HOW APPROPRIATE IS *K*-ANONYMITY FOR ADDRESSING THE CONFLICT BETWEEN PRIVACY AND INFORMATION UTILITY IN MICRODATA ANONYMISATION

Marek P. Zielinski [1], Martin S. Olivier [2]

University of Pretoria, South Africa

1: marek.zielinski@sap.com

2: molivier@cs.up.ac.za

ABSTRACT

Before statistical data, such as microdata, can be released to the public, it needs to be anonymised. Anonymisation protects the privacy of the individuals whose data is released. However, as microdata is anonymised, its level of privacy increases, while its level of information utility decreases.

K-anonymity is often used to address the conflict between privacy and information utility in microdata anonymisation. In this paper, we determine the extent to which *k*-anonymity is appropriate for addressing this conflict. We argue that the way in which *k*-anonymity is currently used to address this conflict does not necessarily lead to an optimum balance between privacy and information utility. We also provide recommendations for an appropriate solution for addressing the conflict between privacy and information utility.

KEY WORDS

Privacy, information utility, *k*-anonymity, microdata

HOW APPROPRIATE IS K-ANONYMITY FOR ADDRESSING THE CONFLICT BETWEEN PRIVACY AND INFORMATION UTILITY IN MICRODATA ANONYMISATION

1 INTRODUCTION

Microdata is one way in which statistical data can be released to the public. However, before it can be released to the public, it needs to be anonymised. Anonymisation ensures the privacy of the individuals whose data is released. As microdata is anonymised, data is removed (to some extent) from the identifying variables. As more data is removed from the identifying variables, it becomes increasingly difficult to infer sensitive data and to perform re-identification. Therefore, as microdata is anonymised, the level of privacy in the microdata increases. However, removing data from the identifying variables also reduces the accuracy and / or the completeness of the released microdata. Therefore, as microdata is anonymised, its level of information utility also decreases.

Ideally, we would like to release microdata that has high levels of privacy and information utility. However, the protection of privacy implies that we should hide and obscure data. On the other hand, releasing usable and useful data implies that we should provide data that is accurate, complete and precise (Zielinski, 2007a, 2007b). Clearly, a conflict between the needs of privacy and information utility exists. This conflict needs to be resolved before a microdata set can be released to the public.

K-anonymity is often used to address the conflict between privacy and information utility in microdata anonymisation. In this paper, we determine the extent to which *k*-anonymity is appropriate for addressing this conflict. We argue that the way in which *k*-anonymity is currently used to address this conflict does not necessarily lead to an optimum balance between privacy and information utility.

This paper is organised as follows. In Section 2, we provide preliminary definitions of microdata, *k*-anonymity, and the "optimum" balance between privacy and information utility. In Section 3, we discuss the appropriateness of the current way in which *k*-anonymity is used for addressing the conflict between privacy and information utility. In Section 4, we provide specific examples of how *k*-anonymity is used to address this conflict. In Section 5, we provide recommendations for a solution that will appropriately address the conflict between privacy and information utility. We discuss related work in Section 6 and conclude the paper in Section 7.

2 PRELIMINARIES

The focus of this paper is on determining the extent to which *k*-anonymity is appropriate for addressing the conflict between privacy and information utility in microdata anonymisation. Therefore, we will define the concepts used, namely microdata, *k*-anonymity, and the "optimum" balance between privacy and information utility. However, we first provide definitions for the *data owner* and the *data user*. We define the *data owner* as a person or an organization that releases microdata about individuals. For example, a data owner may be a hospital that releases microdata that contains information on its patients. We also define the *data user* as a person or an organization that requires the released microdata in order to perform specific types of data analysis.

2.1 Microdata

Statistical data can be disseminated in three main ways (Hundepool et al., 2007; Domingo-Ferrer, Sebe, & Solanas, 2008; Willenborg & De Waal, 2001). These include Dynamically Queryable Databases, Tabular Data, and Microdata.

Microdata is the most basic form in which statistical data can be released. It is the "raw" data from which all other statistical data outputs are derived. A microdata set may be represented as a single data matrix, where the rows correspond to the entities of the database (e.g. an individual person or a respondent) and the columns correspond to the variables of an each entity.

In the existing literature, different names for the different categories of variables of a microdata set are used by different authors. In this paper, we shall adapt the naming conventions used by Willenborg and De Waal (2001). However, for completeness of this discussion, we also provide the alternative names used by other authors.

There are four, not necessarily disjoint, categories into which the variables of a microdata set can be classified. Before the microdata is anonymised, the data owner should determine the category of each variable.

- *Direct identifiers.* These variables are those that uniquely identify a respondent in a microdata set. A person's Passport Number, or ID Number are examples of a direct identifier. Direct identifiers are sometimes simply referred to as *Identifiers* (Ciriani et al., 2007; Hundepool et al., 2007). Before microdata is anonymised, direct identifiers are removed from the microdata set.

- *Indirect identifiers.* These variables are not necessarily unique for each respondent. However, the combination of the values of one or more indirect identifiers of a single record may create a relatively rare, or even a unique combination. Indirect identifiers are those variables on which an intruder will try to re-identify an individual respondent in a microdata set. Examples include the Date of Birth, Marital Status, or Zipcode of a person. Indirect identifiers are also sometimes referred to as *quasi-identifiers* (Samarati, 2001), or *key variables* (Hundepool et al., 2007). However, throughout this paper, we shall refer to an indirect identifier as an *identifying variable,* as has been done by Willenborg and De Waal (2001).

- *Sensitive variables.* These variables are those that contain sensitive information of a respondent. For example, a sensitive variable can be a person's disease that he sought treatment for in a hospital. These variables are also referred to as *confidential outcome variables* (Domingo-Ferrer et al., 2008; Hundepool et al., 2007), since they contain confidential information about the respondents.

- *Non-sensitive, non-identifying variables.* These variables are those that do not fall into any of the above categories. These are also referred to as *non-confidential outcome variables* (Domingo-Ferrer et al., 2008; Hundepool et al., 2007). An example of a non-sensitive, non-identifying variable may be a person's gender.

However, in combination with other variables, such as the marital status, a person's gender could also be an indirect identifier. Therefore, we mentioned earlier that the four variable categories are not necessarily disjoint.

2.2 *K*-anonymity

The concept of *k*-anonymity was introduced by Samarati and Sweeney (Samarati, 2001; Sweeney, 2002a, 2002b) for anonymising microdata. A microdata set satisfies the property of *k*-anonymity if every record in the microdata set is indistinguishable from at least *k* - 1 other records in the same microdata set, where *k* is greater than 1. The inability to distinguish between different records is based on the values of the identifying variables (or quasi-identifiers - an equivalent term commonly used in the literature on *k*-anonymity). That is, given a record with a particular set of values for the identifying variables, the same set of values will be present in the identifying variables of at least of *k* - 1 other records in the same microdata set.

2.3 The "optimum" balance between privacy and information utility

In our research work, we regard the optimum balance between privacy and information utility as has been defined by Zielinski and Olivier (2009a). That is, the optimum balance between privacy and information utility is one in which the levels of privacy and information utility are maximised while satisfying a set of constraints that capture the data owner's and the data user's preferences. These preferences refer to the preferences that exist between each identifying variable in the microdata set, as well as the preference between the resulting levels of privacy and information utility.

The preferences between each identifying variable in the microdata set are directly related to the usefulness of the data. The usefulness of the data should be considered from both the data user's and the data owner's points of view. In the case of the data user (whose main goal is to ensure utility of data), the preferences for identifying variables should reflect the extent to which each identifying variable will be useful for the user's tasks. In the case of the data owner (whose main goal is to protect the privacy of the respondents in the microdata), the preferences are considered from a potential intruder's point of view, in terms of the perceived way in which an intruder may use the released data to infer sensitive information. In this case, the preferences for identifying variables should reflect the extent to which we perceive that each identifying variable will be useful for the intruder in inferring sensitive data.

The preference between the resulting levels of privacy and information utility must be decided and agreed upon by the data user and the data owner. That is, it is necessary to determine if protection of privacy is considered to be equally important as providing useful data, or if privacy should assume a greater or lower importance compared to information utility. For example, if the microdata is released to only a selected group of data users, under strict confidentiality agreements made with this group, then it is certainly possible that the data owner's preference for privacy may be lower in comparison to cases where the microdata is made available to the public.

Therefore, we state our optimisation problem as follows: "Maximise privacy and information utility subject to the constraints imposed by the data user's and the data owner's preferences". In the next Section, we discuss the extent to which *k*-anonymity is appropriate in finding the optimum balance between privacy and information utility.

3 HOW APPROPRIATE IS *K*-ANONYMITY FOR ADDRESSING THE CONFLICT BETWEEN PRIVACY AND INFORMATION UTILITY

The use of *k*-anonymity is seen as a "clean way" of addressing the conflict between privacy and information utility (Domingo-Ferrer & Torra, 2005, 2008). It is seen as a "clean way" because, it is assumed that, if for a given *k* value, *k*-anonymity will provide sufficient privacy, then it allows one to concentrate on only determining how to minimise information loss (or maximise information utility) such that the given level of *k*-anonymity will be achieved. However, we argue that if this is assumed and if *k*-anonymity is used in this fashion, then it does not fully capture the objective of the optimisation problem.

First of all, it is unclear (from the literature stemming from *k*-anonymity) how to determine the optimum value for *k* that will provide "sufficient privacy" for the particular set of circumstances in which anonymisation takes place. Before we can find the optimum value for *k*, we need to know what the optimum balance between privacy and information utility is for the given set of circumstances in which anonymisation takes place. Moreover, under the above assumption, when a certain *k* value is provided as input to anonymisation, it is provided without knowing if the given value will in fact lead to an optimal balance between privacy and information utility.

Under the above assumption, the complexity of the optimisation problem is reduced to only maximising information utility when given a certain level of privacy that needs to be achieved (i.e. a *k* value for *k*-anonymity). However, we believe that such an assumption does not take into account the whole complexity of the optimisation problem (as stated in Section 2.3). That is, such an approach does not take into account that it is *both* privacy and information utility that have to be maximised in the optimisation problem.

When the above assumption is used to solve this optimisation problem, maximising privacy is *no longer an objective function of the optimisation problem*. Instead, under the above assumption, privacy is reduced to *only a constraint under which optimisation occurs*. When privacy becomes just a constraint under which optimisation occurs, then the optimisation does not necessarily lead to a truly optimum solution. Information utility is optimised only to satisfy a given level of privacy, rather than being optimised whilst being aware of the fact that the goal of maximising information utility is in direct conflict to the goal of maximising privacy. In other words, information utility is optimised subject to a given level of privacy that is considered "sufficient".

Nevertheless, the given "sufficient" level of privacy may not necessarily be the optimum level, since the privacy level was decided upon through a means other than during the optimisation itself. This is not to say that, the optimum level of privacy will occur below the required "sufficient" or minimum level. It cannot occur below the minimum level, since otherwise the constraint of the minimum level of privacy would not be met. It is, however, possible that the *optimum* level of privacy will occur above the required minimum privacy level, but this will not be known unless privacy is optimised as well.

Note that we are not discrediting the usefulness of *k*-anonymisation for anonymising microdata. We are, however, stating that when *k*-

anonymisation is used to find the optimum balance between privacy and information utility, then the optimisation problem should be approached from both angles: the need to maximise both information utility and privacy. If this problem is approached from both these angles, then during the process of optimisation, the *k* value will actually be *calculated*. First, the optimum balance will be determined. Thereafter, in a second step, the optimum balance will be used to determine how the microdata should be anonymised. If *k*-anonymity is used as the anonymisation technique, then during the second step, the value for *k* will be calculated and then the microdata set will be *k*-anonymised with this value. In other words, the value for *k* will no longer be an input into the optimisation problem. The only input into the optimisation problem will be the constraints under which the optimisation should occur. These constraints are the preferences that were stated in Section 2.3.

The limitation of the way in which *k*-anonymity is used to address the conflict between privacy and information utility, as discussed above, relates to the objectives of the optimisation problem. Another limitation of *k*-anonymity, with regards to how it is currently used to address the conflict between privacy and information utility, is related to the definition of the constraints under which optimisation is performed.

In the original definition of *k*-anonymity, anonymisation is performed without taking into account the data user's preferences between the different identifying variables. Therefore, the anonymisation does not consider that information loss should be minimised in those identifying variables that a data user considers useful. Some enhancements of *k*-anonymity have addressed this shortcoming, as discussed in the next Section. In a similar way, the original definition of *k*-anonymity also disregards the (perceived) preferences between identifying variables that a potential intruder may have. That is, anonymisation does not necessarily ensure that the most information loss occurs in those identifying variables that we perceive to be most useful for a potential intruder. Furthermore, *k*-anonymity also does not take into account the preference between privacy and information utility. When we need to determine the optimum balance between privacy and information utility, these preferences should be taken into account as constraints under which the optimisation is performed. However, the original *k*-anonymity definition does not take these into account.

To summarise, although *k*-anonymity shows potential as a good way to address the conflict between privacy and information utility, we argue that the way in which it is currently used is not appropriate to address this conflict. That is, the way in which *k*-anonymity is currently used fails to find a truly optimum balance between privacy and information utility for two main reasons. The first reason relates to the way in which the objective of the optimisation is defined. That is, the objective of the optimisation problem focuses on only maximising information utility, such that a certain level of privacy (*k* value) is met. To find the optimum balance between privacy and information utility, the objective of the optimisation should focus on maximising both privacy and information utility. The second reason relates to the way in which the constraints of the optimisation are defined. That is, the preferences between privacy and information utility, as well as the data user's preferences and the data owner's preferences (in terms of the perceived intruder's preferences) between identifying variables are not taken into account when optimisation is performed.

4 SPECIFIC EXAMPLES OF HOW K-ANONYMITY IS USED TO ADDRESS THE CONFLICT BETWEEN PRIVACY AND INFORMATION UTILITY

In this Section, we present a number of specific examples of how *k*-anonymity has been recently used to address the conflict between privacy and information utility.

Stark, Eder and Zatloukal (2006) propose a priority-driven anonymisation technique to achieve *k*-anonymity. The proposed technique allows specifying the degree of acceptable information loss for each variable seperately. Variables that are considered useful for the data user can be protected from extensive generalization. Those variables that have been assigned low priorities are generalized first. Variables that have been assigned higher priorities are only generalized when no other solution may be found to achieve *k*-anonymity. Although this approach is able to take into account the user's preferences with respect to which variables will be useful to him, it is unable to take into account other constraints of the optimisation problem, namely the data owner's preferences between variables (from the perspective of a potential intruder) and also the preferences between privacy and information utility. Moreover, the optimisation problem is addressed by considering only the need to maximise information utility such that a certain level of *k*-anonymity is provided.

Other utility-based anonymisation approaches were also proposed. For example, LeFevre, DeWitt, and Ramakrishnan (2006) propose algorithms that will generate anonymous data such that the utility of the data is preserved with respect to the workload for which the data will be used. Xu et al. (2006) also study the problem of utility-based anonymisation and present a framework to specify the utility of variables. Zhang, Jajodia and Brodsky (2007) propose a model and an algorithm that will guarantee safety under the assumption that the intruder knows the disclosure algorithm and the generalization sequence. Nevertheless, these works address the conflict between privacy and information utility from only one angle, namely the need to maximise information utility subject to a given *k* value (i.e. a level of privacy that is considered as "sufficient"). As we argued in the previous Section, considering the optimisation problem from this limited perspective does not lead to a truly optimum balance between privacy and information utility

In a more recent work, Gionis and Tassa (2009), study how to achieve *k*-anonymity with minimal loss of information (i.e. an optimum *k*-anonymisation). The authors provide an improvement on the best-known *O*(*k*)-approximation provided by Aggarwal et al. (2005) to an approximation of *O*(ln *k*). Nevertheless, the authors also do not consider the optimisation problem from the perspective of maximising both privacy and information utility. Instead, they aim to determine how to achieve *k*-anonymity with such that information utility is maximised. That is, the algorithm proposed expects that the value for *k* will be provided as input. However, as we argued in the previous Section, if we are to obtain a truly optimum balance between privacy and information utility, by using *k*-anonymisation as the anonymisation technique, then the value for *k* will actually be calculated during the optimisation process.

Loukides and Shao (2008) consider how a *k*-anonymisation can be produced with an optimum trade-off between information utility and privacy. In their paper, the needs of both privacy and information utility are considered. The optimisation problem is addressed from both these angles when an optimal anonymisation is determined. However, the proposed measure for information utility is based on the average amount of generalizations that each group of records incurs - the smaller this number, the higher the utility. This proposed measure does not consider the preferences that a specific data user may have between different identifying variables. Therefore, this measure will not be able to take into account the purpose for which the user requires the data and hence does not provide a meaningful measure for information utility. Therefore, an anonymised microdata set will not necessarily have the optimal level of information utility for a specific user and the purpose for which the data is released.

Although a number of approaches based on *k*-anonymity have been proposed to address the conflict between privacy and information utility, none are able to find a truly optimum balance between and information utility. The concept of *k*-anonymity itself is also currently being used inappropriately to address this conflict. In the next Section, we present recommendations for an appropriate solution that will ensure that the optimum balance between privacy and information utility is achieved when microdata is anonymised.

5 RECOMMENDATIONS FOR AN APPROPRIATE SOLUTION

When we consider the definition of the "optimum" balance between privacy and information utility provided in Section 2.3, it is clear that the way in which *k*-anonymity is currently used to address the conflict between privacy and information utility is not appropriate. We now provide recommendations for developing a solution that will be appropriate for determining the optimum balance between privacy and information utility.

We argue that if we are to find a truly optimal balance between privacy and information utility, then the goal of maximising *both* privacy and information utility should be regarded as *the objective function of the optimisation problem*. This stems from the fact that both privacy and information utility are desired, although they may be desired in different proportions. This is our recommendation with respect to the objective of the optimisation problem.

We also need to make recommendations that address the constraints under which optimisation should be carried out. These constraints should reflect the preferences between privacy and information utility. The constraints should also reflect the data user's and the data owner's preferences between identifying variables. In the case of the data owner, the preferences between identifying variables should be considered from the perspective of the potential intruder (i.e. what identifying variables are considered to be most useful for an intruder in deriving sensitive data).

Therefore, a challenge exists to develop a solution that will appropriately capture the above objective and constraints and thereafter find the optimum balance between privacy and information utility. Moreover, once the optimum balance has been determined, the solution should also determine how to anonymise the microdata such that the optimum levels are achieved. Therefore, the solution should have two components: an optimisation component, in which the optimum levels of privacy and information utility are determined, and an anonymisation component, during which the microdata is anonymised.

In cases where *k*-anonymity is used as the anonymisation technique, the optimisation component of the solution will determine the optimum level of privacy and information utility. Thereafter, the anonymisation component of the solution will calculate the optimum value for *k* with which the microdata set should be *k*-anonymised.

6 RELATED WORK

Other approaches for addressing the conflict between privacy and information utility have also been proposed. In addition to *k*-anonymity, Domingo-Ferrer and Torra (2005) identify two other approaches. These include the *score*, and R-U confidentiality maps.

Domingo-Ferrer and Torra (2001) introduced the *score* as a way to evaluate the trade-off between information loss and disclosure risk. It was subsequently used in several other works, for example, by Medrano-Gracia et al. (2007), Nin, Herranz, and Torra (2008a, 2008b), Yancey, Winkler, and Creecy (2002). The *score* is useful in that it allows us to regard the selection of a masking technique (for microdata protection) and the parameters of the technique as an optimisation problem (Domingo-Ferrer & Torra, 2005). For example, Sebe et al. (2002) applied a masking technique to a microdata set, after which a post-masking optimisation procedure was applied to obtain an improved *score*. The main drawback of the *score*, with reference to how appropriate it is in addressing the conflict between privacy and information utility, is that it is unable to take into account the way in which the released data will be used, or the way in which we perceive the intruder to infer sensitive data. The need to take into account these preferences was one of the requirements we identified for the "optimum" balance between privacy and information utility. The *score* fails to take into account this requirement, and hence we do not consider it appropriate for finding the optimum balance between privacy and information utility.

R-U confidentiality maps (Duncan et al., 2001; Duncan, Keller-McNulty & Stokes, 2001) provide a way in which to graphically represent the conflict between disclosure risk, R, and data utility, U. After the form of the disclosure risk, R, and the data utility, U, have been specified, the task is to determine how R and U are related to the parameter values of the specific masking technique chosen to anonymise the microdata set. An R-U confidentiality map is obtained by plotting, on a two-dimensional graph, a set of paired values, (R, U), which represent the disclosure risk and the data utility that correspond to various strategies for data release.

The graphical representation of the relationship between privacy and information utility allows one to easily determine how a particular masking technique, and its parameters, impacts the balance between privacy and information utility. It is, of course, reasonable to expect that the microdata set should be released with a level of data utility U at which the disclosure risk R will be below the maximum tolerable risk. However, by using the R-U confidentiality map alone, it is still unclear where the optimum balance between R and U occurs. One does not know if the optimum balance occurs *exactly* at the point at which R is just below the maximum tolerable risk. However, it is also quite likely that the optimum balance may, in fact, occur at a lower risk level, much lower than the maximum tolerable risk. This is certainly possible when (R, U) pairs form an exponential graph. In such cases, reducing the utility level by a small factor may result in a relatively large reduction of the disclosure risk. Hence, the optimum balance between R and U may in fact occur lower than the maximum tolerable risk, but this is not known by just examining the R-U confidentiality map.

Nevertheless, R-U confidentiality maps do not actually *determine* the optimum balance between privacy and information utility. That is, it can only *guide* the decision about how to balance the needs of privacy and information utility, by graphically representing the relationship between privacy and information utility. However, the decision where to strike the balance between privacy and information utility is still left up to the user of the R-U confidentiality map.

The research work described in this paper has been done in the context of a larger research project, the aim of which was to develop an optimal microdata anonymisation process. The recommendations provided in this paper were used as the basis for developing a solution for the optimal anonymisation of microdata. In a related paper (Zielinski & Olivier, 2009a), we use the recommendations provided here to address the optimisation aspect of the solution, where we use Economic Price Theory as the basis for determining the optimum levels of privacy and information utility that a microdata set should possess. The anonymisation aspect of the solution is addressed in another related paper (Zielinski & Olivier, 2009b), where we determine how microaggregation and *k*-anonymity should be used to anonymise the microdata such that the identified levels of privacy and information utility are achieved.

7 CONCLUSION

When microdata is anonymised, it needs to satisfy two conflicting goals: privacy and information utility. In this paper, we determined whether *k*-anonymity is appropriate in addressing this conflict. We have shown that the way in which *k*-anonymity is currently used to address this conflict is not appropriate, since it does not necessarily lead to an optimum balance between privacy and information utility. We also provided recommendations for the basis of a solution that will be appropriate for finding the optimum balance between privacy and information utility. We have subsequently used these recommendations to develop such a solution, which first determines the optimum levels of privacy and information utility (Zielinski & Olivier, 2009a) and then anonymises the microdata such that these optimum levels are achieved (Zielinski & Olivier, 2009b). This work focused on the conflict between privacy and information utility in microdata anonymisation. For future work, we aim to explore the conflict between privacy and information utility in other forms of statistical data, such as tabular data.

8 ACKNOWLEDGEMENT

The work presented in this paper is part of a larger research project lead by the author at SAP Research CEC Pretoria, South Africa. The support of SAP Research and the SAP Meraka Unit for Technology Development (UTD) towards this work is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the authors and should not necessarily be attributed to SAP Research or the SAP Meraka Unit for Technology Development (UTD).

9 REFERENCES

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigraphy, R., Thomas, D., et al. (2005). Achieving anonymity via clustering. In *Proceedings of the 10th International Conference on Database Theory*. Chicago, USA.

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., & Samarati, P. (2007). Microdata protection. In *Yu, T., Jajodia, S. (editors) Secure Data Management in Decentralized Systems* (pp. 291 - 321). Springer-Verlag.

Domingo-Ferrer, J., Sebe, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. *Computers and Mathematics with Applications, 55*, 714 - 732.

Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In *Doyle, P., Lane J.I., Theeuwes, J.J., Zayatz, L. (editors) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (pp. 111 - 134). North-Holland, Amsterdam.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery, 11*, 195 - 212.

Domingo-Ferrer, J., & Torra, V. (2008). A critique of k-anonymity and some of its enhancements. In *Proceedings of the 2008 Third*

International Conference on Availability, Reliability and Security. Barcelona, Spain.

Duncan, G. T., Feinberg, S. E., Krishnan, R., Padman, R., & Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In *Doyle, P., Lane J.I., Theeuwes, J.J., Zayatz, L. (editors) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (pp. 135 - 166). North-Holland, Amsterdam.

Duncan, G. T., Keller-McNulty, S. A., & Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, USA.

Gionis, A., & Tassa, T. (2009). k-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 206 - 219.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., et al. (2007). Handbook on statistical disclosure control, Version 1.01.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 277 - 286). Philadelphia, USA.

Loukides, G., & Shao, J. (2008). Data utility and privacy protection trade-off in k-anonymisation. In *Proceedings of the 2008 International workshop on Privacy and Anonymity in Information Society* (pp. 36 - 45). Nantes, France.

Medrano-Gracia, P., Pont-Tuset, J., Nin, J., & Muntes-Mulero, V. (2007). Ordered dataset vectorization for linear regression on data privacy. In *Proceedings of the 4th international conference on Modeling Decisions for Artificial Intelligence* (pp. 361 - 372). Kitakyushu, Japan.

Nin, J., Herranz, J., & Torra, V. (2008a). How to group attributes in multivariate microaggregation. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 161*, 121 - 138.

Nin, J., Herranz, J., & Torra, V. (2008b). On the disclosure risk of multivariate microaggregation. *Data and Knowledge Engineering, 67*(3), 399 - 412.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010 - 1027.

Sebe, F., Domingo-Ferrer, J., Mateo-Sanz, J. M., & Torra, V. (2002). Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In *Domingo-Ferrer, J. (editor) Inference Control in Statistical Databases, From Theory to Practice*, Lecture Notes in Compute Science (Vol. 2316, pp. 163 - 171). Springer-Verlag.

Stark, K., Eder, J., & Zatloukal, K. (2006). Priority-based k-anonymity accomplished by weighted generalisation structures. In *Proceedings of the 8th International Data Warehousing and Knowledge Discovery Conference* (pp. 394 - 404). Krakow, Poland.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 571 - 588.

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

Willenborg, L., & De Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics. Springer-Verlag.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. (2006). Utility-based anonymization for privacy preservation with less information loss. *ACM SIGKDD Explorations, 8*(2), 21 - 30.

Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Domingo-Ferrer, J. (editor) Inference Control in Statistical Databases, From Theory to Practice*, Lecture Notes in Computer Science (Vol. 2316, pp. 135 - 152).

Zhang, L., Jajodia, S., & Brodsky, A. (2007). Information disclosure under realistic assumptions: privacy versus optimality. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*. Alexandria, USA .

Zielinski, M. P. (2007a). Balancing privacy and information utility in microdata anonymisation. In *Proceedings of the 2007 Digital Identity and Privacy Conference*. Maastricht, The Netherlands.

Zielinski, M. P. (2007b). Privacy protection in eParticipation: guiding the anonymisation of microdata. In *Avdic, A., Hedstrom, K., Rose, J., Gronuld, A. (editors) Understanding eParticipation - Contemporary PhD eParticipation studies in Europe* (pp. 57 - 69). Örebro University Library, Sweden.

Zielinski, M. P., & Olivier, M. S. (2009a). On the use of Economic Price Theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymization. *(Submitted for publication).*

Zielinski, M. P., & Olivier, M. S. (2009b). How to determine the optimum number of records per cluster in microaggregation. *(Submitted for publication).*