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 person, 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 that person, given that there is only one person per household with the same age. So, everyone in possession of these data knows the medical information of that person and can identify them. In fact, this happened to the governor of Massachusetts, where his medical information was revealed[1]. This is called deanonymization.

Deanonymization is difficult to prevent, because you generally don't 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 a person or the research subject, 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 with or even one-by-one map to 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]

K-Anonymity means that 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: Generalization can be used to keep datasets but make them more general (see green example regarding age). *Suppression* means to leave out datasets (e.g. removing the red rows).

Flaws: The dataset might have no diversity. If every diagnosis is flu, you can be sure that Daniel in the above example has flu. Also, it does not account for background knowledge. For example, 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 to have identified *L-1* 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*. This is especially the case with rare occurrences in your data, e.g. test results being positive in only 1% of all cases. 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 this probability is 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. If 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 still know the disease class (i.e. any stomach disease).

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 categorical data, a distant metric could be to fit categories into a hierarchy and count how many hierarchy levels one has to traverse to reach the other category. The number of levels is then the distance.

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

Table 3. A 2-close table. The distance between “Gastritis” and “Stomach Cancer” is 1 (both stomach diseases), and their distance to “Flu” is each 2. Shuffling the table or sampling arbitrary table blocks yields the same maximum distance of 2.

[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

Differential Privacy[5]

Differential Privacy encompasses many different techniques which add noise to your data while maintaining the accuracy of empirical results.

With the randomized response method, half of the study participants' real answers is stored and the other half is replaced with random values sampled from a certain distribution (e.g. binary Laplace for boolean answers, so $P(yes)=50\%$ and $P(no)=50\%$). This creates plausible deniability for participants, because you cannot tell whether a participant really answered *yes* or *no*, or whether the answer was at random. At the end, the statistical noise is taken into account for calculating the true number of values.

For example with binary Laplace answers, participants reporting their real answer ($P(truth)=50\%$) would always respond *yes*, if this would be their true answer. If they respond randomly ($P(random)=50\%$), the probability of responding with *yes* or *no* is each 50%. So overall, if a participant would have originally answered *yes* honestly (e.g. because the participant has a certain attribute, disease, …), they now respond *yes* with a probability of 75% and *no* with 25% under the randomized response method. Let p be the true proportion of people who would report *yes* (e.g. they have a certain attribute, disease, …). Then there are $\frac{1}{4}(1-p) + \frac{3}{4}p$ positive responses (incl. the random ones). Solving this formula for p gives $p = 2\cdot\#yes - \frac{1}{2}$, and hence the true proportion of people with a certain attribute, disease, ….

Flaws: The privacy of each participant highly depends on the probability distribution you apply. Choosing a skewed distribution may enable skewness attacks to predict the answers of each participant quite accurately.

[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

DOI: 10.14459/2024md1735453 *(most recent version)*

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

Technical University of Munich
University Library

Arcisstraße 21
80333 Munich
www.ub.tum.de