release,
4. When hashing is performed, random salting should also be done,
5. Hashing is not ideal for use on inputs with known limited structures,
6. Encryption keys need to be secret and secure,

7. If no reversal is required a completely random identifier should be considered for creating a perpetual identifier.

7

Demographic Data

Demographic data is not specific to a field of research. Often when collecting data for any reason there are standard values about the person that get collected as well to provide further information to researchers about the trends in the data. Examples of these types of data would include age or date of birth (dob), gender/sex, race/ethnicity, home address, name, education level, etc.

This section details information about data sets that were de-identified using the demographic data contained within them. The data sets themselves were varied in their content otherwise, some were data sets of only demographic data, others contained health records or other information on the individuals as well.

7.1. Suppression

When protecting demographic data there is typically some suppression required, as some of the information will be direct identifiers. The rules of suppression explained in Section 6.1 apply to this type of data as well. The main idea being that any value unique to an individual should be considered as a potential direct identifier and suppressed if it can identify an individual. Names, full home address, and other things that are unique to the person need to be completely removed to protect identities [49].

7.2. Generalization

The indirect-identifier sets of Table 7.1 show how unique a set of values can be. Once the direct identifiers are suppressed the indirect-identifier set should be studied. These are very important for privacy, according to one study 99.98% of people in Massachusetts would be correctly re-identified in a data set containing any 15 demographic attributes [58].

The studies researched looked at populations from the US, Canada, Netherlands, and Germany, and used different types of attacks to re-identify the data. In some cases a secondary data set was used to find identities, in others the uniqueness of a person within the set was used and if the person was completely unique this broke the privacy guarantee stated by the custodians and was considered a successful re-identification attack. Though the census information may not be available in all places, with enough knowledge of an individual through personal experience or looking for information online it could be possible to find them within the de-identified set. Some of the details of the studies of Table 7.1 are discussed in Section 6.2.

That is not to say that this information cannot be kept in a data set together. Generalization of these values can increase the privacy of the data set. The Canadian study into the identifiability of people in Montreal shows that altering the date of birth to the year makes less of the population unique. The same occurs when less information on someone's historical postal code record is released. From the conclusions of the study changing the date of birth to month and year of birth, as well as altering the postal code to only the first 3 digits reduces the uniqueness to a "very low value" [32].

When looking at the indirect identifier set the external data sources need to also be considered. In the US things like publicly released census data, voter lists, data brokers, and public tax registers can provide a lot of information for an attacker. Though their availability is varied state to state or on the case of property taxes, between municipalities. In one study on air quality the information on heath data was redacted heavily according to HIPAA standard [17], however the redacted data contained enough information to use computer

Table 7.1: Indirect-identifier sets of demographics

Study	Year	Country	List of identifiers	Linked Data	Re-id'd
[58]	2019	USA	ZIP code, date of birth, gender, number of children	Voter registry, public information	99.8%
[61]	2017	USA	Race, gender, date of birth, education level, year they moved into their residence, home ownership status	Property tax registers, data purchased from brokers	28%
[67]	2013	USA	date of birth, ZIP code, gender	Voter list, public records website	49%
[32]	2011	CAN	date of birth, postal code	Uniqueness	98%
[32]	2011	CAN	year of birth, postal code	Uniqueness	85%
[32]	2011	CAN	date of birth, 3 char postal code, gender	Uniqueness	80%
[32]	2011	CAN	Postal code trail of 2 years	Uniqueness	17%
[32]	2011	CAN	Postal code trail of 5 years	Uniqueness	35%
[32]	2011	CAN	Postal code trail of 11 years	Uniqueness	43%
[4]	2010	USA	County, gender, date of birth, race	Voter list	60%
[4]	2010	USA	County, year of birth	Voter list	10%
[4]	2010	USA	Gender, year of birth, race	Voter list	0.25%
[4]	2010	USA	Year of birth	Voter list	0.01%
[34]	2010	NLD	4 char postal code, gender, year/month of birth	public register data	4.8%
[34]	2010	NLD	Municipality, gender, year/month of birth	public register data	0.07%
[26]	2006	USA	Gender, ZIP code, date of birth	2000 Census data	63%
[64]	2002	USA	ZIP code, date of birth, gender	Voter list	N/A
[63]	2000	USA	Gender, ZIP code, date of birth	1990 Census Data	87%
[63]	2000	USA	Gender, municipality, date of birth	1990 Census data	53%
[63]	2000	USA	Gender, county, date of birth	1990 Census data	18%
[2]	2001	GER	Income, year of birth, sex, schooling, weekly work hours, occupation, region, time employed, time unemployed, duration of previous employment, marital status, number of children, nationality	Uniqueness	69%

inference methods to create race and gender estimations and the data contained date of birth. Then using property tax registers and information bought from data brokers residents could be re-identified [61].

Similarly the Personal Genome Project gathers genetic information on participants, but it was the demographic information that allowed for re-identification. The data contained date of birth, full zip code, and gender. The demographical information revealed 22% of people when linking to voter data set and 27% of people when linking to a public records website, for a total of 49% of records [67].

In the Netherlands, research into the identifiability of hospitalization and welfare records was investigated. These records are all available upon request from the governing body that collected them. Using public register data the researchers identified 4.8% of the people in the health care set from a shortened postal code (4 values of 6), gender, and year and month of birth. The welfare data contains information about investigations of welfare fraud; using municipality, gender, year of birth, and month of birth 0.07% of people were uniquely identifiable. The total percentage of citizens identifiable to a group of 10 or less was 2.14% within the data set [34].

A summary of the above into best practices of generalization:

1. Full date of birth, postal code, and gender cannot be left untouched in a data set intended to be de-identified,

2. Generalizing the date of birth to month and year of birth, as well as altering the postal code to only the first 3 digits significantly reduces the identifiability.

8

Health Data

Health data is typically information that would be in a medical record. It could contain information on prescriptions, illnesses, DNA sequences, medical procedures, or any other data that might be collected or created by doctors or nurses in a clinical setting. This data could also be from pharmacy records, insurance records, hospital discharges, or ambulance records. Some of the information collected by devices and apps that track fitness can also be considered health data. Health data is often protected by laws such as HIPAA in the United States [17] or PIPEDA in Canada [74] as well as various state or provincial level laws and standards.

If health data is released that contains information that can be re-identified there are many ways in which the individuals in the data set may be harmed. For example, personal information about the illnesses that they have had or are seeking treatments for could affect their current or future insurance rates. Previous diagnoses of mental illnesses could be used to negatively impact their current or future employment due to stigmas held by employers, as well as negative impacts to their personal lives.

Within the literature of re-identification of health data there are a few main methods that are employed. The first is to leverage the demographical information that is included within the medical records. This is information like age, sex, race, home address, etc and the risks created by this type of data was laid out in Section 7 of this document. Then there is the leveraging of newspaper articles and obituaries. These attacks use the nature of reporting on motor vehicle accidents and deaths that contain the person's name to match to records and find their identity. Other attacks have used unique alleles in genomes or the unique pattern of illnesses to reveal information. There has also been a case of the masking used on health insurance numbers being broken. The breakdown of attacks is displayed in Figure 8.1. This shows that of the 18 health data re-identification attacks that were studied 33.3% were performed by leveraging the demographic information, 11.1% leveraged vulnerabilities in the masking methods used, and the other 55.6% of attacks used information unique to medical data. This is the information that will be focused on in this section.

Table 8.1 contains the information about data that was released, what was leveraged, the levels of suppression and or generalization that was applied to the information, and how that affected the identifiability of the individuals. This is based on the half of attacks that did not solely use demographical information, those attacks were included in Table 7.1.

8.1. Suppression

Often some amount of suppression is required to meet the health data protection standards in the region. For example, the HIPAA standard requires names, social security numbers, insurance plan or group numbers, medical record numbers, medical device identifiers, biometric identifiers, or any other unique identifying number, characteristic, or code, except a code to permit re-identification of the de-identified data by the data custodian to be removed from the data set [17]. Similar expectations are laid out in other legal protections for health data at the federal and provincial/state level. The legislation of HIPAA was enacted in 1996 [17] and PIPEDA in 2000 [74], though not all US States have to follow the HIPAA standard based on its exemptions all of the data sets studied were released after these legislation's were enacted, though only 3 directly mention HIPAA in the study.

Table 8.1: Identifier sets of Health Data

Study	Year	Country	List of identifiers	Linked data	Re-id'd
[28]	2018	USA	Hospital records of vehicle accidents, year of accident, location of accident, patient age, patient sex	Newspaper articles on vehicle accidents	N/A
[53]	2018	USA	15-min aggregated physical activity data	Demographical data	94.3%
[53]	2018	USA	24-hr aggregated physical activity data	Demographical data	87%
[7]	2016	AUS	Date of hospitalization perturbed randomly by up to two weeks, year of birth, treatment	Uniqueness, public information	N/A
[7]	2016	AUS	Date of birth perturbed randomly by up to two weeks, date of child's birth perturbed randomly by up to two weeks	Uniqueness	N/A
[7]	2016	AUS	Breakdown of billing	Uniqueness	100%
[66]	2013	USA	Patient demographics, ZIP codes, diagnoses, procedures, attending physician, hospital, a summary of charges, how the bill was paid	Newspaper articles containing the word "Hospitalized"	43%
[19]	2013	CAN	Province, age at death, gender, and exact date of adverse drug event report	Obituaries	30.78%
[19]	2013	CAN	Age at death, gender, and exact date of adverse drug event report	Obituaries	5.05%
[19]	2013	CAN	Province, age at death, gender, and month and year of adverse drug event report	Obituaries	0.63%
[40]	2010	USA	Pattern of diagnosis codes	Hospital discharge data	96%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits	Hospital discharge data	96%
[40]	2010	USA	Pattern of diagnosis codes with least common 5% removed	Hospital discharge data	75%
[40]	2010	USA	Pattern of diagnosis codes with least common 15% removed	Hospital discharge data	25.6%
[40]	2010	USA	Pattern of diagnosis codes with least common 25% removed	Hospital discharge data	0%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits with least common 5% removed	Hospital discharge data	70.4%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits with least common 10% removed	Hospital discharge data	48.2%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits with least common 15% removed	Hospital discharge data	16.3%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits with least common 20% removed	Hospital discharge data	0.25%
[40]	2010	USA	Pattern of diagnosis codes generalized to 3 digits with least common 25% removed	Hospital discharge data	0%
[18]	2009	CAN	Sex, age (days), postal code (3-char), admission and discharge (day/month/year)	Uniqueness	> 20%
[18]	2009	CAN	Sex, age (weeks), postal code (1-char), admission (yearly quarter), length of stay	Uniqueness	33%
[43]	2006	USA	DNA information with familial connections	Online genealogy data from obituaries	70%
[45]	2004	USA	Genetic illnesses, hospital name	Hospital discharge data, census data	98%
[65]	2003	USA	Diagnosis, inferred ZIP, drug, dosage, refill	Ambulatory data, hospital discharge data, voter list	2.3%
[44]	2000	USA	Inferred gender and illness from DNA, hospital name	Hospital discharge data	98%

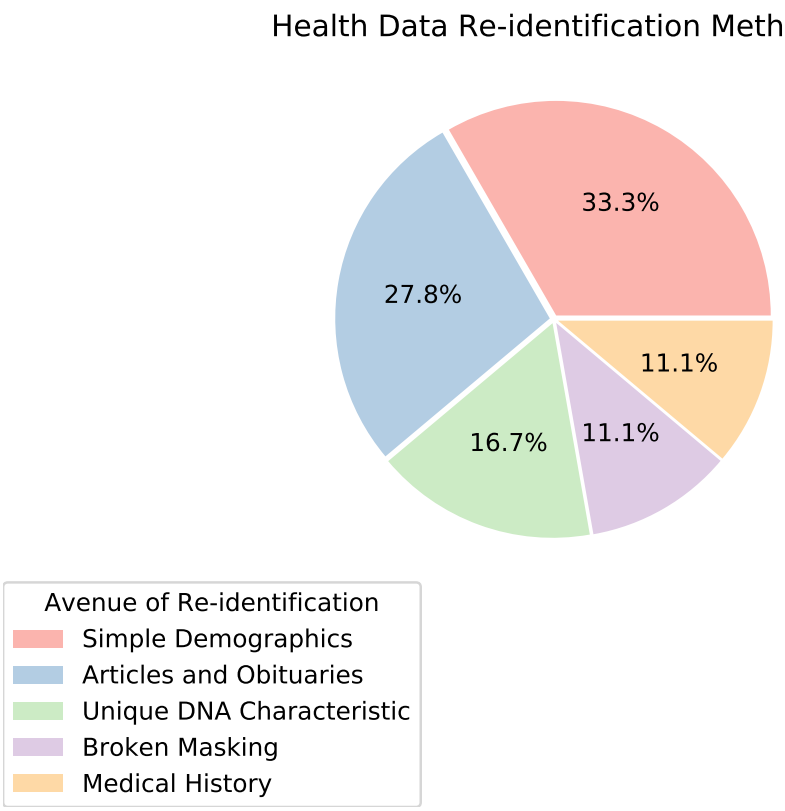


Figure 8.1: Breakdown of information used to re-identify health data

Though all of this information has been removed from these data sets patients can still be identified. For events that are reported on by local news stations, for example, motor vehicle accidents, assaults, fires, arrests, or other unique circumstances, the year and 3 digit zip code providing municipal location of the accident is enough to identify an article about it. News articles often contain the names of people involved especially those that were injured, thus the attacker learns their name. If the data utility does not require the location data then it is recommended to remove it, as this made finding the articles far less likely [28] [66].

In Washington state a data set containing health information on virtually all hospitalizations occurring in the state in a given year, including patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid, was available for $50. It did not contain patient names or addresses, only five-digit ZIP codes. Newspaper articles printed in the state for the same year that contain the word “hospitalized” often included a patient’s name and residential information and explained why the person was hospitalized, such as a vehicle accident or assault. 43% of the health records in the state could be re-identified using newspaper articles about the same events [66]. It should be noted this data was not released to HIPAA standard.

Similarly obituary data can be used to match medical records involving patients dying. Using publicly available data-sets from statistics and health Canada on adverse drug events it was possible to match deaths to obituaries in the newspaper. Disclosing the province, age at death, gender, and exact date of the report has quite a high risk of re-identification, but the removal of province brings down the risk significantly. By only generalizing the date of reporting to month and year and including all other variables, the risk is always low [19].

When dealing with DNA sequencing data it is important to recognize that this data alone can leak information. If the entire DNA sequence is revealed then the genetic sex of the individual is also known. Then due to the nature of genetic disorders and recent breakthroughs in research some diseases can be directly tied to the presence of a specific gene’s allele. It is possible to use this information to find patients that have these types of illnesses and match them to publicly released hospital records, using some demographic data (age, gender, generalized zip code) to confirm the match. Using hospital release records, census data, and DNA sequences released for research purposes patients with diseases such as cystic fibrosis, Huntington’s disease, and sickle cell anemia were 98–100% identifiable [45] [44]. This is one example of how removing information

in one area of the data set might not completely remove it from the attackers hands.

When removing information it is important to consider whether it can be inferred from the data that has been left. For example, in prescription data the patient's zip code was removed, however the pharmacy's zip code was included. Most people stop at a nearby pharmacy on their way home from an appointment. Thus a close approximation of the patient's zip code can be inferred about the patient from the information that was made available about the pharmacy [65].

In the table one study found that they could use a combination of suppression and generalization to remove the possibility of their attack succeeding. Their attack relied on the unique combination of illnesses and diagnosis a single patient may have over their medical history. The problem the researchers found was that to get to 0% identifiability they were suppressing 25% of the least common diagnosis codes from all the patients, and in doing so the information was assessed to be clinically useless for the intended research [40].

A summary of the above into best practices of suppression:

1. News-worthy events require their location information to be removed or strongly generalized,
2. DNA sequences contain information that may have been intended to be removed such as gender,
3. DNA sequences containing alleles for rare genetic based illness should be considered direct identifiers in many data sets.

8.2. Generalization

In all cases of generalization the balance of information and privacy can be difficult to maintain. The more detailed the data the more useful, the more general the data the more private is the basics of this balance. In 2009 when studying whether the prescription data over 18 months could be released for research purposes the Children's Hospital of Eastern Ontario (CHEO), ran a study into the re-identifiability of the requested information. The original request contained the following information generalized to the indicated level if at all: sex, age (days), postal code (3 characters), and admission and discharge dates (day/month/year), as well as the drug and diagnosis information. Using a k-anonymity threshold of 5 and thus risk of re-identification probability of 0.2 it was found that this did not provide sufficient protection of patient identities. When they considered the data usage concerns, the best solution was to raise the threshold of probability to 0.33 and provide: sex, age (weeks), length of stay (days), postal code (1 character), and admission date (quarter and year). To meet the risk threshold for some entries values had to be removed, for instance 11.3% of the age category would still need to be suppressed [18].

When dealing with diagnosis codes in patient records it was found that in a sample of more than 96% of the records are shown to be uniquely identified by their diagnosis codes with respect to an entire population of 1.2 million patients. This was found using ICD-9 diagnosis codes when looking at the re-identifiability of disseminated EMR data. ICD-9 codes are 5 digit diagnosis codes containing three-digit disease codes, followed by two possible digits of further specification. They found that for the majority of patients the set of ICD-9 codes was unique, and thus could be used to identify a patient's record in a set containing private information. Suppressing codes that appeared in less than 5, 15, and 25% patients records was performed as well as generalizing the codes by removing the 2 digit specification. Both of these failed to provide sufficient de-identification [40].

When DNA data is being researched the genealogy can identify the individuals. Often with DNA data researchers are looking at inherited illness. In many cases this means that the familial relation between the DNA sequences are released though identifiers are removed or generalized on the data set. By revealing the familial relationship between individuals however they can be identified. Using genealogy record sites and death records from newspapers a family structure can be built, and approximately 70% of the are unique. They can then be compared to the data revealing the family and then the individuals [43].

In a study from Australia discussed further in Section 8.3 researchers looked into billing records containing information on medical events, year of birth, date of event perturbed to two weeks, and other information. Researchers were able to identify Australian public figures by linking publicly available information about them to the medical records. In doing so they attempted to make the task more difficult by generalizing everyone's year of birth to 5 years, but found this had little effect on their uniqueness.

A summary of the above into best practices of generalization:

1. Sex, age, postal code, and admission dates are a indirect identifier set,

2. Patient diagnosis sets are unique and should be considered an indirect identifier set,

3. The structure of a person's familial tree can also be considered in indirect identifier.

8.3. Perturbation

The methods by which random noise can be added to a data set vary. There are different methods of generation and not all features necessarily require noise to be added to them. In Australia the public healthcare system released a data set containing billing information. Perpetual identifiers were used to identify the same patient across different records and other information included; year of birth, sex, medical events, codes indicating service provided or prescriptions given, the date, the location as State, the price paid, the breakdown of payment sources. The data was de-identified through suppression of some rare events and all dates were perturbed randomly by up to two weeks. Using publicly available information about well known Australians researchers could search for mothers using their date of birth and the birth dates of their children. By querying the data, with the error in reported dates accommodated for, the individuals were shown to be unique in many cases. This indicated that the indicated that the individuals were re-identifiable as they also proved there was enough information to further confirm identity within the medical record. They also found similar results from professional athletes and their known injuries, and news stories about politicians and their medical events. As well they found that the billing breakdown of payments and dates was often unique, thus private insurance companies, banks, and credit card companies, could use their own records to match to the medical records and learn the individual's medical history [7].

A summary of the above into best practices of perturbation:

1. Decreasing the precision of the data, or perturbing it statistically, makes re-identification gradually harder at a substantial cost to utility,

2. A 2-week perturbation of dates makes little impact on sparse data, increasing the perturbation has little effect.

8.4. Aggregation

The only health data set studied that used aggregation did so on physical activity data collected from wearable devices. The aggregation on this set was performed within the attributes, as an individual's average walking intensity for every chosen time interval was calculated and released. Researchers then used a data set containing 6 demographic variables; age, sex, educational level, annual household income, race/ethnicity, and country of birth, about the individuals in the data set to link to the aggregated walking intensity data. This matching was performed using a random forest machine learning algorithm model. It correctly matched 94.3% of adults and 87.2% of children when time intervals of 15-minutes were used. When 24-hours was the time interval it matched 87.0% of adults and 70.2% of children. This study shows that this type of physical activity data can be used to learn more about the individuals within than ever intended by the data custodian [53].

8.5. Access Control

A common method of access control is query control. In one case having data sets of genomic information, but only allowing yes or no responses as to whether a specific allele is in the data set. Some researchers who had access to such a data set found that because an allele's presence can be dependant on other allele's in the genome an individual's presence in the set could be discovered from queries of their single-nucleotide polymorphisms (SNPs) [70]. 5 queries to the set revealed with 95% confidence whether someone was in the set. Removing the ability to query SNPs with less than 5% frequency in populations did not affect their ability to identify presence. Due to the few required queries, limiting the number of queries to the set is ineffective as it would be necessary to limit queries below a number allowing useful analysis of the data. As well it was found that hiding parts of the genome completely caused a similar loss of usefulness to the data set to be effective at preventing the attack.

HIPAA has within its exemptions the limited data set rules. A limited data set under HIPAA is identifiable healthcare information that the HIPAA Privacy Rule permits covered entities to share information with certain entities for research purposes, public health activities, and healthcare operations without obtaining prior authorization from patients, if certain conditions are met, one condition being they signed a data use

agreement that specifies: Allowable uses and disclosures, approved recipients and users of the data, an agreement that the data will not be used to contact individuals or re-identify them, require safeguards to be implemented to ensure the confidentiality of data and prevent prohibited uses and disclosures, state the discovery of improper uses and disclosures must be reported back to the covered entity, state that any subcontractors who are required to access or use the data also enter into a data use agreement and agree to comply with its requirements. With that settled a limited data set cannot also cannot contain any of the following information: names, street addresses or postal address information with the exception of town/city, state and zip code, phone/Fax numbers, e-mail addresses, social Security numbers, medical records numbers, health plan beneficiary numbers, other account numbers, certificate and license numbers, vehicle identifiers and serial numbers, including license plates, device identifiers and serial numbers, URLs and IP addresses, biometric identifiers such as fingerprints, retinal scans and voice prints, full face photos and comparable images.

Even with all of this removed it was found that in Ohio 18.7% of the population is 1-distinct, or unique, and 59.7% are 5-distinct based on their County, Gender, Date of Birth, and Race [4]. This is compared to the risk if the data is kept under the full HIPAA protections, under Safe Harbor, 0.0003% is 1-distinct and 0.002% are 5-distinct This means that though the data is HIPAA compliant the data receivers now hold health data that can be linked back to the individual. Though the receiving entity is bound by the data use agreement to never use it for such a purpose, there is now an increased risk of another entity gaining access to the information from their data center.

A summary of the above into best practices of access control:

1. Dependencies in the attribute values need to be considered as they reveal more information than intended,

2. Access control cannot completely remove risk for the individuals in the data set.

9 Geolocation Data

Geolocation data is information based around an individual's position at a point in time. This information could be traces of their movements over time, such as from a continuous GPS connection providing constant updates of their latitude and longitude coordinates, pings of their location as they access specific services in certain locations, such as when accessing public transit, or a general location such as cell tower connections, where each tower has a range of area that the person connecting could be anywhere inside of. In all cases the location data contains detail of where someone was, and when they were there. Often from this data a trajectory of the individual's movements can be made as they move through an area and connect to different cell towers or connect to services as they use them.

If a geolocation data set is released that contains information that can be re-identified the effect on the individuals within the set can be damaging. Attackers could learn personal information like average income [56], or habits and vices [72], and home address [13]. They could also predict where someone will go and when they will be there [76]. Attackers thus could learn detailed information about the patterns of their life as well as other private information.

Researching attacks on this type of data revealed a lot of information about the protections that were being used. For a majority of the data nothing was done beyond removing direct identifiers to the individual's identity. Most of the data sets contained only the times, locations, and a perpetual identifier used to track traces made by the same individual through the data set.

It also revealed the methods used by the attackers to break this de-identification. The spread of the attacks and used and their course of re-identification is laid out in Figure 9.2b. There were different methods of doing this, many attackers used clustering to find likely home addresses or home and work pairs. They could also use Markov Chains to match the de-identified location patterns to known location patterns that the attacker created themselves from knowledge of the individual or from other public data that contains an identity. Some attacks were only looking at uniqueness of the location patterns and many found that the information was unique to an individual which provides the reasonable assumption of identifiability. The only cases not discussed in the following sections are the two cases of broken masking that were studied as these are discussed in Section 6.3 of this document.

9.1. Suppression

With geolocation data typically suppression will include the removal of names and other known demographic information that can identify the person. Many data holders consider the time stamps and location as well as some form of perpetual identifier as enough de-identification. However in multiple cases it has been found that the location data itself when connected together can leak information about the person, including home address and place of work, which together can identify an individual [21, 23, 27]. From an analysis of the census data in the US it was found that home work pairs at the location granularity of the census block were unique for majority of the population, and less granularity offers more privacy [27].

In some cases the only information released is the address of an individual. For a health care a data set a map containing patient addresses was released with the intention of allowing analysis of illness and

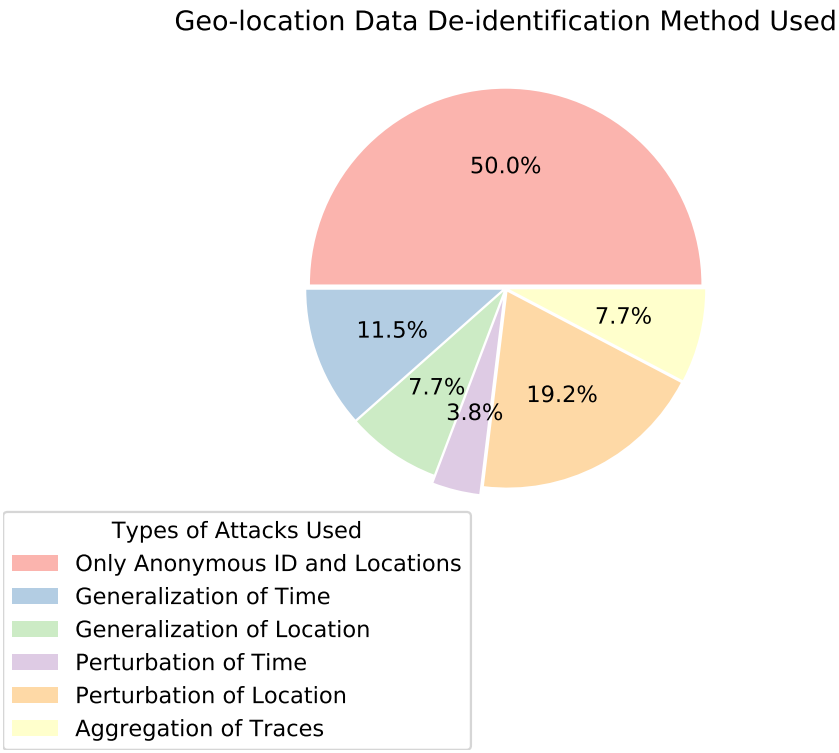
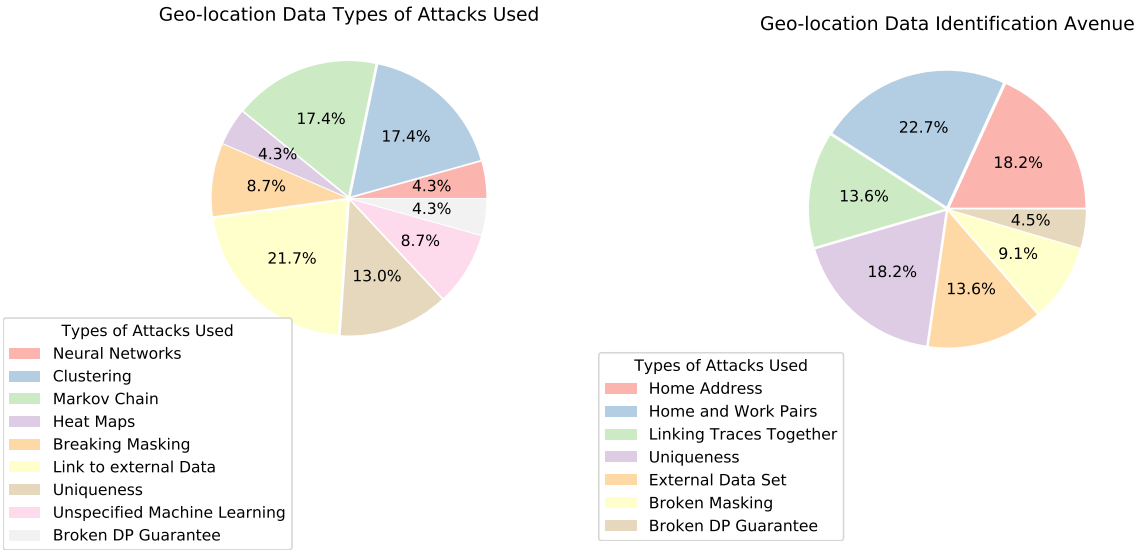


Figure 9.1: Breakdown of the de-identification used on geolocation data sets



(a) Attack methods

(b) Avenue of identification

Figure 9.2: Break down of the attack methods used on geolocation data sets

geographical location through the city of Boston. It was found that the data released was precisely accurate for the home address of 79% of the patients and within 14 meters for all of the addresses. From there researchers determined finding the identity of the patient was possible from an accurate home address [5].

Machine learning models can be used to match individuals between data sets. In a study performed using credit card information that contained the time and location of the transaction it was found that 4 points of an individuals time and location were enough to re-identify their trace within the data set. Overall 90% of individuals were unique with only 4 points [9].

Machine learning models can be used to create models of human behaviour, when it comes to geolocation data a neural network being fed discrete GPS locations and times a map of someone's trajectories and

frequented locations can be built. Using this it is possible to use the location features of social media apps such as twitter, foursquare, weibo, and ISP data to match an anonymous social media account to an identity. Researchers were able to take 21 months of twitter and 48 months of foursquare data and match accounts between the sites without looking at the content of posts or the account names. Similarly they used 1 week of data from Weibo which used GPS coordinates, and 1 week of data from an ISP which provides the coordinate of the base station that the user was connected to at a certain time to connect individuals between the data sets. They were able to accurately match 50% of users between the two services [20]. Similar analysis can be completed using a Markov chain with a 35-45% match rate [24].

Another research group took the GPS coordinates from twitter data and used machine learning models to create clusters of coordinates and then create a centroid of these coordinates. This centroid assumed to be the true location the person was at once the clustering accounted for the error in the coordinates. Using knowledge of humans behaviour between work and home hours the models then predicted the centroid that was likely to be an individuals home and their workplace. Once this was done the contents of the tweets were searched for relevant information about the location the individual was at at the time of the tweet and the model would consider this as well. The model created by researchers was able to predict an individuals home and workplace with 92.5% accuracy [13]. Another study focused on the re-identification of 3 people in this manner and learned names, dob, occupation, family info, home and work address, several facts about their life, info about their web presence and other miscellaneous information [30].

In a case that compared bike share and jogging data it was found that the knowledge of someone's daily routine could easily be leaked to an attacker. Many people have a daily routine that location data covering their morning or evening jog would reveal information about. As would data about their use of a bike share service to and from work, or another form of public transit. Both of these data sets provide near continuous updates of an individuals locations and researchers could create models based on the released data of the individuals movements and predict a bike share users location and time with 75% accuracy [48].

Using a taxi data set from New York that contained no information about the passengers identity it was shown that it is arbitrary to re-identify a passenger in the data set if some information is known about where and when they used a taxi, or their address, as the taxi data is detailed enough to see the address that the taxi picked someone up or dropped them off at. This can reveal information previously unknown to an adversary about the trip that was taken, and the habits and behaviours of an individual. For detached homes, previous knowledge can be minimal as information about the owner can be searched through the address to reveal their identity [72]. A similar study of taxis in San Francisco and Shanghai as well as busses in Shanghai found that 10 pieces of external information were enough to identify a passenger in the data set. Even when the external information was inaccurate [41].

If some external information about a person's location behaviour is known then cell tower data consisting of the entry and exit time of a cell phone from each cell tower's area of coverage can be used to build a model of their location patterns and compared to this external information to find them in the data set. Using this an attacker can achieve 80% accuracy in connecting the identity to the location trace [52].

A summary of the above into best practices of suppression:

1. Home address should be considered a direct identifier for a homeowner and information that reveals a home address is an identifier,
2. The uniqueness of specific locations and patterns of movement to an individual should be considered personal information,
3. Four external locations and times are enough to uniquely identify an individual and thus require de-identification.

9.2. Generalization

Generalization with geolocation data can typically occur in one of two ways, either the location is generalized or the time is generalized. Potentially both of these alterations are occurring. Generalization will alter the specific latitude and longitude to cover a wider area or change the specific time to cover a greater time frame that the person may have been at that location during. Some information about the types of generalization and the identifiability of the data after this was performed is displayed in Table 9.1.

In some areas of the world road cameras are common. In Guangzhou China road cameras along major roads take images of a vehicle's license plate and place it in a data base along with the time stamp, which camera took the photo, and whether the vehicle is local or not. This amounts to a location, time, and perpetual

Table 9.1: Re-identifying Geolocation Data

Study	Year	Country	Type of Data	Time	Location	Identification	Re-id'd
[25]	2019	CHN,CAN	5 Discrete vehicular locations	30 min	Exact	Uniqueness	100%
[25]	2019	CHN,CAN	5 Discrete vehicular locations	1 hr	Exact	Uniqueness	100%
[25]	2019	CHN,CAN	5 Discrete vehicular locations	3 hr	Exact	Uniqueness	98%
[25]	2019	CHN,CAN	5 Discrete vehicular locations	6 hr	Exact	Uniqueness	95%
[25]	2019	CHN,CAN	5 Discrete vehicular locations	12 hr	Exact	Uniqueness	90%
[47]	2017	USA,FRA	Discrete locations 200m apart	Exact	Exact	Uniqueness	68%
[12]	2016	USA	Taxi GPS positions	1 min	Exact	Uniqueness	91%
[12]	2016	USA	Taxi GPS positions	5 min	Exact	Uniqueness	90.3%
[12]	2016	USA	Taxi GPS positions	15 min	Exact	Uniqueness	88.6%
[12]	2016	USA	Taxi GPS positions	30 min	Exact	Uniqueness	86.7%
[12]	2016	USA	Taxi GPS positions	1 min	Census tract	Uniqueness	87.8%
[12]	2016	USA	Taxi GPS positions	5 min	Census tract	Uniqueness	83.5%
[12]	2016	USA	Taxi GPS positions	15 min	Census tract	Uniqueness	81.4%
[12]	2016	USA	Taxi GPS positions	30 min	Census tract	Uniqueness	75.5%
[12]	2016	USA	Taxi GPS positions	1 min	ZIP code	Uniqueness	84.1%
[12]	2016	USA	Taxi GPS positions	5 min	ZIP code	Uniqueness	78.4%
[12]	2016	USA	Taxi GPS positions	15 min	ZIP code	Uniqueness	68%
[12]	2016	USA	Taxi GPS positions	30 min	ZIP code	Uniqueness	54.9%
[12]	2016	USA	Taxi GPS positions	1 min	NYC neighbourhood	Uniqueness	82.5%
[12]	2016	USA	Taxi GPS positions	5 min	NYC neighbourhood	Uniqueness	70%
[12]	2016	USA	Taxi GPS positions	15 min	NYC neighbourhood	Uniqueness	50.5%
[12]	2016	USA	Taxi GPS positions	30 min	NYC neighbourhood	Uniqueness	29.6%
[9]	2015	USA	4 Points of time and location of purchases	Exact	Exact	Uniqueness	90%
[41]	2013	USA,CHN	Taxi and bus GPS positions	1 min	0.01° coordinates	Identity matching	50%
[51]	2012	USA,CHL,BEL	4 Cell tower points covering 0.15km^2 to 15km^2	1hr	Cell tower	Uniqueness	95%
[51]	2012	USA,CHL,BEL	2 Cell tower points covering 0.15km^2 to 15km^2	1hr	Cell tower	Uniqueness	50%
[51]	2012	USA,CHL,BEL	4 Cell tower points covering 0.15km^2 to 15km^2	5hr	5 Cell towers area	Uniqueness	50%
[76]	2011	USA	Cell tower points top location	Exact	Sector	Bin-size uniqueness	372
[76]	2011	USA	Cell tower points top 2 locations	Exact	Sector	Bin-size uniqueness	2
[76]	2011	USA	Cell tower points top 3 locations	Exact	Sector	Bin-size uniqueness	1
[76]	2011	USA	Cell tower points top location	Exact	Cell	Bin-size uniqueness	967

Table 9.1: Re-identifying Geolocation Data

Study	Year	Country	Type of Data	Time	Location	Identification	Re-id'd
[76]	2011	USA	Cell tower points top 2 locations	Exact	Cell	Bin-size uniqueness	9
[76]	2011	USA	Cell tower points top 3 locations	Exact	Cell	Bin-size uniqueness	1
[76]	2011	USA	Cell tower points top location	Exact	ZIP code	Bin-size uniqueness	3125
[76]	2011	USA	Cell tower points top 2 locations	Exact	ZIP code	Bin-size uniqueness	75
[76]	2011	USA	Cell tower points top 3 locations	Exact	ZIP code	Bin-size uniqueness	2
[76]	2011	USA	Cell tower points top location	Exact	City	Bin-size uniqueness	7638
[76]	2011	USA	Cell tower points top 2 locations	Exact	City	Bin-size uniqueness	437
[76]	2011	USA	Cell tower points top 3 locations	Exact	City	Bin-size uniqueness	24
[76]	2011	USA	Cell tower points top location	Exact	County	Bin-size uniqueness	55649
[76]	2011	USA	Cell tower points top 2 locations	Exact	County	Bin-size uniqueness	15628
[76]	2011	USA	Cell tower points top 3 locations	Exact	County	Bin-size uniqueness	3407
[76]	2011	USA	Cell tower points top location	Exact	State	Bin-size uniqueness	720000
[76]	2011	USA	Cell tower points top 2 locations	Exact	State	Bin-size uniqueness	680000
[76]	2011	USA	Cell tower points top 3 locations	Exact	State	Bin-size uniqueness	460000
[27]	2009	USA	Census data home work pairs	N/A	Census block	Bin-size uniqueness	1
[27]	2009	USA	Census data home work pairs	N/A	Census tract	Bin-size uniqueness	21
[27]	2009	USA	Census data home work pairs	N/A	County	Bin-size uniqueness	34980

identifier that can be used to track the vehicle along its route. This means that they can create a trajectory of movement based on these single instances in time. They attempted various time granularity's, but found that even with 12 hours, 5 of these records was enough to uniquely identify 90% of the individuals driving on the roads [25].

One method for generalizing the data in the time domain referred to as Promesse [57] erases user points of interest by using a speed smoothing technique, which assures that between each successive points in the obfuscated trace the distance and time difference are the same. This way someone spending a lot of time in a single location should appear similar on their location trace to somewhere they spent less time. However with this there are methods that can release information to an attacker. Though a place of interest may be harder to determine from a single trace, using a heat map to create points of interest over multiple traces is still possible and will leak enough information to identify a user [47]. This study looked at multiple data sets and re-identification types and found that the same de-identification measures did not result in the same level of de-identification of the records. Despite the structure of the data being the same in all cases the nature of the data resulting from its source, cabs or users on a social network, can alter the effectiveness of an attack due to the patterns of movement. Cabs movements have much less uniqueness than a person and so are more difficult to re-identify once de-identification is applied to the data [47].

In a study of taxi GPS traces from New York taxis researchers tested the privacy of different generalization levels on the data by looking at how unique the traces were. This was considered a successful attack be-

cause they were also able to prove that a trace could be matched to the public medallion information which revealed the driver's identity. With location generalized to the neighbourhood and time generalized to 30 minute intervals it was still possible to identify 30% of the individuals [12].

Looking at cell phone data where the locations are generalized to the area covered by a single cell tower with 1 hr samples only 4 positions are required to uniquely identify someone. From this it was found that statistically, traces are more unique when coarse in one dimension and fine along another than medium grained along both dimensions. Given four points, 40% of individuals are unique in a data set with a temporal resolution of 15 hrs or a spatial resolution of 15 antennas while 60% are unique in a data set with a temporal resolution of 7 hrs and a spatial resolution of 7 antennas. According to their analysis uniqueness decays 1/10 the power of the resolution [51].

Another study looking at cellphone data was generalizing the location starting with the sector of the cell tower's area of coverage, the cell towers area, the zip code area, city, county, and state. They base their analysis on the uniqueness of the set of most common locations starting with one location up to the top 3. From their analysis they determined that a trace longer than 2 weeks reveals the top 2 locations of more than 50% of a population.

A summary of the above into best practices of generalization:

1. Location traces over time should should be de-identified to prevent attacks that look for important places in people's lives,
2. When generalizing locations it is important to note the uniqueness of home work pairs extends to areas not just exact addresses,
3. Generalizing locations so that every point is at least a specific distance apart still allows for points of interest to be found when multiple traces from an individual can be tied together,
4. Data sets from different sources should be considered new data sets when determining the required methods of de-identification as methods used on similarly structured geolocation data will not provide the same protection level,
5. Large data sets with many records for each individual require significant generalization to provide k-anonymous privacy.

9.3. Perturbation

Perturbation is used on coordinates to alter the exact positions slightly and create uncertainty to prevent attackers from knowing specifically where someone was, while researchers can still learn from movement patterns. Two studies of different data sets that added noise to individual geolocation traces found vulnerabilities in the implementation. They found that many of the traces were correlated, people that know each other of course often go to the same places together and immediate family will be in the same location even more often. As such models like Markov chains can be used to compare the traces and find people that are likely to know each other. Using the correlated traces and clustering the points where the individuals were likely together the real location can be found [37, 69].

Specifically, for the case of [37] the noise was applied to geolocation traces from a social networking site using a differential privacy bound and laplacian noise. By leveraging the information about relationships between individuals available on the social media website combined with the geolocation traces specific users could be inferred to know each other. Once that connection was known points on the traces where they were likely in the same location could be exploited using clustering to defeat the noise applied to the positions and learn the exact location [37].

Noise can be generated in different manners to be applied to geolocation traces. One study looking at noise that was drawn from a planar Laplace distribution, as is the case with the Geo-Indistinguishability protocol [1], found weaknesses in this style of noise addition. Using a heatmap to create a pattern of visited locations and frequencies an attacker can still find uniqueness in the trace, and potentially reveal the user from external knowledge of their movements to link to the data set [47].

Other studies were able to use Markov chains to break perturbation applied to the location as the adversary can focus on a subset of transition probabilities. From these the entire transition probability matrix can be recovered. This allowed for the adversary to estimate the locations of the users and thus re-identify them [69].

A summary of the above into best practices of perturbation:

1. Noise added to the location should account for the same person being in the same place multiple times to counteract attacks based on frequency in a location,
2. The method through which the noise is generated and applied should always account for areas in which people would not actually be, such as the middle of a lake,
3. When noise is added to traces without considering the dependencies between the traces relationships between users can be used to learn exact locations through the noise.

9.4. Aggregation

Aggregating similar user's traces together is a method that can add some anonymity to the data set. Traces with similar movement patterns can be merged together to create a single trace that is then released. Essentially this creates a k-anonymous set, if 5 traces were aggregated then any trace could be at least $k = 5$ different people. However using heat maps to match a known trace to the aggregated ones can still leak information about the user that was not previously known to the attacker [47].

10

Browsing History

Web browsing history typically consists of nothing more than accessed links and times, and occasionally a perpetual identifier used to identify the same user. There is typically nothing else in the data set yet browsing history can be used to identify people, as well as personal information about them.

10.1. Suppression

As with all considerations of suppression unique attributes of the data to the person need to be considered. When visiting a site like social media the url is not always the same for all people. Sometimes when accessing features only available for your personal profile the url is unique to you [16]. Other unique urls might include employer websites, or google searches containing personal information like names, ages, and locations [3]. In fact only 10 known searches are required to create a finger print of browsing behaviour that can be used to discover someone's entire browsing history from a data set of just urls and timestamps. These unique searches should be removed from the data or altered before release.

Through social media it is also possible to take the browser history and find someone's identity through matching their browser history to their social media feed. When sites like twitter contain embedded links accessing the secondary website through the provided link creates a url that indicates that this link was opened through a website such as twitter. These types of links create a “finger print" that can then be matched to a twitter feed 70% of the time

11

Call Records

Many consider telephone metadata to be without identifiers because the information contains no values that are traditionally though of as direct identifiers, such as names [49]. However information like locations, relationships between people, and sensitive information can be obtained from the metadata of cellphone records, all of which can contribute to re-identification.

11.1. Suppression

Though telephone metadata contains little information in terms of direct identifiers there is one attribute that can reveal a lot of information about an individual's identity, location, relationships, and sensitive information. Much of the telephone metadata referred to by organizations contains phone numbers. These have been found to be trivially re-identifiable using directories, and social network application programming interfaces [49].

Phone numbers in metadata can reveal not only the identity of the individual the data is from, but the identity of people that they know, because the call and text logs contain recipient and sender phone numbers. Phone numbers can also reveal locations from the business numbers that are called, and learn personal or sensitive information about a person based on the phone calls to businesses, doctor's offices, clinics, religious affiliations, and other organizations. Linking a phone number to a business and physical address is trivial through google places or review services such as yelp. It was found that using 10 phone calls it is possible to predict the individual's location [49].

Telephone numbers should be considered as an identifier and treated as such when de-identifying data sets. If full suppression of the phone number would reduce the utility of the data set below utility thresholds other methods of de-identification could potentially be used to prevent identity leakage.

11.2. Generalization

The relationships between individuals can be inferred from the telephone meta data. From the concentration of call and text volume and length, time of day for call and text volume and length, and comparisons of whether the most called number was the most texted number, and comparisons between most called number and most texted number, it is possible to determine who the individuals are in a relationship with. This is despite metadata containing none of the contents of the calls and text's [49].

12

Social Networks

Social networking data can often be released in the form of undirected graphs with nodes representing the people and edges representing their network of “friends" within the site. More information can be released along with this including usernames or geolocation positions one such case was discussed in Section 9 on Geolocation data.

One of the issues with most social networks is that the data that is released is available to an attacker through the nature of the social network. Even with privacy settings on the site set as high as possible some information about a person on the site is available to anyone looking for it. This would give an adversary a clear external data set to compare the information to.

12.1. Suppression

When these social networking graphs are released they often contain no information about the identity of the individual, and in some cases the social network would not have a real name only a username on record. If there are names associated with the account they are not released and only the structure of the nodes representing people and the edges between them representing relation ships are released.

This graph structure can be used to connect the information back to an external de-identified data source. Looking at networks from Gowalla (a social network from 2009-2012) and Google+ (a social network from 2011-2019), it was found that the uniqueness of the structure of the graph allowed for re-identification of the nodes. From the social network graphs that were released 83.3% of users of Gowalla and 95.5% of users of Google+ could be identified. This privacy leakage would also get worse the more nodes and edges that a network had [31].

A similar attack was done with Twitter and Flickr accounts. By looking at the structure of the networks they were able to match anonymous accounts between the two social networks. The algorithm used found a few matching node pairs between the networks and were able to use these to expand to further nodes and identify 30.8% of the nodes that appeared in both networks [55].

Suppression of the structure does not provide enough privacy for a balance with utility of the data. For instance if the degree of a node (the number of edges, meaning in this case direct relationships to other nodes) is unique, then someone could be re-identifiable from that information. Applying k-anonymity principles to the degree of the nodes so that $k-1$ nodes all have the same degree is a method that has been proposed to de-identify these graphs [38]. When one such method was implemented and attacked it was found that an attack that focuses on the structure of the network was still successful, as there was an overlap of 58% of the edges in the network with the external data. Higher k values would have required the removal of more data that would make the entire data set far less useful [46].

12.2. Perturbation

An interesting method of perturbing the information in a social network graph was proposed that involves adding and deleting edges of the graph at random [75]. The idea is to create noise in the structure of the graph and prevent an attacker from being able to use this structural information to match the de-identified node back to the true node and de-identify the user. However a structural based re-identification attack was

successful for 61% of the nodes in this data until while the added noise was held to an acceptable level of error [46].

12.3. Aggregation

A different way of implementing a k-anonymous measure of privacy to a social network graph is to group similar graph nodes into clusters with a minimum size constraint [71]. This was implemented and then attacked using structure based methods. Despite the de-identification there remained 63% overlap with the external data when the method was implemented on a network at acceptable utility levels [46].

13

Billing Information

Billing information includes details about any products or services that were paid for by one group to another. This typically involves a breakdown of the amount paid, and how it was paid, any tip that might have been added on top of that, and potentially details about what was purchased.

13.1. Suppression

With billing information there is always some suppression of data, credit card information, bank account numbers, and other data that someone could use to fraudulently make transactions must be removed from the data before release.

In a case of credit card data being released that contained times, locations, and amount paid it was found that 4 locations known by an adversary could identify 90% of the individuals in the data set, this is discussed in Section 9, and that the adversary knowing the price of the transaction increased the re-identification risk by 22%. Their attack was performed using machine learning methods applied to the data and looking to match records within the data set to their external data [9].

13.2. Perturbation

When applying noise to a data set the current common method involves some form of differential privacy. In a study three different definitions of differential privacy were implemented on a data set containing customer purchase records corresponding to 100 frequently purchased items. The differential privacy implementation that were tested were naïve composition, advanced composition[15], zero concentrated differential privacy [6], and Rényi differential privacy [50]. The study focused on using the data for machine learning purposes and comparing the utility of the data to the chosen privacy budget ϵ. They concluded that for the machine learning models to create accurate predictions of the data the privacy budget has to be set far too high to provide any protections from a realistic adversary [29].

14

Conclusion

De-identification is the process of lowering the probability of information being used to identify an individual, or a unique pattern in the data set. There are six general ways that this can be done, and their effectiveness depends on the data type, the specific data set, and the external data available to an attacker. The effectiveness of a de-identification method depends on the data type it is being implemented on and the contents of the data set. Key considerations need to be made based on the type of external data available to an attacker and the uniqueness of the values or set of values in the data set.

Though many of the attacks investigated revealed personal information about individuals involved, there is little evidence to suggest that the modern theorems of de-identification are flawed. The majority of data sets had the direct identifiers removed or masked, but little was done beyond that to remove identities. Though there does appear to be an increase of custodians beginning to employ other methods of de-identification in recent years, as all studies involving obfuscation of data are from after 2010, with 10 of the 12 being from the last 3 years (2016) [7, 22, 23, 25, 29, 37, 41, 47, 62, 68–70]. These found that the implementation, not the theory was at fault for the privacy breach.

Though there will be a trade-off in utility when performing operations on data beyond removal of direct identifiers [11] it is important to maintain the public's privacy. Data custodians need to be up to date on de-identification methods and implementations, as well as documented failures of them. Increased adoption and testing of de-identification methods would be beneficial to the entire research community and public. Testing these methods on real world data sets is the best way to determine what methods and standards work and which do not. Testing will guide best practices to be better and give data custodians more information they can use when considering the de-identification methods, standards, and processes their data will require.

Acknowledgment

The author would like to thank Aleksander Essex for his guidance and assistance with understanding concepts within this paper. Thanks also to Elena Novas and Zeev Glauberzon for their insights into industry, as well as their guidance, feedback, and for making this project possible.

Bibliography

[1] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *arXiv preprint arXiv:1212.1984*, 2012.

[2] Johann Bacher, Ruth Brand, and Stefan Bender. Re-identifying register data by survey data using cluster analysis: An empirical study. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):589–607, 2002. doi: 10.1142/s0218488502001661.

[3] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749, Aug 2006. URL `https://www.nytimes.com/2006/08/09/technology/09aol.html`.

[4] Kathleen Benitez and Bradley Malin. Evaluating re-identification risks with respect to the hipaa privacy rule. *Journal of the American Medical Informatics Association*, 17(2):169–177, 2010. doi: 10.1136/jamia.2009.000026.

[5] John S. Brownstein, Christopher A. Cassa, and Kenneth D. Mandl. No place to hide — reverse identification of patients from published maps. *New England Journal of Medicine*, 355(16):1741–1742, 2006. doi: 10.1056/nejmc061891.

[6] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016. URL `http://arxiv.org/abs/1605.02065`.

[7] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. Health data in an open world. *CoRR*, abs/1712.05627, 2017. URL `http://arxiv.org/abs/1712.05627`.

[8] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. Publication of mbs/pbs data commissioner initiated investigation report. *Australian Government Office of the Australian Information Commissioner*, 2018. URL `https://www.oaic.gov.au/privacy/privacy-decisions/investigation-reports/mbspbs-data-publication/`.

[9] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015. ISSN 0036-8075. doi: 10.1126/science.1256297. URL `https://science.sciencemag.org/content/347/6221/536`.

[10] Zach Diamond. We know they are listening, but what do they hear?, Apr 2017. URL `https://medium.com/analytics-for-humans/hey-listen-how-four-big-tech-companies-are-collecting-and-using-your-data-8ddb6ea857dc`.

[11] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM. ISBN 1-58113-670-6. doi: 10.1145/773153.773173. URL `http://doi.acm.org/10.1145/773153.773173`.

[12] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Claudio T. Silva. Anonymizing nyc taxi data: Does it matter? *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016. doi: 10.1109/dsaa.2016.21.

[13] Kostas Drakonakis, Panagiotis Ilia, Sotiris Ioannidis, and Jason Polakis. Please forget where I was last summer: The privacy risks of public location (meta)data. *CoRR*, abs/1901.00897, 2019. URL `http://arxiv.org/abs/1901.00897`.

[14] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, July 2006. ISBN 3-540-35907-9. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/.

[15] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

[16] Svea Ecker and Andreas Dewes. Your 'anonymous' browsing data isn't actually anonymous, Aug 2017. URL https://www.vice.com/en_us/article/gygx7y/your-anonymous-browsing-data-isnt-actually-anonymous.

[17] PF Edemekong and MJ. Haydel. Health insurance portability and accountability act (hipaa). 2019. URL https://www.ncbi.nlm.nih.gov/books/NBK500019/.

[18] Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mark Lysyk. Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian Journal of Hospital Pharmacy*, 62(4), 2009. doi: 10.4212/cjhp.v62i4.812.

[19] Khaled El Emam, Fida K Dankar, Angelica Neisa, and Elizabeth Jonker. Evaluating the risk of patient re-identification from adverse drug event reports. *BMC Medical Informatics and Decision Making*, 13(1), 2013. doi: 10.1186/1472-6947-13-114.

[20] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. *The World Wide Web Conference on - WWW 19*, 2019. doi: 10.1145/3308558.3313424.

[21] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the privacy risk of location-based services. In George Danezis, editor, *Financial Cryptography and Data Security*, pages 31–46, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-27576-0.

[22] Andrea Gadotti, Florimond Houssiau, Luc Rocher, and Yves-Alexandre de Montjoye. When the signal is in the noise: The limits of diffix's sticky noise. *CoRR*, abs/1804.06752, 2018. URL http://arxiv.org/abs/1804.06752.

[23] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and i will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '10, pages 34–41, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0435-1. doi: 10.1145/1868470.1868479. URL http://doi.acm.org/10.1145/1868470.1868479.

[24] Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nunez Del Prado Cortez. De-anonymization attack on geolocated data. *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2013. doi: 10.1109/trustcom.2013.96.

[25] Jing Gao, Lijun Sun, and Ming Cai. Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. *Transportation Research Part C: Emerging Technologies*, 104:78–94, 2019. doi: 10.1016/j.trc.2019.04.022.

[26] Philippe Golle. Revisiting the uniqueness of simple demographics in the us population. *Proceedings of the 5th ACM workshop on Privacy in electronic society - WPES 06*, 2006. doi: 10.1145/1179601.1179615.

[27] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. *Lecture Notes in Computer Science Pervasive Computing*, page 390–397, 2009. doi: 10.1007/978-3-642-01516-8_26.

[28] Victor Janmey and Peter L Elkin. Re-identification risk in hipaa de-identified datasets: The mva attack. *AMIA ... Annual Symposium proceedings*, 2018 1329-1337, 12 2018.

[29] Bargav Jayaraman and David Evans. When relaxations go bad: "differentially-private" machine learning. *CoRR*, abs/1902.08874, 2019. URL http://arxiv.org/abs/1902.08874.

[30] Lukasz Jedrzejczyk, Blaine A. Price, Arosha K. Bandara, and Bashar Nuseibeh. Know what you did last summer : risks of location data leakage in mobile and social computing. 2009.

[31] S. Ji, W. Li, M. Srivatsa, and R. Beyah. Structural data de-anonymization: Theory and practice. *IEEE/ACM Transactions on Networking*, 24(6):3523–3536, December 2016. ISSN 1063-6692. doi: 10.1109/TNET.2016.2536479.

[32] El Emam K, Buckeridge D, Tamblyn R, Neisa A, Jonker E, and Verma A. The re-identification risk of canadians from longitudinal demographics. June 2011. doi: 10.1186/1472-6947-11-46.

[33] P. Knoche. Factual anonymity of microdata from household and person related surveys - the release of microdata for scientific purposes. *Proceedings of the International Symposium on Statistical Confidentiality*, pages 407–413, 1993.

[34] Matthijs Koot, Guido Noordende, and Laat de C. A study on the re-identifiability of dutch citizens. *Journal of Clinical Virology - J CLIN VIROL*, 01 2010.

[35] Arturs Lavrenovs and Karlis Podins. Privacy violations in riga open data public transport system. *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2016. doi: 10.1109/aieee.2016.7821808.

[36] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

[37] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnerable: Differential privacy under dependent tuples. *Proceedings 2016 Network and Distributed System Security Symposium*, 2016. doi: 10.14722/ndss.2016.23279.

[38] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 93–106, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-102-6. doi: 10.1145/1376616.1376629. URL `http://doi.acm.org/10.1145/1376616.1376629`.

[39] Natasha Lomas. Researchers spotlight the lie of 'anonymous' data, Jul 2019. URL `https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/`.

[40] Grigorios Loukides, Joshua C Denny, and Bradley Malin. The disclosure of diagnosis codes can breach research participants privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327, 2010. doi: 10.1136/jamia.2009.002725.

[41] Chris Y. T. Ma, David K. Y. Yau, Nung Kwan Yip, and Nageswara S. V. Rao. Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM Transactions on Networking*, 21(3):720–733, 2013. doi: 10.1109/tnet.2012.2208983.

[42] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24, April 2006. doi: 10.1109/ICDE.2006.1.

[43] Bradely Malin. Re-identification of familial database records. *AMIA Annu Symp Proc.*, page 524–528, 2006.

[44] Bradley Malin and Latanya Sweeney. Determining the identifiability of dna database entries. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 537–41, 02 2000.

[45] Bradley Malin and Latanya Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3):179–192, 2004. doi: 10.1016/j.jbi.2004.04.005.

[46] Jian Mao, Wenqian Tian, Jingbo Jiang, Zhaoyuan He, Zhihong Zhou, and Jianwei Liu. Understanding structure-based social network de-anonymization techniques via empirical analysis. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):279, Dec 2018. ISSN 1687-1499. doi: 10.1186/s13638-018-1291-2. URL `https://doi.org/10.1186/s13638-018-1291-2`.

[47] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. Ap-attack: A novel user re-identification attack on mobility datasets. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous 2017, pages 48–57, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5368-7. doi: 10.1145/3144457.3144494. URL http://doi.acm.org/10.1145/3144457.3144494.

[48] Clément Massart and François-Xavier Standaert. Revisiting location privacy from a side-channel analysis viewpoint. *Progress in Cryptology – AFRICACRYPT 2019 Lecture Notes in Computer Science*, page 333–351, 2019. doi: 10.1007/978-3-030-23696-0_17.

[49] Jonathan Mayer, Patrick Mutchler, and John C. Mitchell. Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences*, 113(20):5536–5541, 2016. doi: 10.1073/pnas.1508081113.

[50] Ilya Mironov. Renyi differential privacy. *CoRR*, abs/1702.07476, 2017. URL http://arxiv.org/abs/1702.07476.

[51] Yves-Alexandre De Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1), 2013. doi: 10.1038/srep01376.

[52] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. Identification via location-profiling in gsm networks. *Proceedings of the 7th ACM workshop on Privacy in the electronic society - WPES 08*, 2008. doi: 10.1145/1456403.1456409.

[53] Liangyuan Na, Cong Yang, Chi-Cheng Lo, Fangyuan Zhao, Yoshimi Fukuoka, and Anil Aswani. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Network Open*, 1(8), 2018. doi: 10.1001/jamanetworkopen.2018.6040.

[54] Arvind Narayanan and Edward W Felten. No silver bullet: De-identification still doesn't work. Jul 2014. URL http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[55] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. *2009 30th IEEE Symposium on Security and Privacy*, 2009. doi: 10.1109/sp.2009.22.

[56] Vijay Pandurangan. On taxis and rainbows, Jun 2014. URL https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

[57] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 539–546, Aug 2015. doi: 10.1109/Trustcom.2015.417.

[58] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-019-10933-3.

[59] Ian Sherr. Apple listens to some siri recordings to make it better, Jul 2019. URL https://www.cnet.com/news/apple-listens-to-some-siri-recordings-to-make-it-better/.

[60] Tom Simonite. Who's listening when you talk to your google assistant?, Jul 2019. URL https://www.wired.com/story/whos-listening-talk-google-assistant/?verso=true.

[61] L. Sweeney, J. S. Yoo, L. Perovich, K. E. Boronow, P. Brown, and J. G. Brody. Re-identification risks in hipaa safe harbor data: A study of data from one environmental health study. *Technology science*, 2017.

[62] L Sweeney, M Von Loewenfeldt, and M Perry. Saying it's anonymous doesn't make it so: Re-identifications of "anonymized" law school data. *Technology Science*, 11 2018. URL https://techscience.org/a/2018111301.

[63] Latanya Sweeney. Simple demographics often identify people uniquely. *Carnegia Mellon University*, 2000. URL https://dataprivacylab.org/projects/identifiability/index.html.

[64] LATANYA SWEENEY. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. doi: 10.1142/S0218488502001648. URL https://doi.org/10.1142/S0218488502001648.

[65] Latanya Sweeney. Patient Identifiability in Pharmaceutical Marketing Data. 1 2011. doi: 10.1184/R1/6625193.v1. URL https://kilthub.cmu.edu/articles/Patient_Identifiability_in_Pharmaceutical_Marketing_Data/6625193.

[66] Latanya Sweeney. Matching known patients to health records in washington state data. *SSRN Electronic Journal*, 2013. doi: 10.2139/ssrn.2289850.

[67] Latanya Sweeney, Akua Abu, and Julia Winn. Identifying participants in the personal genome project by name. *SSRN Electronic Journal*, 2013. doi: 10.2139/ssrn.2257732.

[68] Nazanin Takbiri, Amir Houmansadr, Dennis L. Goeckel, and Hossein Pishro-Nik. Limits of location privacy under anonymization and obfuscation. *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017. doi: 10.1109/isit.2017.8006631.

[69] Nazanin Takbiri, Amir Houmansadr, Dennis L. Goeckel, and Hossein Pishro-Nik. Privacy against statistical matching: Inter-user correlation. *CoRR*, abs/1805.01296, 2018. URL http://arxiv.org/abs/1805.01296.

[70] Nora Von Thenen, Erman Ayday, and A Ercument Cicek. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics*, 35(3):365–371, 2018. doi: 10.1093/bioinformatics/bty643.

[71] Brian Thompson and Danfeng Yao. The union-split algorithm and cluster-based anonymization of social networks. In *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, ASIACCS '09, pages 218–227, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-394-5. doi: 10.1145/1533057.1533088. URL http://doi.acm.org/10.1145/1533057.1533088.

[72] Anthony Tockar. Riding with the stars: Passenger privacy in the nyc taxicab dataset, Sep 2014. URL https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

[73] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 94–94, April 2006. doi: 10.1109/ICDEW.2006.116.

[74] David T.S. Personal information protection and electronic documents act, 2000. URL http://canlii.ca/t/5315t.

[75] Xiaowei Ying and Xintao Wu. Randomizing social networks: a spectrum preserving approach. In *proceedings of the 2008 SIAM International Conference on Data Mining*, pages 739–750. SIAM, 2008.

[76] Hui Zang and Jean Bolot. Anonymization of location data does not work. *Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom 11*, 2011. doi: 10.1145/2030613.2030630.