University Library
Technical University of Munich

Data Anonymization Techniques

Recommendation Guide

Introduction

Anonymizing research data might be a difficult task. Thus proper preparation and counter-measures are important to maintain the privacy of your research participants.

Why Anonymization

Even though the data you are collecting might not be sensitive or identifying, i.e. don't contain medical data, name and address or others, people might be able to enrich your data with their own data and draw conclusions about sensitive or identifying information. For example if you publish the zip code, birth date, sex and some medical information about an anonymous subject A, and someone else enriches this information with zip code, birth date, sex as well as name and address of the public civil registry, there is a high chance that you can combine the name with the medical information of subject A, given that there is only one person per household with the same age. So everyone in possession of the data knows the medical information of subject A, as well as who subject A is – as it happened in a case where medical information of the governor of Massachusetts was revealed[1]. This is called deanonymization.

Deanonymization is difficult to prevent, because you never know which additional data there might be in the future. To tackle deanonymization, researchers came up with some counter-measures.

Attribute Types

In your dataset you often have three different types of attributes. These are *key attributes* (e.g. primary keys or unique-by-definition attributes), *sensitive attributes* you want to prevent being related to an identity, and *quasi-identifiers*. Quasi-identifiers are the groups of identifiers which might lead to an identification of research subjects, because they have a high correlation in numbers with key attributes – like in the example above, the combination of zip code, birth date and sex, identified a single record in the public civil registry database.

1 Latanya Sweeney (2002) k-anonymity: a model for protecting privacy; International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), p. 557-570

K-Anonymity[2]

The information in the released table cannot be distinguished from at least *k-1* individuals, i.e. there are at least *k* people with the same quasi-identifier. Or in other words: The table is *k*-anonymous, if each quasi-identifier appears in at least *k* table records.

Name	Zip Code	Age	Diagnosis
Alice	12345	<30	Gastritis
Bob	12345	<30	Stomach Cancer
Charlie	54321	45	Flu
Daniel	54321	17	Flu

Table 1. Green: 2-anonymous quasi-identifiers, Red: non-anonymous quasi-identifiers

Techniques: Generalisation can be used to keep datasets but make them more general (s. green example regarding age). *Suppression* means to leave out datasets (e.g. removing the red rows).

Flaws: The dataset might have no diversity, so if every diagnosis is flu, you can be sure that Daniel in above's example has flu. Also it does not take account for background knowledge, e.g. if Charlie's neighbor knows that Charlie is the only person at an age of 45, he can conclude that it must be Charlie who has flu.

L-Diversity[3]

Both above mentioned flaws in k-anonymity can be mitigated by introducing artificial diversity. L-diversity means that each set of entries with identical quasi-identifiers has at least *L* different sensitive values. So an attacker would need *L-1* identified rows as background knowledge.

Name	Zip Code	Age	Diagnosis
Alice	12345	<50	Gastritis
Bob	12345	<50	Stomach Cancer
Charlie	54321	<50	Flu
Daniel	54321	<50	Flu

Table 2. Green: 2-diverse datasets, Red: non-diverse datasets

2 Latanya Sweeney (2002) Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, Nr. 05, pp. 571-588, doi.org/10.1142/S021848850200165X

3 Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam (2006) L-diversity: privacy beyond k-anonymity; 22nd International Conference on Data Engineering (ICDE'06), pp. 24-24, doi.org/10.1109/ICDE.2006.1

Flaws: Sometimes it is very difficult to achieve anonymity while maintaining *data quality* e.g. with test results only having the value positive (1%) or negative (99%). Another problem might be *data skewness*, if you know that the person to be deanonymized is in a section of the table, where the probability of a positive test is very high (e.g. 90%), but in the overall table low (e.g. 1%). So assume the people are sorted in order of participation, and 90% of the first 100 people have a positive test, and you know, that the person to be deanonymized is among the first 100 participants, you can say that the person has a 90% chance of having a positive test, even though the overall chance is just 1%. Furthermore L-diversity does not account for *similarity* attacks, i.e. gastritis and stomach cancer are both classified as stomach diseases. So an attacker would know the disease class of every participant with the same quasi-identifier.

T-Closeness[4]

The problem of skewness can be taken into account with t-closeness, which ensures that the difference of distributions of a sensitive attribute in the whole table compared to arbitrary table blocks is less than a distance metric *t*. To assimilate the distributions, you may have to swap rows between table blocks. When using categorial data, it is might be difficult to find proper distance metrics. One approach is to fit categories into a hierarchy, and calculate how many hierarchy levels you have to traverse to reach the other category. The number of levels is your distance.

Differential Privacy[5]

Differential Privacy adds noise to your data while maintaining the accuracy of empirical results. For example half of the participants' real answers is stored, the other half of the participants' answers is replaced with random values sampled from a certain distribution (e.g. binary laplace for boolean answers, so *P(false)=50%* and *P(true)=50%*). This creates plausible deniability for individuals, because you cannot tell whether the individual really answered *true* or *false*, or whether the answer was substituted. At the and you can take the noise into account for calculating the number of *true* values.

Flaws: The privacy of each individual highly depends on the probability distribution you apply. Otherwise it might be possible to use a skewness attack to predict the answers of each individual quite accurately.

4 Ningui Li, Tiancheng Li, Suresh Venkatasubramanian (2007) t-closeness: privacy beyond k-anonymity and l-diversity", Proceedings of the IEEE 23rd International Conference on Data Engineering. ICDE'07

5 Cynthia Dwork, Aaron Roth (2014) The algorithmic foundations of differential privacy; Foundations and Trends in Theoretical Computer Science 2.3, p. 211-407

Author
Maximilian Josef Frank

based on the contents of the lecture „Security Engineering“ (2023)
by Prof. Dr. Alexander Pretschner, Technical University of Munich

Technische Universität München
Universitätsbibliothek

Arcisstraße 21
80333 München
www.ub.tum.de