Personal Data Protection Commission, Singapore

Guide to Basic Data Anonymisation Techniques

Report Review

Anshul Aggarwal
anshulaggarwal987@gmail.com

1 Introduction

The Personal Data Protection Commission of Singapore published a detailed report[1], describing the various technical aspects of anonymisation, and how organizations must deal with data to ensure the privacy of the users is not at risk. It describes in detail the various scenarios and approaches that organizations dealing with sensitive data can take up to safeguard their customers from any potential privacy breach. The report only deals with *static, structured, well-defined, single-level* datasets, and ways to anonymise data in such datasets. This somewhat limits the scope of the report, as today data is available in a large number of formats, each with their own associated privacy risks, but nonetheless it is one of the most widespread form of data.

The report introduces the terminology associated with privacy fairly well, and has defined the scope and target audience comprehensively. Further, it discusses in detail some data anonymisation concepts, such as the purpose and utility of anonymisation, characteristics of anonymisation techniques, inference, audience limitations, and the like. These concepts are well defined. Then, it discusses some data disclosure risks, defining the different ways an adversary can obtain information from a database if it is not sufficiently anonymised.

The report then discusses some data anonymisation techniques in detail. The following section discusses each, along with its strengths and weaknesses.

2 Data Anonymisation Techniques

2.1 Attribute Suppression

This technique involves removing complete attributes, or columns of data from the database. Such attributes are generally either not required in the anonymised dataset, or when the attribute cannot be sufficiently anonymised using any other technique. Examples include attributes which are direct identifiers, such as names and addresses.

[1] https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

The report states that this is "the strongest type of anonymisation technique". This is however dubious, as it has been shown that removing just identifying information does not anonymise the dataset. There are other attributes, called *quasi-identifiers*[1], or attributes which do not directly leak information, but a combination of a few of those narrows the search down to a very small number of people. Research [2] has shown that *linkage attacks* can be performed on these quasi-identifiers using multiple publicly available datasets to break anonymity.

2.2 Record Suppression

This involves removal of entire records from the dataset. This is generally done over outliers that are obtained after using other generalization techniques over the datasets.

While this technique surely does protect the privacy of the users whose data has been removed (as their data no longer exists in the dataset — there is no longer a problem to solve for them), this can significantly impact the dataset statistics, affecting the integrity of the dataset as a whole. This has been rightfully highlighted in the report.

2.3 Character Masking

This technique involves partially changing the data values by using a constant symbol to hide a sequence of characters (For example, zip code 123456 masked to 12XXXX).

This is one of the techniques that can be used to achieve *k-anonymity*[3, 4] in a dataset, as it can be a good binning approach for cases where a large number of sequences are masked (for example 123456, 124589, 126643 all masked to 12XXXX). But it has been shown that k-anonymity is not enough for completely anonymising datasets, especially where all the k-anonymised records end up being identical in the sensitive attributes as well. This makes the dataset not sufficiently anonymised and vulnerable to *homogeneity attacks*[5].

Further, real-life datasets are generally very sparse, and k-anonymity requires each record to have at least k close neighbours. Nearest neighbours are very far from each other. This is called the curse of dimensionality[6]. Projection of the dataset to low dimensions to make k-anonymity possible loses so much information to render the dataset useless. So while this approach may be partially effective, it is not a fool-proof technique.

The report further highlights an important distinction point. Scenarios where data is masked in such a way that the person to whom the data belongs can recognize it, is not part of data anonymisation techniques. For example, partially masked credit card numbers can still be uniquely recognized by the owners combined with possibly just one other attribute. The objective of data anonymisation is that even the owners of the data record cannot identify it themselves uniquely.

2.4 Pseudonymisation

This technique involves replacing identifying attributes of data with pseudonyms, or made-up values, a map of which with the true values is available only with the original owner of the dataset. This approach allows for referential integrity of datasets, in contrast to attribute suppression.

For a standalone release of a dataset, this is roughly equivalent to attribute suppression, as these attributes no longer contribute any value to the dataset. This is equivalent to dropping the attributes as a whole. There is still a threat of linkage attacks using quasi-identifiers, and as discussed above, such techniques are not very effective at preserving privacy as a whole.

2.5 Generalization

This technique is similar to character masking, albeit for all types of data attributes. It involves deliberately reducing the precision of the data in the record. This may involve processes like converting age into an age range, or reducing the precision of the location.

This technique is one of the operations used for the general k-anonymisation approach, and suffers all the flaws that k-anonymity suffers from. The dataset obtained after applying this approach may not be sufficiently anonymised and can be de-anonymised.

2.6 Swapping

This technique involves swapping the values of some between records, to prevent linkage attacks being successful. On the whole, this ensures that the dataset statistics remain consistent when considering the dataset as a whole. However, this technique significantly compromises the integrity of the dataset, as it may now contain completely false data on an individual record level. The correlations between attributes may be significantly altered.

2.7 Data Perturbation

This technique involves changing the values in the original dataset by adding a small random noise, so that the final values are slightly different. It introduces some degree of "error" into the data, making it harder to identify individuals in the dataset. This only works on numerical attributes.

The success or failure of this technique depends heavily on the degree of noise that is added. If the change is too small, it is ineffective, as simple rounding operations will negate the error introduced. If the change is too large, the dataset becomes useless. The correct balance may be hard to find. Further, probabilistic attacks using summary statistics are still possible, as the error will not greatly impact the statistics of the dataset.

One technique that offers strong guarantees for privacy that works by adding noise proportional to a 'privacy budget' is *Differential Privacy*[7]. While the report mentions the concept in passing, this technique must be used to protect user privacy.

2.8 Synthetic Data

This is a special case where synthetic data is generated with patterns similar to the original dataset, when the data is to be used for research purposes. This is relatively effective, as the original data is not used and it may be impossible to concretely identify individual users' sensitive attributes. However, local statistics based attacks[8] to find out the sensitive attributes of a person probabilistically is still possible, as the synthetic dataset mimics the patterns of the original dataset. Further, such datasets may only be useful in very specific scenarios.

2.9 Data Aggregation

This technique involves aggregating information, to get a summary of the data in the record instead of the actual values. For example, the entry and exit time of people into a building can be combined into an attribute describing the total time spent in the building.

This approach however works again only in a limited set of scenarios on numerical and temporal data attributes. Further, probabilistic attacks[8] are still possible, as such attacks tend to exploit such summary statistics, and the approach is generating, not masking such attributes.

3 Conclusion

The report discusses several approaches to anonymise data in structured datasets. The approaches revolve around the concept of k-anonymity. The report further discusses comprehensively in detail k-anonymity and how is it used over datasets, using the anonymisation techniques discussed above.

However it has been shown in research that k-anonymity provides a very weak privacy guarantee. In fact, the report does not even discuss in detail the improved versions of k-anonymity, i.e. *l-diversity* and *t-closeness*, which provide significantly greater privacy guarantee. However, such techniques are still not good enough. The report does highlight the fact that these techniques are not actually anonymisation techniques, but are a measure of risk. The report fails to discuss in detail concepts such as *differential privacy*, which provide a way better privacy guarantee.

The report further discusses some general safeguards and restrictions on data access that the organizations can put in place to prevent leaks of private information. These are good suggestions, and must be enforced in organizations handling extremely sensitive data, but it is not always possible to enforce such restrictions. In case the dataset has to be released publicly, these safeguards are rendered irrelevant. The report also touches on the aspect of IT governance, and the importance of anonymising data before it is released, and keeping track of the anonymised instances to ensure that different anonymisation approaches over different instances of dataset releases cannot be combined to adversarially obtain sensitive information.

References

[1] Tore Dalenius. Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.

[2] Latanya Sweeney. Simple demographics often identify people uniquely. *Data Privacy Working Paper, Carnegie Mellon University*, 2000.

[3] Latanya Sweeney. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.

[4] Latanya Sweeney. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):571–588, October 2002.

[5] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–es, March 2007.

[6] Charu C. Aggarwal. On K-Anonymity and the Curse of Dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, page 901–909. VLDB Endowment, 2005.

[7] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[8] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application (2017)*, 2017.