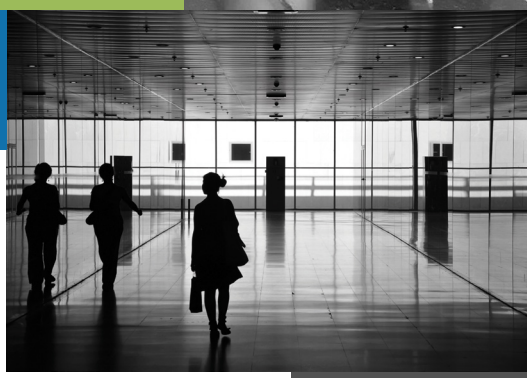
PRIVACY ANALYTICS

WHITE PAPER

The Top 5 Drawbacks to Using Only Data Masking

Although data masking and de-identification are often grouped together for discussion, the two use different approaches in making data anonymous. Masking is used to anonymize direct identifiers while de-identification is used to anonymize quasi-identifiers. In practice, masking and de-identification should be used together to optimize the balance between protecting privacy and maintaining the usefulness of the data. This paper explores the major limitations of using data masking on its own, without de-identification.



PRIVACY ANALYTICS

a QuintilesIMS company

Introduction

REAL LIFE EXAMPLE OF DATA MASKING BY ITSELF:

An organization has replaced patient identifying information in a database by creating pseudonyms, which is a data masking technique. Unfortunately, a data breach occurred and that database was lost. During the subsquent investigation the regulator working on the file concluded that notwithstanding the fact that pseudonyms were utilized, there were other demographic and diagnosis fields in the database that showed the data was still be protected health information (PHI). The risk of re-identification of the patients was still quite high. Now the organization will incur the breach cost of an estimated $208 per affected individual.

The goal of de-identification is to ensure that data cannot be matched to the person it describes so that their privacy is protected. Many people assume that simply masking data and removing names, addresses and other identifiers like Social Security Number should be sufficient to make information anonymous. However, data contains other personal details that, while not obviously identifying, can be used to re-identify a person. This includes information like date of birth, marital status, and occupation.

Data Masking and De-identification are Often Treated as Interchangeable Terms, But This is Not True.

Data masking is part of the broader de-identification process. By only applying masking techniques, data custodians limit the use of the data and opening themselves up to unnecessary risk. Here are our top five drawbacks to using data masking on its own.

1. Data Masking Only Deals With Direct Identifiers

Masking refers to a set of techniques that attempt to eliminate direct identifiers. Direct identifiers are data fields that can be used alone to uniquely identify individuals. This includes elements such as name, email address or Social Security Number, where each of these is generally associated with only one person. Typically, direct identifiers are not used in statistical analyses that are run on health data.

Quasi-identifiers, or indirect identifiers, are fields that can identify individuals but are also useful for data analysis. Examples of these include dates, demographic information, such as race and ethnicity, and socioeconomic variables, like occupation and income. This distinction is important because the drawback of dealing with only direct identifiers is that the risk exposure from the indirect identifiers remains.

2. Masking Effectively Eliminates The Analytic Utility

Many of the masking techniques that are commonly used destroy

the data utility of the masked fields. Masking should only be used on fields that will not require any analytics. Consequently, this means that any relationships among masked variables or between masked and non-masked variables are lost.

Masking techniques should not typically be applied to dates or geographic information because these fields are often used in data analysis. Masking dates could replace them with null values (eg. 00-00-00), which renders them useless for any further analysis.

Similarly, geographic data that is masked makes it difficult to perform any analysis using those fields. A zip code points to a specific state, county and town. Masking part of a person's address, such as a zip code, without consideration of city and state, may render the data unusable. Some approaches leave only the first three digits of the zip code visible. The downside is that this makes the data less useful, for example, the area in the three-digit zone is geographically large, making analysis less exact.

With some masking techniques, such as shuffling, it is possible to have accurate summary statistics about a single field at a time; but this does not hold when you want to look at relationships between variables. For the purpose of most data analytics, this is quite limiting. Masking tends to rely on techniques that get rid of data, distorting the information and reducing the data's usefulness

3. Masking is Not Based on Risk Measurement

Masking techniques do not use metrics to measure the actual risk of re-identification. Therefore, it is not always possible to know whether the transformations performed on the data were considered sufficient to anonymize it and, thus, defensible. Not using metrics is only acceptable if the masking method is guaranteed to ensure a low probability of re-identification.

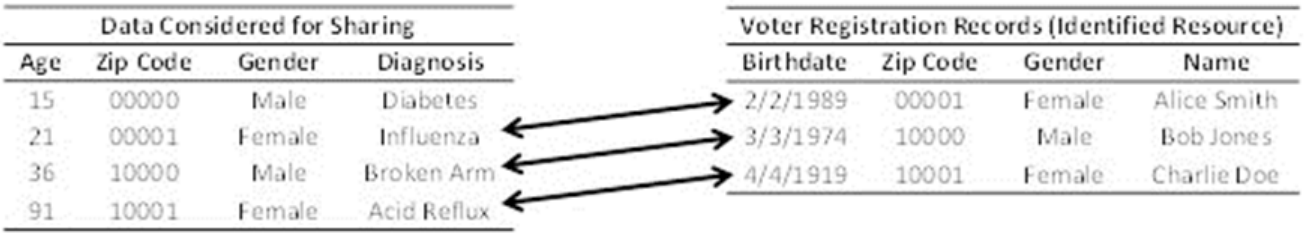
In some instances, we know that the probability of re-identification will be very small. For example, if we do a random replacement of first names in a database that is large (say 10,000 records) and the replacement names are allocated using a uniform distribution, then the probability of guessing the correct name for any record is 1/10000. This is a very small probability and the risk of reverse engineering the randomized names is negligible. The same can be said for the replacement of facility names and addresses. Randomization can be a useful masking technique for gaining authentic looking data, however, risk measurement is still needed to ensure the right amount of randomization is being used. Without knowing the risk in the data, it's easy to over or under randomize.

> Masking should only be used on the fields that will not require any analytics.

More often, combining masking with de-identification techniques provides the risk-measurement-based approach that is needed to safeguard privacy. The excerpt below (Figure 1) describes a scenario where a hospital dataset is matched with the voter registration records. Using a risk-based approach ensures that the correct techniques are used and provides for the best protection. As stated by the HHS, "Patient demographics could be classified as high-risk features. In contrast, lower risk features are those that do not appear in public records or are less readily available."

Example Scenario

"Imagine that a covered entity is considering sharing the information in the table to the left in Figure 1. This table is devoid of explicit identifiers, such as personal names and Social Security Numbers. The information in this table is distinguishing, such that each row is unique on the combination of demographics (i.e., Age, Zip Code, and Gender). Beyond this data, there exists a voter registration data source, which contains personal names, as well as demographics (i.e., Birthdate, Zip Code, and Gender), which are also distinguishing. Linkage between the records in the tables is possible through the demographics. Notice, however, that the first record in the covered entity's table is not linked because the patient is not yet old enough to vote. Thus, an important aspect of



Data Considered for Sharing			
Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Identified Resource)			
Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

Figure 1 - Linking two data sources to identity diagnoses. Source: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

identification risk assessment is the route by which health information can be linked to naming sources or sensitive knowledge can be inferred."[1]

4. Using Masking, it is Not Always Possible to Know Whether the Transformations Performed on the Data Were Sufficient and Defensible

Data masking methods are not necessarily protective of privacy. Protecting against identity disclosure is a legal or regulatory requirement. Complying with the law means that a dataset must not contain personal information when disclosed for secondary purposes without patient consent or authorization. The HIPAA

Privacy Rule states, "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information."[2] A data custodian may put their organization in a position of non-compliance that risks legal action by using certain masking techniques, because these techniques do not use metrics to measure the actual risk of re-identification.

There will also be situations where data masking can result in data releases where the risk of a privacy breach is high.

Methods like cropping should not be used for masking because you cannot know whether the data has received the correct level of protection. Without metrics, an analyst may over or under-crop. The problem is that the organization may find this out at the worst possible time – once a breach has occurred.

5. Masking cannot protect all of the fields in a typical health dataset.

The aim of de-identification is to do as little as possible to alter the data while still effectively making the information anonymous.

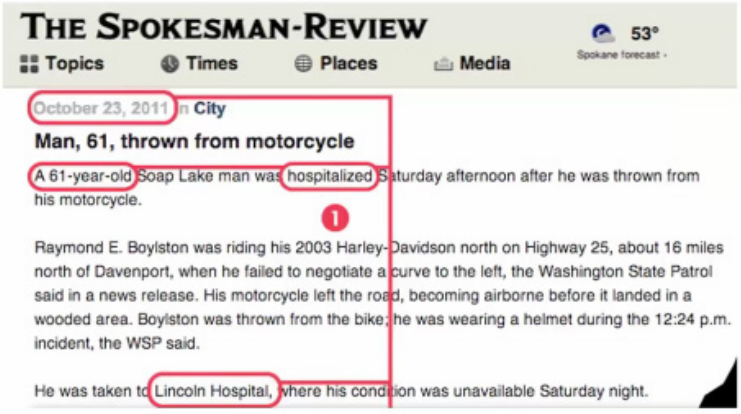
De-identification uses techniques like record suppression, cell suppression, sub-sampling and aggregation to transform the data values while minimally distorting the data. Both masking and de-identification together are needed to protect all of the fields in a typical health dataset.

Masking Gone Wrong

In 2011, a Vietnam veteran named Ray Boylston had a motorcycle accident in Washington State after suffering a diabetic shock while riding. The incident was covered briefly in the local paper (Figure 2).

The paperwork relating to his week-long stay in hospital was added to a database of 650,000 hospitalizations at that particular hospital during the year, which was masked then made available for purchase.

Generally, the market for the resale of health information consists of researchers and insurance companies, but the information is there for anyone who wants to buy it. All a hacker needs is a tiny piece of information to trace the identification trail back to a particular individual. In this case, the newspaper details filled in the gaps needed to re-identify Mr. Boyston's record in the dataset.


THE SPOKESMAN-REVIEW

53°

Topics Times Places Media

October 23, 2011 in City

Man, 61, thrown from motorcycle

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

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 2. Extract of a news story that contains name, age, residential information, hospital, incident date, and type of incident. [Spokesman Review. 10/23/2011]

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

There are many data masking techniques available today, which promise quick methods to deal with privacy and compliance. However, both masking and de-identification are typically required to actually provide meaningful privacy protections.

> Aside from merely identifying a patient through a publicly available database, such information can also be used to discriminate against an individual based on residency in a "risky" area code, or through knowledge of a patient's medical history, for example. "If they're going to release that kind of information, they should consult with the patient," Boylston told Bloomberg Business. "That's personal information about me. It's just not right."[3]

Conclusion

Data masking, when used as the sole means of privacy protection, has major drawbacks in terms of legislative compliance and data utility. The problem of dealing solely with the direct identifiers is that the risk exposure still remains from the indirect identifiers.

Nearly all datasets consist of both indirect identifiers and direct identifiers. In practice, it is important to apply both data protection techniques: masking and de-identification. Data masking is only part of the solution to the puzzle.

To learn more about right way to unlock the value of health data, make sure to visit Privacy Analytics' De-Id University.

SOURCES:
1. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html
2. http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/
3. http://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy