Data anonymization

Josep Domingo-Ferrer,
Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Center for Cybersecurity Research of Catalonia (CYBERCAT)

Jordi Soria-Comas,
coordinator of Technology and Security at the Catalan Data Protection Authority (APDCAT)



Abstract

Data anonymization aims at generating a statistically valid data set whose records can no longer be linked to specific individuals. The trade-off between privacy and the amount of information conveyed by each record makes data anonymization a complex topic in the current context in which there are vast amounts of data about each individual. This is recognized by the GDPR, which considers pseudonymization as a fallback. Data anonymization is still possible, but we must focus on specific analyses and on specific threats. We review the difficulties of data anonymization and the main approaches to it.

Data minimization is one of the principles in the EU's General Data Protection Regulation (GDPR). This basically says that personal data processing should be kept to the minimum necessary to meet the data processing purpose. This means, for instance, that we should aim to reduce the number of individuals whose data is processed, and the number of attributes and their level of detail. Where possible, techniques designed to hide the relationship between data records and the people that generated them should be used; primarily, pseudonymization and anonymization.

Anonymization has other advantages beyond satisfying the data minimization principle. With the goal of protecting people's rights, the GDPR introduces many limitations and obligations, which may make some data processing tasks cumbersome. For example, even when using personal data that is publicly available (open data), the data controller must take the necessary steps to inform data subjects. If anonymized data can be used instead of personal data, the GDPR's provisions no longer apply, which makes data processing much simpler.

However, anonymization is not a silver bullet that can be systematically used to bypass the GDPR. It has many downsides

If anonymized data can be used instead of personal data, the GDPR's provisions no longer apply, which makes data processing much simpler. However, anonymization is not a silver bullet



that have to be taken into consideration. ~~Among these, the limited utility loss of anonymized data stands out as especially problematic.~~

Anonymization and pseudonymization in the GDPR

The GDPR uses the term *personal data* to refer to any information related to an identified or identifiable data subject. Identification can be performed directly (via attributes such as name, email address and identification number) or indirectly by combining the values of attributes that we know correspond to a particular person. Although surprising at first, many common attributes have a strong potential for re-identification. For example, a study on the 2000 US census

Although surprising at first, many common attributes have a strong potential for re-identification

shows that 63% of the population is uniquely identifiable by their combination of ZIP, sex, and birth date. If we reduce the accuracy of the birth date to month and year, the percentage of identifiable people is reduced to a (still significant) 4.7%. If such little information can re-identify so many people, what could we not do with the large amount of personal information that current databases hold?

Two main techniques are used to reduce the risk of re-identification while still allowing us to analyse the data: pseudonymization and anonymization. Data encryption, another technique mentioned in the GDPR, does not fit in this category because it thwarts any data analysis.[1]

Both pseudonymization and anonymization have the same goal: to reduce the risk of leaking personal data while preserving the analytical validity of the data as much as possible.

Pseudonymization replaces identifying attributes with pseudonyms. The use of pseudonyms makes it possible to hide the identity of the subjects to whom the records correspond while maintaining the coherence of the data. The relationship between original values and pseudonyms is maintained by the pseudonymizer separately from the pseudonymized data set. In this way, only the pseudonymizer

Both pseudonymization and anonymization have the same goal: to reduce the risk of leaking personal data while preserving the analytical validity of the data as much as possible

1 Some advanced encryption techniques, such as homomorphic encryption, may allow some reduced data analyses to be conducted (such as computing a sum over encrypted data). However, because of the limited analysis capabilities they offer, most of the time they are not a feasible solution.

can reverse the process. For example, name, email, and IP address can be pseudonymized by replacing each value with a unique random value.

Although data anonymization techniques are much more diverse than pseudonymization techniques, their purpose is the same. According to the GDPR, the difference between pseudonymization and anonymization is that anonymization entails a process that cannot be reversed. In other words, once a data set has been anonymized, there should be no way to re-identify any particular record.

According to the GDPR, the difference between pseudonymization and anonymization is that anonymization entails a process that cannot be reversed

However, the reality is that both anonymized and pseudonymized data sets may be the target of attacks by intruders seeking to find out personal information. The GDPR acknowledges this fact by requiring data protection to take into account all the means reasonably available to directly or indirectly identify the natural person. This statement suggests that the GDPR is more concerned with the risk of re-identifying the subjects of records (*re-identification disclosure*). However, properly protecting the privacy of individuals also requires us to prevent intruders from learning confidential attributes of subjects (*attribute disclosure*). In fact, attribute disclosure can occur without re-identification disclosure and vice versa. Re-identifying the subject of a record that contains no confidential attributes ~~(e.g. a record with only address, age and gender)~~ does not result in attribute disclosure. On the other hand, if a confidential attribute, say

Properly protecting the privacy of individuals also requires us to prevent intruders from learning confidential attributes of subjects



salary, has very similar values in a group of records, it is enough to know that your neighbour's record is in that group to accurately estimate your neighbour's salary: thus, in this case, attribute disclosure occurs without re-identification.

Preventing record re-identification is hard

To prevent record re-identification attacks, we must make sure that the values of the attributes that are externally available are appropriately masked. Otherwise, they could be used to mount a re-identification attack. However, several famous privacy breaches show us just how difficult it is to get this right.

In 2006, AOL published a list of 20 million web search queries performed by its users. The goal was to make the data available for research. As AOL wanted to preserve the privacy of its users, it only published a random identifier followed by the list of queries performed by that user. AOL believed that the absence of personal features would prevent any re-identification. What AOL did not take into account was that oftentimes the queries posted by a person contain personal information details that may lead to re-identification. Indeed, this is what two journalists from *The New York Times* exploited (Barbaro & Teller, 2006): they searched for a user whose queries contained enough personal information that could be useful in re-identification (i.e. enough personal information that was publicly available). The rest of queries not used in the re-identification process revealed that user's personal interests.

In 2006, Netflix launched a contest endowed with a one-million-dollar prize to improve its recommender system. To support the contest, Netflix published





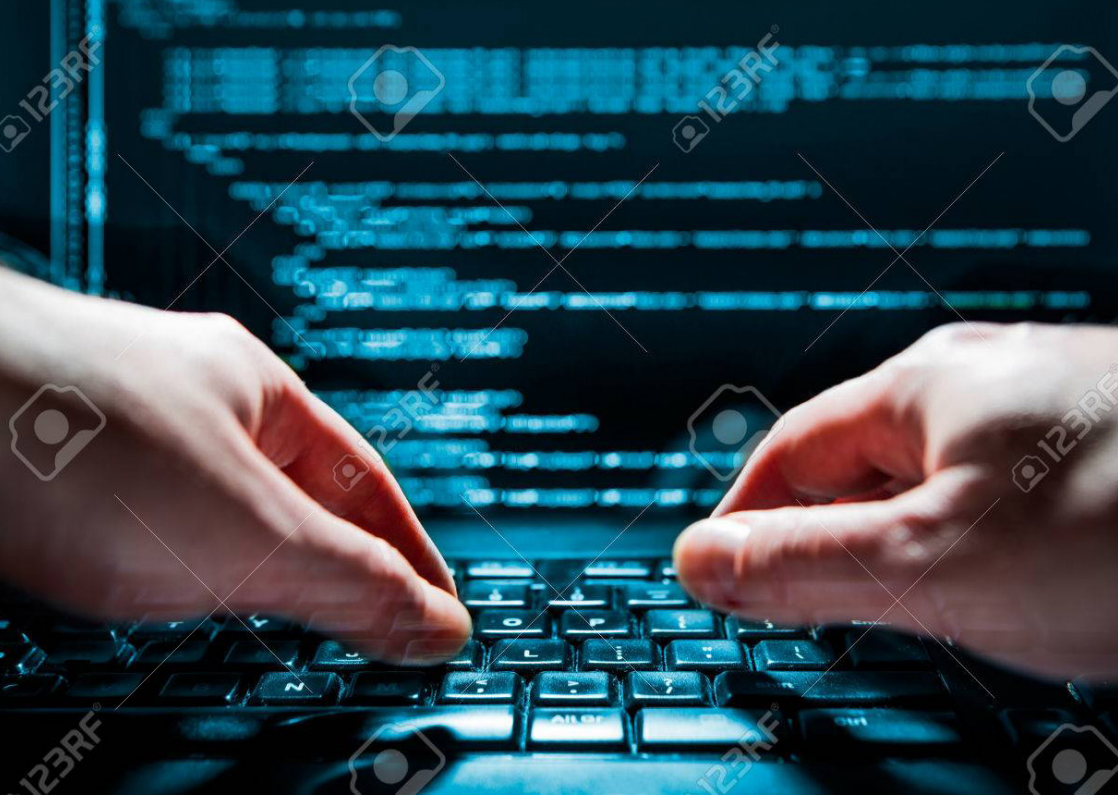
a data set containing 100 million movie ratings submitted by around 500,000 customers. To protect subscribers' privacy, only the movie ratings and dates were published. Netflix argued that all customer-identifying information had been removed. According to Netflix, the rating posted by a customer was not enough to re-identify them. However, two researchers (Narayanan & Shmatikov, 2008) showed that a small number of ratings was enough to re-identify people, even if some of the ratings or the dates were wrong. They took advantage of the publicly-available movie ratings in IMDB to re-identify users in the Netflix data set.

Anonymization

Anonymization is a data masking process. Given an original data set, we want to generate data that remain useful for analyses and preserve people's privacy.

Anonymization techniques are subject to a trade-off between privacy and information loss. To increase privacy protection, we need to apply a more thorough masking, which degrades the utility of the resulting data.

Anonymization is a data masking process. Given an original data set, we want to generate data that remain useful for analyses and preserve people's privacy



Anonymization techniques

We can broadly classify anonymization techniques according to the approach used to protect the data in the following categories (for further details see Hundepool *et al.*, 2012):

- **Data masking** techniques alter the content of the original data records. They can be further classified according to the truthfulness of the masked data in perturbative and non-perturbative masking.
  - **Non-perturbative masking** techniques reduce the amount of information in the original data; for instance, by

replacing exact numerical values with intervals.

- **Perturbative masking** techniques alter the truthfulness of the data. For instance, noise addition adds some random noise to numerical values, microaggregation groups records and replaces each group with a representative value of the group (such as the group average).

- **Synthetic data generation** aims at generating a new data set from scratch that preserves some of the statistical properties of the original data. The fact the synthetic records are based on a statistical model of the original data (and do not have a direct dependence on original data records) reduces the risk of record re-identification. However, re-identification and attribute disclosure are both possible if the model is overfitted to the original data.

- **Tabular data protection** modifies the contingence and magnitude tables to be released in such a way that no individual information can be inferred from the tables. Among other rules, cells that correspond to a single subject should not be released.

- **Interactive query protection** limits a database's answers to interactive queries by users. The goal is to prevent users from isolating single records by chaining several queries.

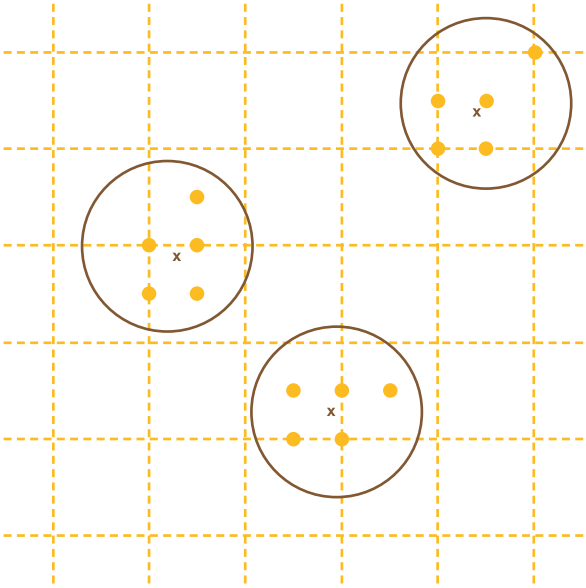


Figure 1. Example of perturbative masking: microaggregation. Records are grouped in clusters of at least k records and cluster averages are released instead of the original record. In the figure, k=5, original records are depicted as yellow dots and the average record of each cluster is depicted as a cross.

Although an attacker will use a different approach for each of the previous categories, all of them are subject to the trade-off between privacy and utility. Thus, regardless of the method, increasing the accuracy of the data increases privacy risks.

Privacy Models

One issue with many of the above techniques is that they do not provide guarantees with respect to the level of privacy protection attained. Privacy models have been introduced for this reason. A privacy model describes a



precondition that data must satisfy for privacy to be sufficiently protected. They come in different flavours.

k-Anonymity

k-Anonymity (Samarati & Sweeney, 1998) is a privacy model for data set releases. Its typical use is when the aim is to publish an entire data set (as, for example, in open data).

k-Anonymity seeks to prevent the re-identification of records. It assumes that re-identification occurs through a given combination of attributes, called quasi-identifiers. As a measure of protection, this privacy model requires each combination of values of these

A privacy model describes a precondition that data must satisfy for privacy to be sufficiently protected

attributes to be shared by a group of at least *k* records (this group is called a *k-anonymous class*). In this way it is not possible to re-identify a record with certainty: the probability of getting it right is at most 1/k.

Height (cm)	Weight (kg)	Blood pressure (syst. mmHg)	AIDS (Y/N)
175	76	117	Y
175	76	131	N
175	76	122	N
180	81	115	N
180	81	122	Y
180	81	146	N
190	95	110	N
190	95	115	Y
190	95	125	N
190	95	140	N

Figure 2. Example of a k-anonymous data set with k=3, quasi-identifiers height and weight, and confidential attributes blood pressure and AIDS (Y/N).



l-Diversity and t-Closeness

k-Anonymity protects against identity disclosure. However, attribute disclosure is still possible if the variability of confidential attributes among the records of a k-anonymous class is small. In other words, even if the intruder cannot determine whether a record corresponds to a certain person, they may be able to learn accurate information about that person.

To avoid this issue, other privacy models strengthen k-anonymity by requiring a certain minimum level of variability in the confidential attributes within each k-anonymous class. Among these we find l-diversity (Machanavajjhala *et al*., 2007) and t-closeness (Li *et al*., 2007).

Differential Privacy

Differential privacy (Dwork *et al*., 2006) is a privacy model designed to limit disclosure in database queries. The underlying idea is that privacy is well protected if the presence or absence of any record does not alter query responses in a significant manner. In other words, since my data did not alter the response in a significant manner, no intruder can use the response to infer anything about my data with enough confidence.

Differential privacy is widely recognized as offering strong protection. Indeed, it has some very beneficial properties. These notably include the fact that, unlike k-anonymity, it does not rely on any assumptions about the intruder's

Differential privacy is widely recognized as offering strong protection. Indeed, it has some very beneficial properties

knowledge, i.e. it is not necessary to assume a specific set of quasi-identifiers.

While the strong privacy guarantees of DP are very convenient, it has a significant downside: it has a significant impact on the utility of the data. Indeed, the number of real-world applications of differential privacy is very small, and most of them are very specific applications to well-behaved problems. Despite this observation, because of its academic clout, it is commonly advertised by many software products. When we analyse the details, we find that marketing has a strong component in these advertisements. The reality is that very often strong relaxations are made to the original notion of differential privacy in an attempt to keep the utility of the data within a reasonable range.

Keeping the risk low enough is necessary to comply with the GDPR, but to preserve data utility we should only aim for the level of protection that is necessary to deal with likely threats

Conclusions

We have briefly introduced the notion of data anonymization and described the motivations behind it. We have also described some anonymization techniques and privacy models.

There are several software packages available that implement these methods and models. The difficult part is not applying them, but instead to strike a good balance between privacy and utility. In other words, each of these methods can be applied more or less thoroughly, which has an impact on the privacy and the utility of the resulting data. Keeping the risk low enough is necessary to comply with the GDPR, but to preserve

data utility we should only aim for the level of protection that is necessary to deal with likely threats. ▮

References

Barbaro, Michael, & and Zeller Jr, Tom. (2006). **A Face is Exposed for AOL Searcher No. 4417749**. *The New York Times*. August 9, 2006.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, & Smith, Adam. (2006). **Calibrating Noise to Sensitivity in Private Data Analysis**. *Lecture Notes in Computer Science*, vol. 3, 876. Berlin, Heidelberg: Springer.

Hundepool, Anco, Domingo-Ferrer, Josep, Franconi, Luisa, Giessing, Sarah, Nordholt, Eric Schulte, Spicer, Keith, & de Wolf, Peter-Paul. (2012). ***Statistical Disclosure Control***. Hoboken: Wiley.

Li, Ninghui, Li, Tiancheng, & Venkatasubramanian, Suresh. (2007). ***t-Closeness: Privacy Beyond k-Anonymity and l-Diversity***. IEEE 23rd International Conference on Data Engineering, Istanbul, 106-115.

Machanavajjhala, Ashwin, Kifer, Daniel, & Gehrke, Johannes. (2007). L-*diversity: Privacy beyond* k-*anonymity*. *ACM Transactions on Knowledge Discovery from Data, 1* (1). **https://doi.org/10.1145/1217299.1217302**

Narayanan, Arvind, & Shmatikov, Vitaly. (2008). ***Robust De-anonymization of Large Sparse Datasets***. Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP'08), 111-125.

Samarati, Pierangela, & Sweeney, Latanya. (1998). ***Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression***. Cambridge (USA): Harvard Data Privacy Lab.