WP.8
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

MICROAGGREGATION HEURISTICS FOR *P*-SENSITIVE *K*-ANONYMITY

Invited Paper

Prepared by Josep Domingo-Ferrer, Francesc Sebé and Agusti Solanas
Rovira I Virgili University, Catalonia, Spain

Microaggregation Heuristics for p-Sensitive k-Anonymity.

Josep Domingo-Ferrer, Francesc Sebé and Agusti Solanas

Rovira i Virgili University, Dept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. ({josep.domingo,francesc.sebe,agusti.solanas}@urv.cat)

Abstract. p-Sensitive k-anonymity is a sophistication of k-anonymity requiring that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. Like for k-anonymity, the computational approach originally proposed to achieve this property is based on generalization and suppression; this has several data utility problems, such as turning numerical key attributes into categorical, injecting new categories, injecting missing data, etc. We present and evaluate two heuristics for p-sensitive k-anonymity which, being based on microaggregation, overcome most of such drawbacks, while offering a smooth information loss increase as p and k grow.

1 Introduction

What is meant by database privacy largely depends on the context where this concept is being used. In official statistics, it normally refers to the privacy of the respondents to which the database records correspond (*respondent privacy*). In cooperative market analysis, it is understood as keeping private the databases owned by the various collaborating corporations (*data owner privacy*). In healthcare, both respondent and owner privacy are implicitly required: patients must keep their privacy and the medical records should not be transferred from a hospital to, say, an insurance company. In the context of dynamically queryable databases and, in particular, Internet search engines, the most rapidly growing concern is *user privacy*, that is, the privacy of the queries submitted by users (especially after scandals like the August 2006 disclosure of 658000 queries by the AOL search engine). Thus, what makes the difference is whose privacy is being sought.

Statistical disclosure control (SDC, see Dalenius, 1974; Willenborg and De Waal, 2001; Hundepool *et al.*, 2006) was born in the statistical community as a discipline to achieve respondent privacy. Privacy-preserving data mining (PPDM) appeared simultaneously in the database community (Agrawal and Srikant, 2000) and the cryptographic community (Lindell and Pinkas, 2000) with the aim of offering owner privacy: several database owners wish to compute queries across their databases in such a way that only the results of the queries are revealed to each other, not the contents of each other's databases. Finally, private information retrieval (PIR; Chor *et al.*, 1995) originated in the cryptographic community as an attempt to guarantee the privacy of user queries to databases.

Thus, the technologies to deal with the above three privacy dimensions (respondent, owner and user) have evolved in a fairly independent way within research communities with surprisingly little interaction. Fortunately, it turns out that some developments are useful for more than one privacy dimension, even if all three dimensions are independent (Domingo-Ferrer, 2007). Such is the case for k-anonymity and p-sensitive k-anonymity, which are useful properties both for respondent and owner privacy. Furthermore, in combination with private information retrieval, those two properties make all three privacy dimensions compatible. Thus, presenting efficient computational methods to meet those two properties is an especially relevant objective, which will be treated in this paper. Section 2 discusses k-anonymity for respondent and owner privacy, and recalls how to achieve it using microaggregation. Section 3 discusses p-sensitive k-anonymity and presents two heuristics to achieve this property via microaggregation. Section 4 contains an empirical performance evaluation of both heuristics. Conclusions are drawn in Section 5.

2 k-Anonymity for respondent and owner privacy

k-Anonymity is an interesting approach to face the conflict between information loss and disclosure risk, suggested by Samarati and Sweeney (1998). To recall the definition of k-anonymity, we need to enumerate the various (non-disjoint) types of attributes that can appear in a microdata set $\mathbf{X}$:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in $\mathbf{X}$ have been removed/encrypted.

- *Key attributes*. Borrowing the definition from Dalenius (1986), key attributes are those in $\mathbf{X}$ that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in $\mathbf{X}$ refer. Examples are job, address, age, gender, etc. Unlike identifiers, key attributes cannot be removed from $\mathbf{X}$, because any attribute is potentially a key attribute.

- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

We now have:

Definition. *A protected data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of key attributes, at least k records exist in the data set sharing that combination.*

If, for a given k, k-anonymity is assumed to be enough protection for respondents, one can concentrate on minimizing information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility. The original computational approach in Samarati

and Sweeney (1998) to achieve k-anonymity relies on suppressions and generalizations, so that minimizing information loss translates to reducing the number and/or the magnitude of suppressions.

The drawbacks of partially suppressed and coarsened data for analysis were highlighted in Domingo-Ferrer and Torra (2005):

1. Satisfying k-anonymity with minimum data modification using generalization (recoding) and local suppression was shown to be NP-hard in Meyerson and Williams (2004) and Aggarwal *et al.* (2004);

2. Using global recoding for generalization causes too much information loss, and using local recoding complicates data analysis by causing old and new categories to co-exist in the recoded file;

3. There is no standard way of using local suppression (at the tuple level, at the attribute level, with blanking, with replacement by neutral values, etc.);

4. Analyzing partially suppressed data usually requires specific software (imputation software, censored data analysis, etc.);

5. Last but not least, when numerical attributes are generalized, they become non-numerical.

Joint multivariate microaggregation (in the way of Domingo-Ferrer and Mateo-Sanz, 2002) of all key attributes with minimum group size k was proposed in Domingo-Ferrer and Torra (2002) as an alternative to achieve k-anonymity; besides being simpler, this alternative has the advantage of yielding complete data without any coarsening (nor categorization in the case of numerical data). As a reminder, microaggregation seeks to split a data set into groups of records such that each group contains at least k records and groups are as homogeneous as possible; then records within a group are replaced with the average of all records in the group. Clearly, the higher the homogeneity of records in a group, the lower the information loss caused by replacement of those records by their average. In the case of the k-anonymity application, microaggregation is performed on the projection of records on key attributes, rather than on the entire records. If the microaggregated attributes are numerical, group homogeneity can be measured by the within-groups sum of squares SSE: the smaller SSE, the more homogeneous are the groups.

In Aggarwal and Yu (2004), masking through condensation (actually a special case of multivariate microaggregation) is proposed to achieve k-anonymity in the context of privacy-preserving data mining, and thus with the aim of owner privacy.

3 p-Sensitive k-anonymity via microaggregation

k-Anonymity is able to prevent identity disclosure, *i.e.* a record in the k-anonymized data set cannot be mapped back to the corresponding record in the original data set. However, in general, it may fail to protect against attribute disclosure. In Truta and Vinay (2006), an evolution of k-anonymity called p-sensitive k-anonymity was

presented. Its idea is that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. The following example illustrates a case where p-sensitive k-anonymity is useful because k-anonymity alone does not offer enough protection.

Example. *Imagine that an individual's health record is k-anonymized into a group of k patients with k-anonymized key attributes values* Age = "30", Height = "180 cm" *and* Weight = "80 kg". *Now, if all k patients share the confidential attribute value* Disease = "AIDS", *k-anonymization is useless, because an intruder who uses the key attributes* (Age, Height, Weight) *can link an external identified record*

(Name="John Smith", Age="31", Height="179", Weight="81")

with the above group of k patients and infer that John Smith suffers from AIDS (attribute disclosure).

Based on the above remarks, the following definition can be given

Definition. *A data set is said to satisfy p-sensitive k-anonymity for $k > 1$ and $p \leq k$ if it satisfies k-anonymity and, for each group of tuples with the same combination of key attribute values that exists in the data set, the number of distinct values for each confidential attribute is at least p within the same group.*

The computational approach proposed in Truta and Vinay (2006) and Truta *et al.* (2007) to achieve p-sensitive k-anonymity is an extension of the generalization/suppression procedure proposed in the original k-anonymity papers. Therefore it shares the same shortcomings pointed out in Domingo-Ferrer and Torra (2005) and listed above.

Like we did for k-anonymity in Domingo-Ferrer and Torra (2005), in Domingo-Ferrer (2006) we showed a way to achieve p-sensitive k-anonymity via microaggregation.

The goal is to obtain p-sensitive k-anonymous data sets without coarsened nor partially suppressed data. This makes their analysis and explotation easier, with the additional advantage that numerical continuous attributes are not categorized.

Note. *In addition to p-sensitive k-anonymity, a number of other sophistications of k-anonymity for protecting against attribute disclosure have recently been proposed, such as l-diversity (Machanavajjhala, 2006), (α, k)-anonymity (Wong* et al.*, 2006), t-closeness (Li* et al.*, 2007) and m-confidentiality (Wong* et al.*, 2007). All of them rely on generalizations, so the microaggregation approach proposed in this paper would be a novelty in all of them. For the sake of concreteness, we will focus here on p-sensitive k-microaggregation.*

We next present two different heuristics for microaggregation-based p-sensitive k-anonymity. The first one starts by achieving k-anonymity and then achieves p-sensitivity. The second one first achieves p-sensitivity and then k-anonymity.

3.1 k-Anonymity first

The heuristic in this section was described in Domingo-Ferrer (2006) without performance analysis and is as follows:

Algorithm 1 (k-Anonymity first)

1. *If $p > k$, signal an error ("p-sensitive k-anonymity infeasible") and exit the Algorithm.*
2. *If the number of distinct values for any confidential attribute in $\mathbf{X}$ is less than p over the entire dataset, signal an error ("p-sensitive k-anonymity infeasible") and exit the Algorithm.*
3. *k-Anonymize $\mathbf{X}$ using the MDAV microaggregation algorithm described in Domingo-Ferrer and Torra (2005). Let $\mathbf{X}'$ be the microaggregated, k-anonymized dataset.*
4. *Let $\hat{k} := k$.*
5. *While p-sensitive k-anonymity does not hold for $\mathbf{X}'$ do:*
   (a) *Let $\hat{k} := \hat{k} + 1$.*
   (b) *$\hat{k}$-Anonymize $\mathbf{X}$ using microaggregation. Let $\mathbf{X}'$ be the $\hat{k}$-anonymized dataset.*

The above algorithm is based on the following facts:

- A $k+1$-anonymous dataset is also k-anonymous;
- By increasing the minimum group size, the number of distinct values for confidential attributes hopefully increases (in the extreme case, if there is a single group as large as the entire dataset, all distinct values for all attributes are in the group).

3.2 p-Sensitivity first

The heuristic below first achieves p-sensitivity and then completes groups in order for them to include k or more records.

Algorithm 2 (p-Sensitivity first)

1. *Let $x_1, x_2, \ldots, x_n$ be the records in the original data set $\mathbf{X}$; let L be the set of confidential attributes. Let Q be the set of key attributes and let $x_j(Q)$ be the projection of record x_j on its key attributes.*
2. *Let P be an initially empty partition.*
3. *While there are at least k records in $\mathbf{X}$ and such records contain at least p different values for each attribute in L do:*
   (a) *Compute the average record $\overline{x}(Q)$ of the projections $x_1(Q), \ldots, x_n(Q)$.*
   (b) *Consider record $x_r \in \mathbf{X}$ so that Euclidean distance between $x_r(Q)$ and $\overline{x}(Q)$ is maximum.*

(c) *Create a new group C that initially contains record x_r.*

(d) *While confidential attributes of the records in C do not satisfy p-sensitivity:*

i. *Take $x_s \in \mathbf{X}$ so that $x_s(Q)$ is the nearest record to $x_r(Q)$ such that $x_s(L)$ contributes to the compliance of p-sensitivity by $C(L)$ (records in C projected on the confidential attributes in L).*

ii. *Add x_s to C and remove it from $\mathbf{X}$.*

(e) *While C does not contain at least k records:*

i. *Take $x_s \in \mathbf{X}$ so that the distance between $x_s(Q)$ and $x_r(Q)$ is minimum.*

ii. *Add x_s to C and remove it from $\mathbf{X}$.*

(f) *Add C to P.*

4. *For each record x remaining in $\mathbf{X}$:*

- *Add x to the group $C \in P$ satisfying that the distance from $x(Q)$ to $Centroid(C)(Q)$ (the centroid of the projections on Q of records in C) is minimum.*

5. *For $i = 1$ to n:*

- *Let x'_i be x_i with $x_i(Q)$ replaced by $Centroid(C)(Q)$, where C is the group in P to which x_i has been assigned.*

6. *The microaggregated, p-sensitive, k-anonymous data set X' is formed by records $x'_1, \cdots x'_n$.*

4 Empirical results

The test data set was generated from the "Census" data file, which was used in the European CASC project (Brand *et al.*, 2002) and in several papers in the microaggregation literature (Domingo-Ferrer *et al.*, 2001; Dandekar *et al.*, 2002; Yancey *et al.*, 2002; Laszlo and Mukherjee, 2005; Domingo-Ferrer and Torra, 2005; Domingo-Ferrer *et al.*, 2007). This data set contains 1080 records with 13 numerical attributes.

The following procedure was used to generate the test data set:

1. The first six attributes of "Census" were taken as key attributes. These attributes were standardized by subtracting their mean and dividing by their standard deviation.

2. Three categorical confidential attributes with 15 categories each were generated from the next three continuous attributes from "Census", respectively. To generate a categorical attribute with 15 categories from a numerical attribute, the range comprised between the minimum value and the maximum value of the attribute is divided into 15 intervals of the same length. Continuous values falling into the first interval are recoded into the first category, those falling into the second interval are recoded into the second category, and so on.

Algorithms 1 and 2 were run for different values of k and p. Results are presented in Tables 1 and 2, respectively.

Table 1: $100 \times SSE/SST$ ratio of the p-sensitive k-anonymous microaggregations yielded by Algorithm 1

	p				
k	**1**	**3**	**5**	**7**	**10**
3	3.69	44.96			
5	6.20	44.96	58.85		
7	7.93	44.96	58.85	71.67	
10	9.71	44.96	58.85	71.67	100

Table 2: $100 \times SSE/SST$ ratio of the p-sensitive k-anonymous microaggregations yielded by Algorithm 2

	p				
k	**1**	**3**	**5**	**7**	**10**
3	3.69	23.13			
5	6.20	23.28	47.15		
7	7.93	22.31	47.15	57.63	
10	9.71	23.13	47.15	57.63	100

Some comments on Tables 1 and 2 follow:

- For $p = 1$, Algorithm 1 is equivalent to the MDAV microaggregation algorithm (Domingo-Ferrer and Torra, 2005). For $p = 1$, Algorithm 2 is equivalent to Centroid-based fixed-size microaggregation (CBFS, Laszlo and Mukherjee, 2005), an algorithm very similar to MDAV. This similarity explains that for $p = 1$ and all values of k reported in Tables 1 and 2, the same SSE/SST ratio is obtained.

- However, for higher values of p, Algorithm 1 behaves clearly worse than Algorithm 2, with higher SSE/SST ratios. The explanation is that the average size of the computed groups is greater for Algorithm 1 than for Algorithm 2: indeed, not caring about p-sensitivity from the start results in a penalty in terms of group size and, consequently, of SSE/SST ratio.

- The reason that Table 1 does not seem to reflect a dependency on k for $p = 3, 5, 7, 10$ is that the actual minimum group size $\hat{k}$ computed by Algorithm 1 turns out to be always greater than or equal to $k = 10$.

- In Table 2, such a lack of dependency on k seems to occur for $p = 5, 7, 10$ also for Algorithm 2: again the explanation is that, for those values of p, the minimum group size if greater than or equal to 10.

- For $p = 10$, both algorithms yield an SSE/SST ratio equal 100. This means that they yield a partition with a single group. This is not really surprising, because the categorical attributes in our data set have only 15 different categories, some of which are rare (*e.g.* those intervals corresponding to the tails of the attribute range).

5 Concluding discussion

p-Sensitive k-anonymity is a sophistication of k-anonymity, whose idea is to avoid that all records sharing a combination of key attributes in a k-anonymous data set also share the values for one or more confidential attributes. The computational approach originally proposed to achieve this new property is based on generalization and suppression and has a number of data utility problems enumerated above.

We have proposed and evaluated two heuristics for p-sensitive k-anonymity which, being based on microaggregation, preserve the numerical nature of key attributes, do not introduce missing data and gracefully degrade data utility as p grows. The first heuristic starts by ensuring k-anonymity and then seeks to achieve p-sensitivity. The second heuristic proceeds the other way round: first p-sensitivity is satisfied and then k-anonymity. This second strategy seems to be clearly better in terms of within-groups homogeneity and, consequently, of data utility.

Disclaimer and acknowledgments

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Ministry of Education through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446.

References

Aggarwal, C. C. and Yu, P. S. (2004) "A condensation approach to privacy preserving data mining". In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari, editors, *Advances in Database Technology - EDBT 2004*, volume 2992 of *Lecture Notes in Computer Science*, pages 183–199, Berlin Heidelberg.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D.,and Zhu, A. (2004) "k-Anonymity: Algorithms and hardness". Technical report, Stanford University.

Agrawal, R., and Srikant, R. (2000) "Privacy preserving data mining". In *Proceedings of the ACM SIGMOD*, pages 439–450. ACM.

Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002) "Reference data sets to test and compare SDC methods for protection of numerical microdata". European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.

Chor, B., Goldreich, O., Kushilevitz, E., and Sudan, M. (1995) "Private information retrieval". In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 41–50.

Dalenius, T. (1974) "The invasion of privacy problem and statistics production. an overview". *Statistik Tidskrift*, 12: 213–225.

Dalenius, T. (1986) "Finding a needle in a haystack - or identifying anonymous census records". *Journal of Official Statistics*, 2(3):329–336.

Dandekar, R., Domingo-Ferrer, J., and Sebé, F. (2002) "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection". In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, Springer.

Domingo-Ferrer, J. (2006) "Microaggregation for database and location privacy". In O. Etzion, T. Kuflik, and A. Motro, editors, *Next Generation Information Technologies and Systems-NGITS 2006*, volume 4032 of *Lecture Notes in Computer Science*, pages 106–116, Berlin Heidelberg.

Domingo-Ferrer, J. (2007) "A three-dimensional conceptual framework for database privacy". In *Secure Data Management-4th VLDB Workshop SDM'2007*, volume 4721 of *Lecture Notes in Computer Science*, pages 193–202, Berlin Heidelberg, 2007.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002) "Practical data-oriented microaggregation for statistical disclosure control". *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001) "Comparing SDC methods for microdata on the basis of information loss and disclosure risk". In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg: Eurostat.

Domingo-Ferrer, J., Sebé, F., and Solanas, A. (2007) "A polynomial-time approximation to optimal multivariate microaggregation", *Computers & Mathematics with Applications*, (to appear).

Domingo-Ferrer, J., and Torra, V. (2005) "Ordinal, continuous and heterogeneous k-anonymity through microaggregation". *Data Mining and Knowledge Discovery*, 11(2):195–212.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., and DeWolf, P.-P. (2006) *Handbook on Statistical Disclosure Control (version 1.0)*. Eurostat (CENEX SDC Project Deliverable).

Laszlo, M., and Mukherjee, S. (2005) "Minimum spanning tree partitioning algorithm for microaggregation". *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911.

Li, N., Li, T., and Venkatasubramanian, S. (2007) "T-closeness: privacy beyond k-anonymity and l-diversity". In *Proceedings of the IEEE ICDE 2007*.

Lindell, Y., and Pinkas, B. (2000) "Privacy preserving data mining". In *Advances in Cryptology - CRYPTO'00*, volume 1880 of *Lecture Notes in Computer Science*, pages 36–53, Berlin Heidelberg.

Machanavajjhala, A., Gehrke, J., Kiefer, D., and Venkitasubramanian, M. (2006) "L-diversity: privacy beyond k-anonymity". In *Proceedings of the IEEE ICDE 2006*, 2006.

Meyerson, A., and Williams, R. (2004) "On the complexity of optimal k-anonymity". In *Proc. of the ACM Symposium on Principles of Database Systems-PODS'2004*, pages 223–228, Paris, France, ACM.

Samarati, P., and Sweeney, L. (1998) "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". Technical report, SRI International.

Truta, T. M., Campan, A., and Meyer, P. (2007) "Generating microdata with p-sensitive k-anonymity". In *Secure Data Management-4th VLDB Workshop SDM'2007*, volume 4721 of *Lecture Notes in Computer Science*, pages 124–141, Berlin Heidelberg.

Truta, T. M. and Vinay, B. (2006) "Privacy protection: p-sensitive k-anonymity property". In *2nd International Workshop on Privacy Data Management PDM 2006*, page 94, Berlin Heidelberg. IEEE Computer Society.

Willenborg, L., and DeWaal, T. (2001) *Elements of Statistical Disclosure Control*. Springer-Verlag, New York.

Wong, R. C. W., Fu, A. W. C., Wang, K. and Pei, J. (2007) "Minimality attack in privacy preserving data publishing". In *Proceedings of the VLDB 2007*, Vienna.

Wong, R. C. W., Li, J., Fu, A. W. C., and Wang, K. (2006) "(α,k)-Anonymity: an enhanced k-anonymity model for privacy-preserving data publishing". In *Proceedings of the ACM KDD*, pages 754–759, New York.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002) "Disclosure risk assessment in perturbative microdata protection". In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, Springer.