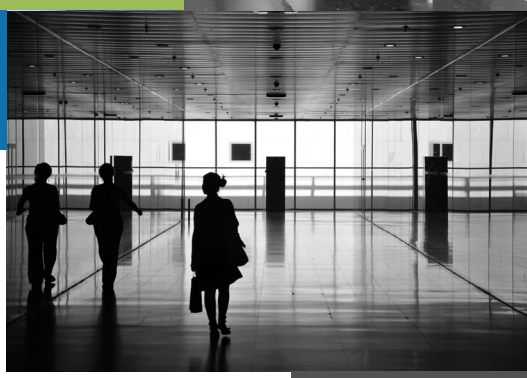
PRIVACY ANALYTICS

WHITE PAPER

De-identification 301

Data breaches and re-identification attacks compromise the personal privacy of individuals and both are on the rise. Re-identification results when a record is correctly tied to the person behind that data, even if the data was thought to have been made anonymous.

Re-identification attacks occur because an attacker has the skills, resources and the motivation to do so. Motivations can vary but are usually the result of curiosity or a desire for personal gain. Demonstration attacks happen when a researcher or journalist aims to show that a dataset has been insufficiently de-identified and wants to prove that re-identification is possible.



PRIVACY ANALYTICS

a QuintilesIMS company

Three Adversaries Who Could Attack Your Data

Understanding the risk of re-identification requires an understanding of the data's uniqueness. A combination of quasi-identifiers that is unique is called an equivalence class. If there are very few records in an equivalence class, e.g. one, then correctly identifying the person associated with that record may be easy.

There are three different types of attacks that can be made to try to deliberately re-identify data. These are termed prosecutor risk, journalist risk and marketer risk. Prosecutor risk aims to re-identify a specific person and relies upon pre-existing knowledge about a person known to exist in the de-identified database. Journalist risk also aims to re-identify an individual but instead uses access to another source of public information about an individual or individuals that are also present in the de-identified dataset. Marketer risk wishes to re-identify as many people a possible from the de-identified data even if this means some of them will be incorrectly identified. All three impact the overall risk of re-identification for a dataset.

Data Breaches and Re-identification Attacks

Data breaches garner significant media attention when they happen at a large public company or healthcare organization. A breach occurs when someone successfully accesses information to which they do not have rights. Often, breaches are the result of external forces exploiting lax network and security protocols or a consultant or employee who leaves an unencrypted device containing protected health information (PHI) in a public place. However, privacy can also be compromised when data that was thought to be made anonymous is re-identified, revealing personal and perhaps sensitive information about people. Unfortunately, security breaches and re-identification attacks are both on the rise[1].

While it is possible that re-identification can occur inadvertently, it is more likely the result of an intentional attempt to find a person within the data. These re-identification attacks fall under two broad categories: demonstration attacks and deliberate attacks. Deliberate attacks can be classed one of three ways depending on the risk being posed. These are termed prosecutor, journalist and marketer risk.

This is the third paper in a series that explores the issues around de-identification and the techniques used to protect patient privacy. Here we will delve into the problem of deliberate re-identification attacks, examine the motivations to re-identify data, discuss the use of equivalence classes to assess risk, and look at specific types of attacks.

The Motivations for Re-identifying Data

In order to re-identify a person from a de-identified database, an attacker must have the resources to carry out the attack, sufficient technical knowledge and a motivation to do so. The motivations to re-identify data can be grouped into three categories:

1) Breaches that are malicious with the intent to steal information, often for personal gain;

2) Breaches that are the result of curiosity, often about a friend, family member or public figure;

The Motivations for Re-identifying Data

Managing the Motives of Re-identification

One way to try to manage the motives for re-identification is to put contracts in place with data recipients that clearly lay out the limitations on the use and disclosure of the data. These contracts should include specific clauses that:

- Prohibit re-identification.
- Require the prohibition on re-identification be passed along to any other party with whom the information is shared.
- Prohibit any attempt to contact any of the patients in the dataset; and,
- Require an audit, either by the data owner or by a third party, that allows spot checks to be conducted to ensure compliance with the agreement.

3) Breaches resulting from a person who wants to prove that a dataset can be re-identified. This is referred to as a demonstration attack.

In the first case, the motivation to steal data is often tied to the potential to profit from reselling it on the black market. In its Fifth Annual Study on Medical Identity Theft, the Ponemon Institute reported a 21% increase over the previous year in the number of cases of medical identity theft in the U.S.[2] Compared to the theft of credit card numbers, medical identity theft reaps a much better payday for criminals. A recent New York Times article reported that, at one online auction, a complete medical record sold for $251 while credit card records sold for a mere 33 cents[3].

When the privacy breach is the result of curiosity, it often stems from an abuse of privilege by someone who has access to sensitive data. In 2010, nine individuals who worked for a Department of Education contractor were indicted for inappropriately accessing Barack Obama's student loan records[4]. All nine either pled guilty or were convicted, despite the fact that none of them had disclosed the information from the breach.

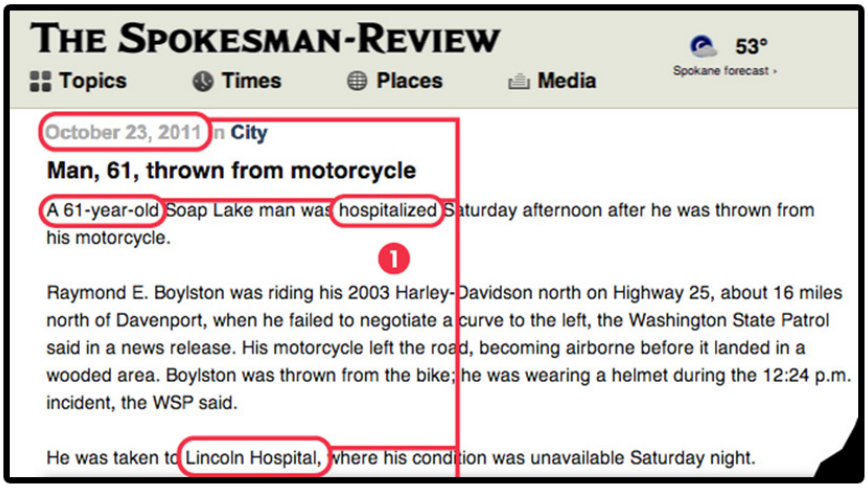
In the case of a demonstration attack, the attacker wants to show that re-identification is possible. Proving the point requires the re-identification of only a single record. The next section looks at how a demonstration attack can be perpetrated with the use of public information.

An Example of a Demonstration Attack

Even when a dataset is de-identified, personal privacy can still be threatened if information from the dataset can be matched to other information sources, such as publicly available databases (e.g. census data) or information like news reports. This is how a Washington State man was re-identified from his hospital admissions data[5].

In 2011, a Vietnam veteran named Ray Boylston had a motorcycle accident when he suffered a diabetic shock while riding. The incident was covered briefly in the local Washington paper (Fig. 1). The record relating to Ray's week-long stay at Lincoln Hospital was subsequently included in the hospital's inpatient database. As part of a larger

Managing Re-identification Risk



THE SPOKESMAN-REVIEW

53°
Spokane forecast ›

Topics Times Places Media

October 23, 2011 n City

Man, 61, thrown from motorcycle

A 61-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle.

1

Raymond E. Boylston was riding his 2003 Harley-Davidson north on Highway 25, about 16 miles north of Davenport, when he failed to negotiate a curve to the left, the Washington State Patrol said in a news release. His motorcycle left the road, becoming airborne before it landed in a wooded area. Boylston was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident, the WSP said.

He was taken to Lincoln Hospital, where his condition was unavailable Saturday night.

Figure 1: News report of Ray Boylston's motorcycle accident with quasi-identifiers highlighted.

statewide project, the hospital's inpatient database of 650,000 records was made available for purchase. While the information was bought mainly by researchers and insurance companies it was available to anyone who wished to buy it.

This allowed a researcher who had access to the data to conduct a demonstration attack. By scanning the local news items from the same year as the hospitalizations, he was able to find the report of Ray's accident in the Spokesman-Review. The report contained key identifying information such as Ray's gender, age, admission date and cause of trauma. This same information was contained in the de-identified hospital data and provided enough details that the researchers were able to pinpoint Ray's record.

The concern here is not that the researcher was able to attach Ray's name to his record in the dataset – the fact that he had had a motorcycle accident was already public knowledge. The problem arises from the fact that the inpatient database contains additional information about Ray's hospital stay; information that we can learn about him that was not part of the public report and that he may not want to be known. Addressing this problem requires gaining a better understanding of how to manage re-identification risk.

Managing Re-identification Risk

Re-identification risk is measured by finding the unique combinations of quasi-identifiers in a de-identified dataset. Ray Boylston's re-identification was facilitated by the fact that his record presented a unique case within the data. A unique combination of quasi-identifiers is referred to as an equivalence class.

Equivalence Classes

The size of the smallest equivalence class in the dataset is a key factor to determine the risk for re-identification. To illustrate this, we will look at a simple de-identified dataset containing three quasi-identifies: gender, year of birth (which has been generalized to decade) and nationality. The dataset also contains the person's results for a genetic mutation test, which is sensitive personal information.

In this dataset, we have three equivalence classes:

1) American males born in the 1930's (records 1 and 3)

The Risks of Disclosing Personal Data

De-identified Database				
Quasi Identifiers			Sensitive Data	
ID	Gender	Year of Birth	Nationality	Gene Mutation
1	M	1930-1939	American	-ve
2	F	1960-1969	American	+ve
3	M	1930-1939	American	+ve
4	F	1960-1969	American	+ve
5	M	1960-1969	American	+ve
6	F	1960-1969	American	+ve

Table 1: De-identified database with three equivalence classes

2) American females born in the 1960's (records 2, 4 and 6)

3) American males born in the 1960's (record 5)

Let us assume the attacker is trying to find the results of the genetic mutation test of her friend Sue Storm. She knows that Sue is an American female born in the 1960's. In this de-identified dataset, the attacker has a 33% chance (1/3) of identifying Sue's record correctly. However, if the attacker is trying to find the test results for her colleague Peter Parker, an American male born in the 1960's, then there is a perfect match since only one record exists in that equivalence class. The attacker now knows for certain that Peter's genetic mutation test was positive.

Since we cannot know for certain which equivalence class the attacker will attempt to match, we must assume the worst-case scenario – that the person they want to re-identify is a member of the smallest equivalence class. When de-identifying a dataset in the real world it is recommended to have a minimum of five records in the smallest equivalence class. By making the probability of successfully re-identifying a record very small, we reduce the interest of the attacker.

Three Types of Re-identification Attacks

With a better understanding of re-identification risk, we can now look more specifically at the risks posed by different types of attacks. These are referred to as prosecutor risk, journalist risk and marketer risk.

In both the prosecutor and journalist scenarios, the attacker is attempting to re-identify a specific individual in a de-identified database. In the marketer scenario, the attacker wants to re-identify as many individuals as possible in a database.

Three Types of Re-identification Attacks

Public Database				
	Direct Identifier	Quasi-Identifiers		
ID	Secret Identity	Gender	Year of Birth	Nationality
1	Bruce Wayne	M	1939	American
2	Barbara Gordon	F	1961	American
3	Clark Kent	M	1938	American
4	Natasha Romanoff	F	1964	American
5	Peter Parker	M	1962	American
6	Sue Storm	F	1961	American
7	Reed Richards	M	1961	American
8	Billy Batson	M	1939	American
9	Bruce Banner	M	1962	American
10	Janet Van Dyne	F	1953	American
11	Tony Stark	M	1963	American
12	Clint Barton	M	1964	American

Table 2: Publicly available database

Prosecutor Risk

In this scenario, the attacker wants to re-identify someone they know and whose information is known to exist somewhere within the dataset. This is the situation we witnessed above in the discussion of equivalence classes where our attacker looked for Sue Storm and Peter Parker in the data. The pre-existing knowledge she has about these individuals enables her to search for them in the data and potentially learn additional information (e.g. whether or not they have a genetic mutation).

Even though we are looking at the re-identification of a single record, every record is potentially at risk since we cannot predict which one will be targeted for re-identification. As a result, it is reasonable to apply the average risk of all the equivalence classes in the set as the overall risk of re-identification.

Journalist Risk

This is similar in nature to prosecutor risk since it targets a single record, but the journalist does not know for certain whether a specific individual exists in the de-identified dataset. The attacker in this case has access to a separate source of information, such as a public database, which includes some or all of the people that also exist in the de-identified dataset. In the case of journalist risk, the attacker is looking to match individuals from the public data to the de-identified data but is not particularly concerned

Equivalence Class			De-id Database		Public Database	
Gender	Year of Birth	Nationality	Count	ID	Count	ID
M	1930-1939	American	2	1,3	3	1, 3, 8
F	1960-1969	American	3	2,4,6	4	2, 4, 6, 10
M	1960-1969	American	1	5	5	5, 7, 9, 11,12

Table 3: Registry Table showing equivalence class counts

Three Types of Re-identification Attacks

with who they are able to re-identify.

The risk profile, in this case, differs from that of the prosecutor risk. Here, the smallest equivalence class found in the publicly available database that maps to the de-identified dataset measures the risk of re-identification.

Table 1 (on Page 4) is the de-identified version of a dataset. The individuals in that dataset are a subset of the individuals who are also members of the larger, public database shown in Table 2.

As noted above, there are three equivalence classes in the de-identified dataset of Table 1. These equivalence classes can now be mapped to the records in Table 2, giving us the following registry table.

Table 3 above shows that the smallest equivalence class in the public database that maps to the de-identified dataset is a male born in the 1930's, where we have 3 records. Therefore, there is a one in three chance (33%) of correctly re-identifying a record that is part of this equivalence class.

The challenge in assessing journalist risk in the real world is that the entire content of a public database is rarely known.

Equivalence Class			De-id Database		Public Database		Expected Correct Matches
Gender	Year of Birth	Nationality	Count	ID	Count	ID	
M	1930 - 1939	American	2	1,3	3	1,3,8	2/3 (0.67)
F	1960 - 1969	American	3	2,4,6	4	2,4,6,10	3/4 (0.75)
M	1960 - 1969	American	1	5	5	5,7, 9,11,12	1/5 (0.20)
Expected number of identified records							**1.62**

Table 4: Expected Correct Matches for Marketer Risk

Marketer Risk

With marketer risk, the attacker wishes to re-identify as many individuals as possible in the database and is unconcerned if some of the records are misidentified. Here, the risk pertains to everyone in the dataset.

As an example, the research company Oscorp Industries has purchased the de-identified dataset.

They can then try to match this data to their own internal database in order to create a marketing campaign targeting genetic mutants. The marketer is not concerned if some of the recipients of the campaign receive the information in error. For the purpose of this example, we'll assume the marketer's database is the same as the public database shown in Table 2. The marketer risk is based on the probability of matching a record from an equivalence class in the de-identified dataset with one in the same

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

equivalence class in their own database. The expected number of records that the marketer can properly identify in the de-identified dataset is shown in the Expected Correct Matches column in Table 4. To determine the overall probability of correctly re-identifying any record in the de-identified dataset, we need only add up the probabilities for each equivalence class and divide it by the number of records in the de-identified dataset; in this case, approximately 27% (1.62/6).

Conclusion

Re-identification attacks pose serious risks not only for the organizations that have been attacked but also for the individuals whose personal information has been compromised. Once a person has been identified in a database it is possible to learn new information about them. Such information could be used to discriminate against them based on their medical history, for example.

Attackers will try to re-identify data for various reasons including curiosity, for personal gain or simply to prove that they are able to do it. Understanding the risk of re-identification requires knowing the uniqueness of the information in the dataset and the various types of attacks that can be perpetrated. Commercial software solutions employ algorithms that can effectively measure the level of risk for each type of re-identification attack.

Complete the journey by reading the final paper in this series De-identification 401: An Optimal Approach to Data De-identification.

Appendix: Terminology

Dataset	A collection of related data records. Most commonly, a dataset refers to the contents of a database with many tables of data, where every column in the table represents a particular variable.
De-identification	A process that removes or suppresses, and/or alters personally identifiable information in a data collection so that it may be shared within the organization, with other organizations, or individuals for secondary purposes. This term is sometimes used interchangeably with the term anonymization.
Demonstration Attack	An attack on a dataset where the attacker wants to make a point of showing that individuals in a dataset can be re-identified. The attack is considered successful if even a single individual is correctly re-identified.
Direct Identifier	The fields within a dataset that can easily be used alone to uniquely identify individuals. This includes information such as name or email address.
Equivalence Class	A group of unique combinations of identifiers in a dataset. For example, all 40-year-old female engineers in a dataset could be members of a single equivalence class; 40-year-old male engineers would be members of a separate equivalence class.
Journalist Risk	The risk of re-identification posed by an attacker attempting to re-identify a single individual in the data using information they know about the individual, usually a family member, friend, colleague or a well-known person. The attacker does not know with certainty that the individual they are trying to identify is in the dataset.
Marketer Risk	The risk of re-identification posed by an attacker who attempts to re-identify as many individuals as possible in a dataset. The attacker is not concerned if some of the individuals are incorrectly re-identified only that as many of the individuals as possible are re-identified.
Medical Identity Theft	Occurs when someone uses an individual's name and personal identity in order to fraudulently receive medical services, prescription drugs or other goods used for health and well-being. This includes attempts to commit fraudulent billing.
Protected Health Information (PHI)	Health data that can be used to uniquely identify or locate an individual. Examples of protected health information include health plan numbers, disease diagnoses, hospital admissions information or lab results.
Privacy Breach	The result of unauthorized access to, or collection, use and disclosure of personally identifiable information for one or more individuals whose information is contained in a dataset. Privacy breach may be inadvertent (security vulnerability) or intentional (hacker).
Prosecutor Risk	The risk of re-identification posed by an attacker attempting to re-identify a single individual in the data using information they know about the individual, usually a family member, friend, colleague or a well-known person. The attacker knows with certainty that the individual they are trying to identify is in the dataset.
Quasi-identifier	Fields within a dataset that can be used in combination with one another to identify individuals. For example, birth date or postal code. Quasi-identifiers are also referred to as indirect identifiers.
Re-identification	The identification of a unique individual within a dataset that was supposed to have been de-identified.

Sources

1 McCann, Erin. (2015, March 12). HIPAA breaches: The list keeps growing. Healthcare IT News. Retrieved from http://www.healthcareitnews.com/news/list-biggest-hipaa-data-breaches-2009-2015

2 Ponemon Institute LLC. (2015, January). Fifth Annual Study on Medical Identity Theft. Ponemon Institute Research Report. Retrieved from http://www.identityfinder.com/us/Business/Downloads/WhitePapers

3 Abelson, Reed & Julie Creswell. (2015, February 6). Data Breach at Anthem May Forecast a Trend. New York Times. Retrieved from http://www.nytimes.com/2015/02/07/business/

4 Hill, Kashmir. (2010, October 08). Celebrity Data Breaches. Forbes. Retrieved from http://www.forbes.com/2010/10/08/lohan-obama-paris-hilton-business-entertainment-celebrity-data-breaches.html

5 Dawes, Terry (2015, March 19). Ottawa's Privacy Analytics Participates in HITRUST Health De-identification Framework. Cantech Letter. Retrieved from http://www.cantechletter.com/2015/03/ottawas-privacy-analytics-participates-in-hitrust-health-de-identification-framework/