An Anonymity Model Achievable Via Microaggregation[⋆]

Josep Domingo-Ferrer, Francesc Sebé, and Agusti Solanas

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

Abstract. k-Anonymity is a privacy model requiring that all combinations of key attributes in a database be repeated at least for k records. It has been shown that k-anonymity alone does not always ensure privacy. A number of sophistications of k-anonymity have been proposed, like p-sensitive k-anonymity, l-diversity and t-closeness. We identify some shortcomings of those models and propose a new model called (k, p, q, r)-anonymity. Also, we propose a computational procedure to achieve this new model that relies on microaggregation.

Keywords: Microaggregation, statistical databases, privacy, microdata protection, statistical disclosure control.

1 Introduction

Database privacy can be described in terms of three dimensions, as suggested in [6]: respondent privacy, data owner privacy and user privacy. Respondent privacy is about ensuring the respondents to which the database records correspond that no re-identification is possible. The need for data owner privacy arises in a context with several organizations engaged in collaborative computation and consists of each organization keeping private the database it owns. User privacy refers to the privacy of the queries submitted by users to a queryable database or search engine.

Statistical disclosure control (SDC, [4, 18]) was born in the statistical community as a discipline to achieve respondent privacy. Privacy-preserving data mining (PPDM) appeared simultaneously in the database community [1] and the cryptographic community [11] with the aim of offering owner privacy: several database owners wish to compute queries across their databases in 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, [3]) originated in

[⋆] The authors with the UNESCO Chair in Data Privacy, but the views expressed in this paper are those of the authors and do not commit UNESCO. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES" and by the Government of Catalonia under grant 2005 SGR 00446.

W. Jonker and M. Petković (Eds.): SDM 2008, LNCS 5159, pp. 209–218, 2008.

the cryptographic community as an attempt to guarantee privacy for user queries to databases. Although the technologies to deal with the above three privacy dimensions have evolved in a fairly independent way within research communities with surprisingly little interaction, it turns out that some developments are useful for more than one privacy dimension, even if all three dimensions are independent (see [6]). Such is the case for k-anonymity and its enhancements. Thus, improving k-anonymity-related privacy models both conceptually and computationally is an especially relevant objective that will be pursued in this paper. Section 2 is a critical review of k-anonymity and its enhancements. Section 3 presents the (k, p, q, r)-anonymity model. A microaggregation-based heuristic to achieve (k, p, q, r)-anonymity is described in Section 4. Empirical results are reported in Section 5. Finally, Section 6 lists some conclusions and future research issues.

2 A Critical Review of *k*-anonymity and Its Enhancements

We review in this section the definition and the limitations of the following privacy models: k-anonymity, p-sensitive k-anonymity, l-diversity, (α, k)-anonymity and t-closeness.

2.1 *k*-anonymity

k-Anonymity is an interesting approach suggested by Samarati and Sweeney [14] to deal with the tension between information loss and disclosure risk. It can also be understood as a kind of indistinguishability, as suggested in [22]. 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 X:

- *Identifiers.* These are attributes that *unambiguously* identify the respondent.
- *Key attributes.* Borrowing the definition from [5,15], key attributes are those in X that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in X refer.
- *Confidential attributes.* These are attributes which contain sensitive information on the respondent.

Definition 1 (k-Anonymity). *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.

k-Anonymity is able to prevent identity disclosure, *i.e.* a record in the k-anonymized data set will be correctly mapped back to the corresponding record in the original data set with a probability at most $1/k$. However, in general, it may fail to protect against attribute disclosure.

2.2 p-Sensitive k-anonymity

In [17], an evolution of k-anonymity called p-sensitive k-anonymity was presented. Its purpose is to protect against attribute disclosure by requiring that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. The formal definition is as follows.

Definition 2 (p-Sensitive k-anonymity). *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.*

An attacker trying to obtain the confidential value for a given record that has been linked to the p-sensitive k-anonymous data set will not be able to determine which of the p different values inside the group is the corresponding one. p-Sensitive k-anonymity may cause a huge data utility loss in some data sets. In some cases, p-Sensitive k-anonymity is insufficient to prevent attribute disclosure due to the skewness attack and the similarity attack.

2.3 l-Diversity

Like p-sensitive k-anonymity, l-diversity [12] attempts to solve the attribute disclosure problem that can happen with k-anonymity.

Definition 3 (l-Diversity). *A data set is said to satisfy l-diversity if, for each group of records sharing a combination of key attributes, there are at least l "well-represented" values for each confidential attribute.*

According to [12] "well-represented" can be defined in several ways:

1. *Distinct l-diversity.* There must be at least l distinct values for the confidential attribute in each group of records sharing a combination of key attributes.
2. *Entropy l-diversity.* The entropy of a group G for a particular confidential attribute with domain C can be defined as
$$H(G) = -\sum_{c \in C} p(G,c) \log p(G,c)$$
in which $p(G,c)$ is the fraction of records in G which have value c for the sensitive attribute. A data set is said to satisfy entropy l-diversity if for each group G, $H(G) \geq \log l$.
3. *Recursive (c,l)-diversity.* This model makes sure that the most frequent values do not appear too frequently and the least frequent values do not appear too rarely. Let m be the number of values of the confidential attribute in a group G and r_i, for $1 \leq i \leq m$, be the number of times that the i-th most frequent value appears in G. Then G is said to satisfy recursive (c,l)-diversity if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$. A data set is said to satisfy recursive (c,l)-diversity if all of its groups satisfy recursive (c,l)-diversity.

Distinct l-diversity may be vulnerable to skewness and similarity attacks in the same way p-sensitive k-anonymity is. Regarding entropy l-diversity and (c, l)-diversity, both models aim at preventing skewness attacks, but the risk of similarity attacks still remains. All three variants may introduce high information loss in some cases.

2.4 (α, k)-Anonymity

(α, k)-Anonymity was proposed in [19] as follows:

Definition 4 ((α, k)-Anonymity). *A data set is said to satisfy (α, k)-anonymity if it is k-anonymous and, for each group of records sharing a combination of key attributes, the proportion of each sensitive value is at most α, where $\alpha \in [0, 1]$ is a user parameter.*

For $\alpha = 1/k$, this model becomes k-sensitive k-anonymity. This model prevents attribute disclosure (an upper-bound of α on the probability of a correct attribute value estimation is provided) but skewness and similarity attacks are still possible. High information loss may also be incurred during the anonymization procedure.

2.5 (k, e)-Anonymity

Models discussed in this section so far are designed for categorical confidential attributes. In [23], the following model is proposed for numerical attributes:

Definition 5 ((k, e)-Anonymity). *A data set D is said to satisfy (k, e)-anonymity if, given D and any public database P, any association cover that an attacker can derive satisfies: (i) the size of the association cover is no less than k; (ii) the range of the confidential attribute values in the association cover is no less than e.*

Clearly, what is called association cover in Definition 5 can be assimilated to a group of records sharing a combination of key attribute values.

(k, e)-Anonymity tries to overcome the similarity attack described above by requiring a minimum range in the values of the confidential attribute. Besides potentially inflicting a substantial information loss to the data, this model does not guard against skewness attacks.

2.6 t-Closeness

In [10], a new privacy model called t-closeness is defined as follows.

Definition 6 (t-Closeness). *A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the distance between the distribution of the confidential attribute in the group and its distribution in the whole data set is no more than a threshold t.*

t-Closeness solves the attribute disclosure vulnerabilities inherent to previous models (i.e. skewness attack and similarity attack). However, some criticisms can be made to t-closeness:

- No computational procedure to enforce t-closeness is given.
- If such a procedure were available, it would greatly damage the utility of data. The only way to decrease the damage is to increase the threshold t, that is, to relax t-closeness.

2.7 Other Models

The models discussed so far share with k-anonymity the lack of assumptions on the intruder's capabilities or the public databases available to the intruder. If such assumptions are made, a number of other evolutions of k-anonymity can still be found in the recent literature: m-confidentiality [20], personalized privacy preservation [21] and (c, k)-safety [13]. These evolutions are mentioned for completeness, but our focus will be to enhance the k-anonymity-like models recalled in the previous sections, which do not make assumptions about the intruder.

3 The (k, p, q, r)-anonymity Model

From the lessons learned on the limitations of the models described in Section 2, we can define a new model as follows:

Definition 7 ((k, p, q, r)-Anonymity). *A data set is said to satisfy (k, p, q, r)-anonymity if it is k-anonymous and satisfies that:*

- *It is p-sensitive only for those groups where values of confidential attributes appear whose relative frequency is less than q in the overall data set.*
- *For groups where p-sensitivity holds, the ratio between the within-group variance of confidential attributes and their variance over the entire data set is at least r.*

The rationale of the model in Definition 7 is explained in the rest of this section. The variance for numerical attributes is the standard statistical variance. For categorical attributes (ordinal or nominal), specific variance definitions are needed, which can be found in [7] and [9].

(k, p, q, r)-Anonymity guarantees k-anonymity for the key attributes in the data set. Regarding the confidential attributes:

- It guarantees p-sensitivity in those groups where "rare" values of a confidential attribute are present (with relative frequency less than q).
- Disclosure of non-rare values of confidential attributes is not considered a privacy problem. The advantage of suppressing the p-sensitivity requirement for very frequent confidential attribute values is that smaller groups (of size closer to the lower bound k) are feasible, which causes less data utility loss as far as key attributes are concerned. In this data utility respect, the new model outperforms p-sensitive k-anonymity, l-diversity, (α, k)-anonymity, (k, e)-anonymity and t-closeness which may all yield in very large

groups, because they attempt to prevent attribute disclosure even for very frequent values of the confidential attribute.

- Finally, enforcing a lower bound for the within-group variance of confidential attributes is meant to thwart the similarity attack which is possible against k-anonymity, p-sensitive k-anonymity, l-diversity, (α, k)-anonymity and (k, e)-anonymity.

4 A Heuristic for (k, p, q, r)-Anonymity

In this section we present a computational procedure to achieve (k, p, q, r)-anonymity for data sets with numerical key attributes and one confidential attribute. Let $x_1, x_2, \ldots, x_n$ be the records in the original data set X. Let L be the confidential attribute and Q be the set of key attributes. Let $x_j(Q)$ denote the projection of record x_j on its key attributes and $x_j(L)$ denote the projection of record x_j on its confidential attribute.

The proposed heuristic procedure is as follows:

```
1. Label as ‘sensitive’ those records in $X$ whose confidential attribute takes a value
   appearing less than $q \cdot n$ times in $X$. Let $Y \subseteq X$ be the subset of sensitive records;
2. Compute $Var := Variance(Y(L))$;
3. Compute $MinVar := r \cdot Var$;
4. While $NotEmpty(Y)$ loop
   (a) Let $C$ be a new empty group;
   (b) Let $x_s$ be a random sensitive record from $Y$;
   (c) Add $x_s$ to $C$ and remove it from $Y$ and $X$;
   (d) While elements in $C(L)$ do not satisfy $p$-sensitivity loop
         i. Take $x_t \in X$ such that $x_t(Q)$ is the nearest record to $x_s(Q)$ which:
              – contributes to the compliance of $p$-sensitivity by $C(L)$;
              – increases $Variance(C(L))$ if added to $C$;
        ii. If no record satisfying the above two conditions is found, take $x_t$ such that
            $x_t(Q)$ is the nearest record to $x_s(Q)$ that contributes to the compliance of
            $p$-sensitivity by $C(L)$;
       iii. Add $x_t$ to $C$ and remove it from $X$ (and from $Y$ if $x_t \in Y$);
       end loop
   (e) While $Variance(C(L)) < MinVar$ loop
         i. Take $x_t \in X$ such that $x_t(Q)$ is the nearest record to $x_s(Q)$ which increases
            $Variance(C(L))$ if added to $C$;
        ii. Add $x_t$ to $C$ and remove it from $X$ (and from $Y$ if $x_t \in Y$);
       end loop
   (f) While $Cardinality(C) < k$ loop
         i. Take $x_t$ such that $x_t(Q)$ is the nearest record to $x_s(Q)$ which keeps
            $Variance(C(L)) \geq MinVar$ if added to $C$;
        ii. Add $x_t$ to $C$ and remove it from $X$ (and from $Y$ if $x_t \in Y$);
       end loop
   (g) If ($Variance(Y(L)) < MinVar$) or ($p$-sensitivity of $Y(L) < p$) then
         i. Add the remaining records from $Y$ to $C$;
        ii. Remove from $X$ all records in $Y$;
       iii. Remove all records from $Y$;
```

```
            iv. If ($Variance(C) < MinVar$) then
                  Remove from $C$ those records not having a 'sensitive' value and return
                    them to $X$;
                end if
            end if
      (h) Add $C$ to partition $P$;
      end loop
  5. Apply MDAV [8] to build a $k$-partition of records in $X$ and add the MDAV-
     generated groups to $P$;
  6. Microaggregate the records, that is, for $i = 1$ to $n$ replace $x_i(Q)$ by the centroid of
     $C_i(Q)$, where $C_i$ is the group in $P$ to which $x_i$ has been assigned.
```

Each iteration of Step (4) constructs one group containing 'sensitive' records. These groups are those that must satisfy the constraints given by parameters k, p, q and r. Such constraints are satisfied by the loops nested inside Step (4):

- Each new group C is initialized by assigning a random 'sensitive' record to it (Substep (4c));
- Next, Substep (4d) is iterated until C satisfies p-sensitivity, this is, the records in C contain at least p different values for the confidential attribute; if possible, records to be added to C are chosen so that they increase the variance of $C(L)$;
- After that, Substep (4e) ensures the variance of $C(L)$ is at least the one specified by parameter r; this step iterates until this condition is satisfied;
- Then, Substep (4f) is iterated until C has at least k records (in this way, the constraint specified by k is satisfied); once we get out of this loop, C is guaranteed to satisfy the properties of the (k, p, q, r)-anonymity model;
- Finally, Substep (4g) checks that the remaining records in X will be able to form a new group satisfying the model; if this is not the case, they are added to the last group C.

Once no more 'sensitive' records are left in X, the remaining ones are clustered at Step (5) using the MDAV heuristic [8]. Finally, Step (6) replaces each record x_i with its microaggregated version.

5 Empirical Results

In this section, empirical results on the proposed heuristic are reported and compared with those obtained with k-anonymization (based on microaggregation [8]) and p-sensitive k-anonymization (based on the random initial point variant of the microaggregation heuristic [16]). The information loss is reported as $100 \cdot SSE/SST$, where SSE is the within-groups sum of squares and SST is the total sum of squares. A synthetic data set obtained from the "Census" benchmark file [2] has been used. In our first experiment we have used a data set with 1080 records. Each record has 12 continuous numerical key attributes that have been standardized. The confidential attribute takes integer values in the range from 1 to 10; the attribute has been initialized so that each value appears

Table 1. Information loss of our (k, p, q, r)-anonymity heuristic for $k = 5$, $p = 4$, $q = 0.2$ and different values of r

r	0.1	0.3	0.5	0.7	0.9
$Inf.loss$	11.98	12.09	13.01	30.85	68.518

Table 2. Information loss under several models for several values of k and p and fixed $q = 0.2$ and $r = 0.5$ (unskewed confidential attribute)

k	p	k-anonymity	(k, p, q, r)-anonymity	p-sensitive k-anonymity
3	2	5.58	11.87	7.24
4	3	7.52	11.58	9.81
7	5	11.53	14.69	14.42

in exactly 108 records. Parameter q was set to 0.2 so that all values of the confidential attribute were considered as 'sensitive'. Table 1 shows the information loss of (k, p, q, r)-anonymity for $k = 5$ and $p = 4$ and different values of parameter r. As expected, information loss increases with r. This is due to the fact that higher values of r force the heuristic to form groups with a higher variance of its confidential attribute. The higher r, the more constrained are groups, which increases information loss. The k-anonymous version of the data set used had an information loss of 9.21 and the k-anonymous p-sensitive version had 12.31.

Using the same data set of the previous experiment and for fixed $q = 0.2$, $r = 0.5$ and different values of k and p, a second experiment was carried out to compare the information loss incurred by k-anonymity, p-sensitive k-anonymity and (k, p, q, r)-anonymity. Results are shown in Table 2. It can be seen that k-anonymity presents the lowest information loss. The reason is that it is the model with least restrictions. For p-sensitive k-anonymity and (k, p, q, r)-anonymity the information loss is roughly similar; strictly speaking it is a bit higher for (k, p, q, r)-anonymity due to the additional constraint introduced by parameter r which forces a minimum variance of the confidential attributes in a group.

In the third experiment, we modified the distribution of the values of the confidential attribute. Values from 1 to 9 appeared 10 times each while value 10 appeared 990 times. We took $q = 0.2$, so that records with confidential value 10 were considered 'non sensitive' by the (k, p, q, r)-heuristic. The results are shown in Table 3. It can be seen that (k, p, q, r)-anonymity outperforms p-sensitive k-anonymity.

Table 3. Information loss under several models for several values of k and p and fixed $q = 0.2$ and $r = 0.5$ (skewed confidential attribute)

k	p	k-anonymity	(k, p, q, r)-anonymity	p-sensitive k-anonymity
3	2	5.58	9.47	16.42
4	3	7.52	12.13	22.72
7	5	11.53	18.97	30.42

6 Conclusions and Future Research

We have presented (k, p, q, r)-anonymity as a new security model which outperforms most current security models in the literature: it guarantees p-sensitivity for rare values and offers protection against the similarity attack, one of the most difficult to thwart. The model behaves in a pragmatic way (no p-sensitivity for frequent values) in order to reduce information loss. The only attack for which no defense is offered is skewness, but we have shown that such an attack can only be countered at the expense of a very substantial information loss (using the t-closeness model). Future research will involve designing other heuristic procedures, which can accommodate non-numerical quasi-identifiers and can deal with more than one confidential attribute.

References

1. Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proceedings of the ACM SIGMOD, pp. 439–450 (2000)
2. Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.M.: Reference data sets to test and compare SDC methods for protection of numerical microdata., European Project IST-2000-25069 CASC (2002), `http://neon.vb.cbs.nl/casc`
3. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 41–50 (1995)
4. Dalenius, T.: The invasion of privacy problem and statistics production. An overview. Statistik Tidskrift 12, 213–225 (1974)
5. Dalenius, T.: Finding a needle in a haystack - or identifying anonymous census records. Journal of Official Statistics 2(3), 329–336 (1986)
6. Domingo-Ferrer, J.: A three-dimensional conceptual framework for database privacy. In: Jonker, W., Petković, M. (eds.) SDM 2007. LNCS, vol. 4721, pp. 193–202. Springer, Heidelberg (2007)
7. Domingo-Ferrer, J., Solanas, A.: A measure of variance for nominal attributes (manuscript, 2008)
8. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14, 189–201 (2002)
9. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
10. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE ICDE (2007)
11. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–53. Springer, Heidelberg (2000)
12. Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkatasubramanian, S.: l-Diversity: privacy beyond k-anonymity. In: Proceedings of the IEEE ICDE 2006 (2006)
13. Martin, D.J., Kiefer, D., Machanavajjhala, A., Gehrke, J.: Worst-case background knowledge for privacy-preserving data publishing. In: Proceedings of the IEEE ICDE 2007 (2007)

14. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression., Tech. Report, SRI International (1998)
15. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
16. Solanas, A., Sebé, F., Domingo-Ferrer, J.: Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond. In: Extending Database Technology, EDBT 2008 (2008)
17. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: 2nd International Workshop on Private Data Management PDM 2006. IEEE Computer Society Press, Los Alamitos (2006)
18. Willenborg, L., DeWaal, T.: Elements of Statistical Disclosure Control. Springer, Heidelberg (2001)
19. Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K.: (α, k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In: Proceedings of the KDD 2006 (2006)
20. Wong, R.C.-W., Fu, A.W.-C., Wang, K., Pei, J.: Minimality attack in privacy preserving data publishing. In: Proceedings of the VLDB 2007, pp. 543–554 (2007)
21. Xiao, X., Tao, Y.: Personalized privacy preservation. In: SIGMOD Conference 2006, pp. 229–240 (2006)
22. Yao, C., Wang, L., Wang, X.S., Jajodia, S.: Indistinguishability: The Other Aspect of Privacy. In: Jonker, W., Petković, M. (eds.) SDM 2006. LNCS, vol. 4165, pp. 1–17. Springer, Heidelberg (2006)
23. Zhang, Q., Koudas, N., Srivastava, D., Yu, T.: Aggregate query answering on anonymized tables. In: Proceedings of the IEEE ICDE 2007, pp. 116–125 (2007)