

---

# BALANCING PRIVACY AND INFORMATION UTILITY IN DATA ANONYMISATION

*Marek P. Zielinski (University of Pretoria / SAP Research CEC Pretoria)*

---

**Abstract:** In order to protect the privacy of individuals in sensitive data that has been released for analysis, it is not sufficient to de-identify the data by removing explicit identifiers, such as names and addresses. The data must also be anonymised, so as to prevent compromising the privacy of individuals through the manipulation of the data and / or by matching it with other sources of data. Anonymised data must satisfy two conflicting goals: privacy and information utility (e.g. the precision and completeness of the data). When data is anonymised, ideally both privacy and information utility levels should be maximised. However, it is difficult to determine the optimal levels of privacy and information utility when anonymising data, since the higher the required level of privacy, the lower the utility value of the anonymised data. To date, we have not found a model that has been proposed in the literature that could be used to guide the selection of an optimal balance between privacy and information utility. This paper describes current research work undertaken to address this challenge; a data anonymisation model will be developed during the course of the research work to guide the process of microdata anonymisation.

**Keywords:** security, privacy, microdata, information utility, association rules

## 1. Introduction

Statistical data often serves as the basis for creating knowledge that is used by different organisations to assist in their planning and decision-making activities. When confidential statistical data is released for use, it needs to satisfy different requirements with regards to protection of privacy of the individuals whose data is released. Furthermore, the released data should be such that it is useful and usable to the users of the data.

It is clear that a conflict arises in this case since protection of privacy implies hiding and obscuring data, while making data usable and useful implies providing data that is accurate, complete and precise. Ideally, both privacy and information utility levels should be maximised. However, the higher the required level of privacy, the lower the utility value of the released data. Without guidelines to guide the selection of optimal levels of privacy and information utility (taking into account the purpose for which the released data will be used and in which type of environment), it is difficult to find a good balance between the two goals.

In this paper, we describe current research work undertaken to address the challenge of anonymising microdata such that it possesses an optimal balance between privacy and information utility. This paper is organized as follows: In Section 2, we describe the main research aim of the undertaken work by discussing the need to guide the anonymisation of microdata to balance the needs of privacy and information utility. To address this need, we propose a model to guide the anonymisation of microdata. We discuss this model in Section 3 and conclude the paper in Section 4.

## 2. Guiding the anonymisation of microdata

Statistical data can be released either in the form of tables or as microdata files. Data released in the form of tables is usually summarised and aggregated (e.g. in the form of averages, means, frequency counts, etc.). Microdata files, on the other hand, contain records of non-aggregated data and each record contains a set of attribute values that are associated with a single individual or other entity [9].

If the data is released in the form of microdata rather than in the form of tables, it may be of greater value to users of the data since the users are offered greater flexibility with regards to what the data can be used

for. For example, users can compute specific statistics that are of interest to them, rather than being restricted to using pre-computed aggregated statistics.

Due to the need to protect individuals' privacy, confidential statistical data should be anonymised in addition to being de-identified. The process of microdata de-identification, which involves the removal of explicit identifiers (e.g. names and addresses) does not necessarily produce anonymous microdata. That is, the microdata can be manipulated or matched with other sources of data to compromise the privacy of an individual. This can occur through different forms of disclosure [2]: identity disclosure, which occurs when the identity of an individual is revealed from the released microdata; attribute disclosure, which occurs when sensitive information about an individual is obtained from the released microdata; and inferential disclosure, which occurs when the value of a particular characteristic of an individual can be determined, from the microdata, more accurately than it would otherwise have been possible.

The concept of  $k$ -anonymity has been proposed by Samarati and Sweeney [6, 7, 8] to anonymise microdata through the use of data generalization and suppression. In order for a microdata set to satisfy the requirement of  $k$ -anonymity, every record in the microdata must be related to at least  $k$  other records. Different algorithms [1, 3, 6, 7] may be used to anonymise microdata such that it satisfies the requirement of  $k$ -anonymity.

The privacy level of anonymised microdata is affected by the value chosen for  $k$ . The higher the value for  $k$ , the greater the privacy level of the anonymised microdata. However, choosing a high value for  $k$  requires that data is anonymised to a high degree, which reduces the accuracy and completeness of the data. Therefore, the higher the value chosen for  $k$ , the lower the information utility level of the anonymised microdata.

Although  $k$ -anonymity is able to protect against identity disclosure, it does not prevent attribute disclosure. This limitation has been addressed in [5] by introducing the concept of  $l$ -diversity, which requires that a microdata set is such that, for every block of tuples that have the same value for the quasi-identifier, there are at least  $l$  "well-represented" values for the sensitive attribute. However,  $l$ -diversity is also insufficient to prevent attribute disclosure, as has been discussed in [4] where the authors present the concept of  $t$ -closeness to address this limitation.

Before microdata is anonymised, the required levels of privacy and information utility should be determined. Ideally, the microdata should be anonymised such that it will provide a level of privacy that is sufficient to prevent disclosure, while also providing an acceptable level of information utility. However, since privacy and information utility are two conflicting requirements, it is difficult to determine what should be the optimal value for  $k$  when microdata is anonymised for different uses such that it satisfies  $k$ -anonymity. Similarly, it is also difficult to determine the optimal values for  $l$  when satisfying  $l$ -diversity, or the value for  $t$  when satisfying  $t$ -closeness.

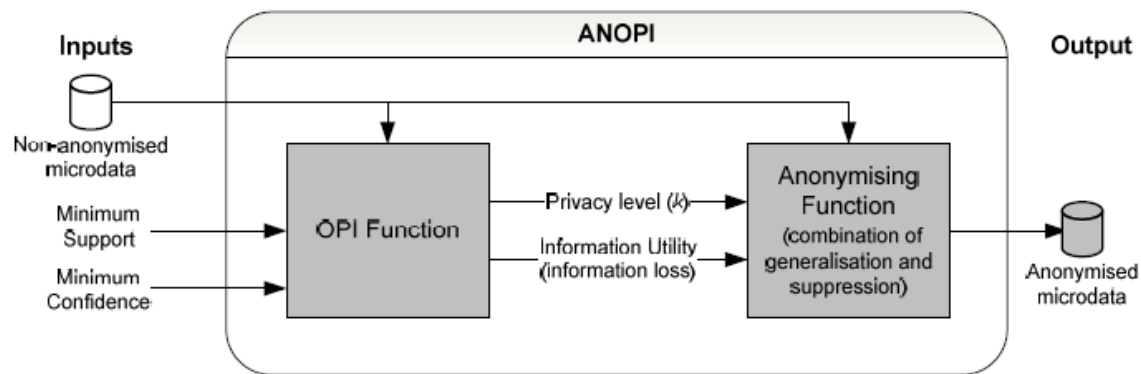
To date, we have not found a method proposed in the literature that could be used for the purpose of guiding the selection of an optimal balance between privacy and information utility when anonymising microdata for different uses, and which would be sufficiently formal and yet simple enough to use. Without such guidelines, it is difficult to select the optimal levels of privacy and information utility when microdata is anonymised for different uses in different environments.

The study discussed in this paper aims to address the above problem by answering the following research question: "How can the processes of microdata anonymisation be guided such that there will exist an optimal balance between privacy and information utility in the anonymised microdata?"

The research question will be addressed by developing an anonymisation model. This model, called the ANOPI Model, or the ANonymisation with Optimal Privacy and Information utility Model, will ensure that the anonymised microdata possesses an optimal balance between privacy and information utility, based on the content of the microdata, the purpose for which the microdata will be used, as well as the environment in which the microdata will be used. The model will be the main result of this study and it is briefly discussed in the next section.

### 3. The ANOPI model

The ANOPI Model, or the ANonymisation with Optimal Privacy and Information utility Model, will be developed during the course of this research. The purpose of this model is to guide the anonymisation of microdata such that the microdata will possess an optimal balance between privacy and information utility. The level of privacy and information utility will be based on the content of the microdata, the purpose for which the microdata will be used, as well as the environment in which the microdata will be used. The model anonymises microdata when it is given, as inputs, a non-anonymised microdata set as well as the minimum support and confidence values of sensitive association rules that can be mined from the microdata. During the course of this research, the ANOPI model will be implemented in a prototype data anonymisation system. The model is shown in Figure 1.



**Figure 1.** The ANOPI model, shown with its inputs, its two functions and its output

The ANOPI model is composed of two functions: the first function determines the optimal balance between privacy and information utility, and the second function anonymises microdata such that it possesses this optimal balance.

The OPI function, or the **O**ptimal **P**rivacy and **I**nformation utility function is used to determine the optimal balance between privacy and information utility. The function produces two values as outputs: privacy and information utility. The value for privacy is represented in terms of the value for  $k$  used for subsequent  $k$ -anonymisation, while the value for information utility is represented in terms of information loss. The operation of the OPI function is based on the fact that different association relationships exist among attribute values and individual records in the microdata. These relationships are represented as association rules that have certain support and confidence values.

The OPI function works as follows. When a non-anonymised microdata set is given as input, all sensitive association rules, that can be mined from the non-anonymised microdata set, are found. The association rules are known as *sensitive* association rules because they can be used to disclose confidential information. The association rules must have certain minimum support and confidence values, which are provided as inputs to the function. Based on these rules and their support and confidence values, the optimal values for  $k$  (privacy level) and information loss (information utility) are determined and provided as outputs of the function. These values for  $k$  and information loss, together with the non-anonymised microdata, are then used as inputs for the anonymising function.

The anonymising function is the second function of the model and its purpose is to perform the actual anonymisation of the microdata. To anonymise the microdata, a combination of generalisation and suppression is applied on the non-anonymised microdata such that it satisfies  $k$ -anonymity and possesses the identified level of information utility.

## 4. Conclusion

In this paper, we presented current research work that has been undertaken to address the challenge of finding an optimal balance between privacy and information utility when anonymising microdata. The aim of this research is to develop a model that can be used to guide the process of anonymising microdata. To date, we have not found a model that has been proposed in the literature and which could be used to guide the selection of an optimal balance between privacy and information utility. Therefore, it can be quite difficult to determine the optimal way in which to release microdata such that it possesses an optimal balance between privacy and information utility. We believe that, by undertaking this study, the process of releasing microdata (with respect to resulting levels of privacy and information utility) can be significantly improved through the use of a formal model to guide the anonymisation of microdata.

## Acknowledgement

The study described in this paper is part of a larger research project lead by the author at SAP Research CEC in Pretoria, South Africa. The support of SAP Research towards this work is hereby acknowledged.

## References

1. Bayardo, R. J.; Agrawal, R. (2005). *Data Privacy Through Optimal K-Anonymisation*, Proceedings of the 21st International Conference on Data Engineering.
2. Duncan, G. T.; Fienberg, S.E.; Krishnan, R.; Padman, R.; Roehrig, S.F. (2001). *Disclosure Limitation Methods and Information Loss for Tabular Data*. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135-166.
3. Le Fevre, K.; De Witt, D. J.; Ramakrishnan, R. (2005). *Incognito: Efficient Full-Domain k-Anonymity*. Proceedings of the 2005 Sigmod Conference, Baltimore, USA, pp. 49-60.
4. Li, N.; Li, T.; Venkatasubramanian, S. (2007). *t-Closeness: Privacy Beyond k-Anonymity and L-Diversity*. Proceedings of the 23rd IEEE International Conference on Data Engineering, Istanbul, Turkey.
5. Machanavajjhala, J. Gehrke; Kifer D. (2006). *l-Diversity: Privacy Beyond k-Anonymity*. In: Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta, USA, 2006.
6. Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, pp. 1010-1027.
7. Sweeney, L. (2002a). Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 571-588.
8. Sweeney, L. (2002b). k-Anonymity: a Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 557-570.
9. Willenborg, L.; De Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, Springer-Verlag, New York, USA.

## Author

Marek Piotr Zielinski  
SAP Research CEC Pretoria

Unit 3 ProPark Building; 29 De Havilland Crescent  
0020 Perseus Park  
South Africa  
[marek.zielinski@sap.com](mailto:marek.zielinski@sap.com)