Micro-aggregation-Based Heuristics for p-sensitive k-anonymity: One Step Beyond

Agusti Solanas
CRISES Research Group
UNESCO Chair in Data Privacy
Dept. Computer Engineering and Mathematics
Rovira i Virgili University
Av. Països Catalans, 26
E- 43007 Tarragona.
Catalonia. Spain
agusti.solanas@urv.cat

Francesc Sebé
CRISES Research Group
UNESCO Chair in Data Privacy
Dept. Computer Engineering and Mathematics
Rovira i Virgili University
Av. Països Catalans, 26
E- 43007 Tarragona.
Catalonia. Spain
francesc.sebe@urv.cat

Josep Domingo-Ferrer
CRISES Research Group
UNESCO Chair in Data Privacy
Dept. Computer Engineering and Mathematics
Rovira i Virgili University
Av. Països Catalans, 26
E- 43007 Tarragona.
Catalonia. Spain
josep.domingo@urv.cat

ABSTRACT

Micro-data protection is a hot topic in the field of Statistical Disclosure Control (SDC), that has gained special interest after the disclosure of 658000 queries by the AOL search engine in August 2006. Many algorithms, methods and properties have been proposed to deal with micro-data disclosure. p-Sensitive k-anonymity has been recently defined as a sophistication of k-anonymity. This new property requires that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. Like k-anonymity, the algorithm originally proposed to achieve this property was based on generalisations and suppressions; when data sets are numerical this has several data utility problems, namely turning numerical key attributes into categorical, injecting new categories, injecting missing data, and so on. In this article, we recall the foundational concepts of micro-aggregation, k-anonymity and p-sensitive k-anonymity. We show that k-anonymity and p-sensitive k-anonymity can be achieved in numerical data sets by means of micro-aggregation heuristics properly adapted to deal with this task. In addition, we present and evaluate two heuristics for p-sensitive k-anonymity which, being based on micro-aggregation, overcome most of the drawbacks resulting from the generalisation and suppression method.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Storage, Information Search and Retrieval; E.1 [**Data Structures**]: Records

General Terms

Privacy, Security, k-Anonymity, Micro-data protection, p-Sensitive k-Anonymity

1. INTRODUCTION

British politicians gasped of astonishment when they were told on November 20th, 2007, that two computer disks full of personal data of 25m British individuals had gone missing. The fate of the disks is unknown and the privacy of the individuals, whose personal data are lost, is in danger. Unfortunately, this is the latest in a series of similar nonsenses. In October, Her Majesty's Revenue and Customs (HMRC) lost another disk containing pension records of 15.000 people, and it also lost a laptop containing personal data on 400 people in September. Data on 26.5m people were stolen from the home of an employee of the Department of Veterans Affairs in America in 2006, and 658000 queries were disclosed by the AOL search engine in August of the same year. These pitfalls are not new; however, due to the great advances in the Information and Communication Technologies (ICTs), it is very easy to gather large amounts of personal data, and mistakes such as the previously explained are magnified.

There are many real-life situations in which personal data is stored: (i) Electronic commerce results in the automated collection of large amounts of consumer data. These data, which are gathered by many companies, are shared with subsidiaries and partners. (ii) Health care is a very sensitive sector with strict regulations. In the U.S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA,[8]) requires the strict regulation of protected health information for use in medical research. In most western countries, the situation is similar, (see *e.g.* [2]). (iii) Cell phones have become ubiquitous and services related to the current position of the user are growing fast. If the queries that a user submits to a location-based server are not securely managed, it could be possible to infer the consumer habits of the user [18]. (iv) The massive deployment of the Radio Frequency IDentification (RFID) technology is a reality. On the one hand, this technology will increase the efficiency of supply chains and will eventually replace bar codes. On the other hand, the existence of RFID tags in almost every object could be seen as a privacy problem [17].

In addition to the aforementioned real-life situations, most

countries have legislation which compels national statistical agencies to guarantee statistical confidentiality when they release data collected from citizens or companies; see [13] for regulations in the European Union, [14] for regulations in Canada, and [23] for regulations in the U.S. Thus, protecting individual privacy is a key issue for many institutions, namely statistical agencies, Internet companies, manufacturers, etc; and many efforts have been devoted to develop techniques guaranteeing some degree of personal privacy.

In some situations, information must be stored as it is and no modification is allowed (*e.g.* information on the taxes that a given individual should pay cannot be modified, especially when an authority must control whether the individual is really paying). In this case, data encryption and access policies seem to be the only way to protect data from being stolen. On the contrary, there exist situations in which data can be slightly altered in order to protect the privacy of data owners (*e.g.* medical data can be modified previous to their release, so that researchers are able to study the data without jeopardising the privacy of patients). In the latter case the problem is how to modify data to minimise the information loss whilst guaranteeing the privacy of the respondents.

1.1 Contribution and plan of the article

In this article we propose two heuristics for p-sensitive k-anonymity which are based on micro-aggregation. With these heuristics we overcome the problems of previous heuristics based on suppression and generalisation (*cf.* Section 2.2).

The rest of the paper is organised as follows. In Section 2 we provide the reader with some important concepts and foundational ideas. Specifically in Section 2.1 we recall some concepts on micro-aggregation, next in Section 2.2 we recall the definition of k-anonymity and we show how to achieve it by means of micro-aggregation. To conclude with Section 2, Section 2.3 provides some concepts of p-sensitive k-anonymity. Our heuristics are explained in detail in Section 3. In Section 3.1 we propose a heuristic based on the MDAV micro-aggregation method [9]. Next, in Section 3.2 we present an improvement of the previous heuristic based on the random selection of initial records. Section 4 contains some experimental results. Finally, the article concludes in Section 5.

2. BACKGROUND

The anonymity problem is not new. Many techniques and methods have been proposed to deal with this problem. In this section, we summarise some fundamental concepts of this field. First, we take a look at some basic micro-aggregation concepts. Then we show how to apply the ideas of micro-aggregation to achieve k-anonymity, and finally, we recall the definition of p-sensitive k-anonymity.

2.1 Micro-aggregation

Statistical Disclosure Control (SDC), also known as Statistical Disclosure Limitation (SDL), seeks to transform data in such a way that they can be publicly released whilst preserving data utility and statistical confidentiality, where the latter means avoiding disclosure of information that can be linked to specific individual or corporate respondent entities.

Micro-aggregation is an SDC technique consisting in the aggregation of individual data. It can be considered as an SDC sub-discipline devoted to the protection of individual data, also called micro-data. Micro-aggregation can be seen as a clustering problem with constraints on the size of the clusters. It is somehow related to other clustering problems (*e.g.* dimension reduction or minimum squares design of clusters). However, the main difference of the micro-aggregation problem is that it does not consider the number of clusters to generate or the number of dimensions to reduce, but only the minimum number of elements that are grouped in the same cluster.

When we micro-aggregate data we have to keep two goals in mind: (i)*Preserving data utility.* To do so, we should introduce as little noise as possible into the data *i.e.* we should aggregate similar elements instead of different ones. In the example given in Figure 1 for a security parameter $k = 3$, groups of three elements are built and aggregated. Note that elements in the same aggregation group are similar. (ii)*Protecting the privacy of the respondents.* Data have to be sufficiently modified to make re-identification difficult *i.e.* by increasing the number of aggregated elements, we increase data privacy. In the example given in Figure 1, after aggregating the chosen elements, it is impossible to distinguish them, so that the probability of linking any respondent is inversely proportional to the number of aggregated elements.

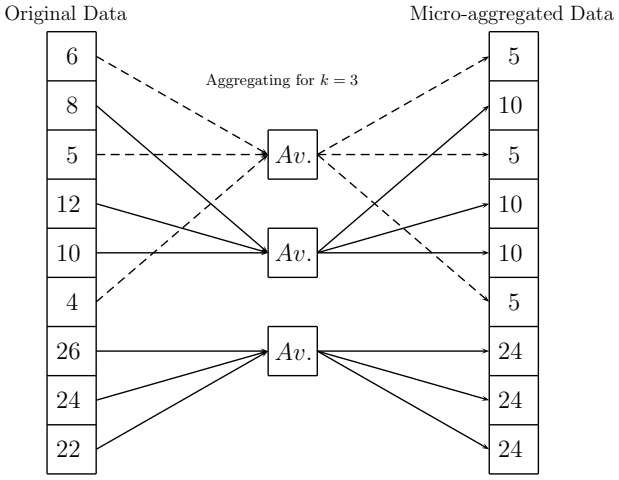

Figure 1: k-Aggregation example with $k = 3$

In order to determine whether two elements are similar, a similarity function such as the Euclidean Distance can be used. The Sum of Squared Errors (SSE) is also a common choice *cf.* Expression (1).

$$\mathbf{SSE} = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})'(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i}) \qquad (1)$$

where s is the number of subsets, n_i is the number of elements in the i-th subset, $\mathbf{x_{ij}}$ is the j-th element in the i-th subset and $\bar{\mathbf{x}}_\mathbf{i}$ is the average element of the i-th subset.

Given a homogeneity measure such as the SSE and a security parameter k, which determines the minimum cardinality of the subsets, the micro-aggregation or k-micro-aggregation problem can be enunciated as follows:

> *Given a data set* $\mathbf{D}$ *built up of* $\mathbf{n}$ *elements in a characteristic space* $\mathbb{R}^d$*, the problem consists in obtaining a* k*-partition*[1] $\mathcal{P}$ *of* $\mathbf{D}$ *so that the homogeneity of* $\mathcal{P}$ *is maximised. Once* $\mathcal{P}$ *is obtained, each element of every part of* $\mathcal{P}$ *is replaced by the average element of the part.*

This problem is known to be NP-hard [12] for multivariate data sets, so heuristic methods must be used to solve it.

2.2 k-Anonymity

k-Anonymity is an interesting approach to face the conflict between information loss and disclosure risk, suggested by Samarati and Sweeney (1998) [15, 16, 19, 20]. To recall the definition of k-anonymity, we need to enumerate the various (non-disjoint) types of attributes that can appear in a micro-data 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) [4], 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.

From these previous ideas, the concept of k-anonymity can be defined:

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 minimising 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) [15, 16, 19, 20] to achieve k-anonymity relies on suppressions and generalisations, so that minimising information loss translates to reducing the number and/or the magnitude of suppressions and generalisations.

The drawbacks of partially suppressed and coarsened data for analysis were highlighted in Domingo-Ferrer and Torra (2005) [7]:

1. Satisfying k-anonymity with minimum data modification using generalisation (recoding) and local suppression was shown to be NP-hard in Meyerson and Williams (2004) [11] and Aggarwal *et al.* (2004) [1];
2. Using global recoding for generalisation causes too much information loss, and using local recoding complicates data analysis by causing old and new categories to coexist 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. Analysing partially suppressed data usually requires specific software (imputation software, censored data analysis, etc.);
5. Last but not least, when numerical attributes are generalised, they become non-numerical.

Joint multivariate micro-aggregation (in the way of Domingo-Ferrer and Mateo-Sanz, 2002) [5] of all key attributes with minimum group size k was proposed in Domingo-Ferrer and Torra (2002) [6] as an alternative to achieve k-anonymity; besides being simpler, this alternative has the advantage of yielding complete data without any coarsening (nor categorisation in the case of numerical data). Other proposals [3, 10] generalise ordinal numerical data replacing numerical data by intervals. In the case of the k-anonymity application, micro-aggregation is performed on the projection of records on key attributes, rather than on the entire records. If the micro-aggregated 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.

2.3 p-Sensitive k-anonymity

k-Anonymity can prevent identity disclosure, *i.e.* a record in the k-anonymised 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) [22], 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*-anonymised into a group of* k *patients with* k*-anonymised 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*-anonymisation*

[1] A k-partition of $\mathbf{D}$ is a partition where its parts have, at least, k elements of $\mathbf{D}$.

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) [22] and Truta *et al.* (2007) [21] to achieve p-sensitive k-anonymity is an extension of the generalisation/suppression procedure proposed in the original k-anonymity papers. Therefore it shares the same shortcomings listed above.

We next present two different heuristics for micro-aggregation-based p-sensitive k-anonymity, where data sets have numerical quasi-identifiers and discrete confidential attributes.

3. HEURISTICS

Our aim is to obtain p-sensitive k-anonymous data sets without coarsened nor partially suppressed data. This makes their analysis and exploitation easier, with the additional advantage that numerical continuous attributes are not categorised. To do so, we propose an algorithm based on micro-aggregation (*cf.* Algorithm 1).

Our algorithm receives as input a micro-data set $\mathbf{X}$ consisting of n records having Q numerical key attributes and L discrete confidential attributes each. The result of the algorithm is a k-partition used to micro-aggregate the original micro-data set and to generate a micro-aggregated data set $\mathbf{X}'$ that fulfils the p-sensitive k-anonymity property.

The first part of the algorithm (lines 2:16) builds the initial clusters that fulfil the p-sensitive k-anonymity property. To build a cluster, a starting point x_r is selected (line 3). Depending on the selection method, we distinguish two heuristics that we discuss in the next sections. Once the initial record is selected, the algorithm looks for other records which are close to x_r and, at the same time, contribute to the p-sensitive property, *i.e.*records having different values in the confidential attributes (lines 6:9). When a group C_i fulfilling the p-sensitive property is obtained, the algorithm adds records to it until it reaches a minimum cardinality k (lines 10:13). After repeating this process several times, a set of clusters fulfilling the p-sensitive k-anonymity property is obtained. However, a number of records can remain unassigned, and they must be distributed amongst the previously created clusters (line 17:20). Finally, the algorithm micro-aggregates the original micro-data set $\mathbf{X}$ by replacing each record in $\mathbf{X}$ by the centroid of the group to which it belongs (lines 21:23).

In the next sections we show two different ways of choosing the records to start building clusters. Figure 2 shows a synthetic micro-data set having 100 records, that will be used to illustrate how the heuristics work. The data set depicted in Figure 2 consists of records having two key attributes (x and y in the Figure) and one confidential attribute (represented in the figure by means of different shapes and colours).

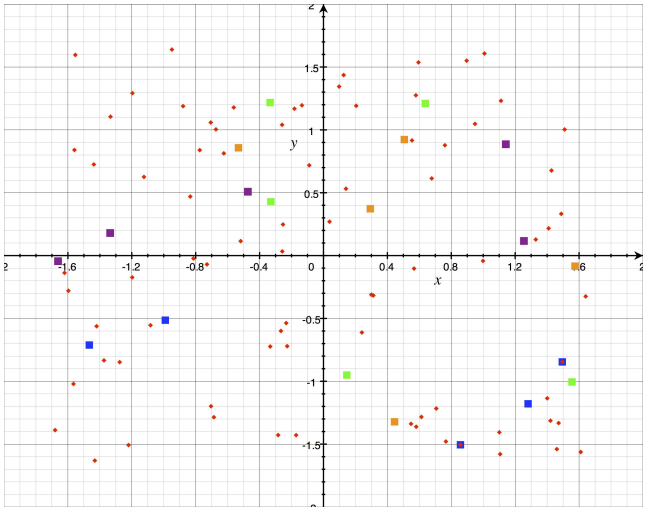

Figure 2: Running example

3.1 p-Sensitive k-anonymity with MDAV

The first way that we propose to select initial points consists in computing the average vector of the records that remain unassigned and select the record which is furthest from the average. This heuristic is inspired in the Maximum Distance to Average Vector (MDAV) micro-aggregation heuristic.

In Algorithm 3.1 we detail this process. In addition, Figure 3 shows the initial clusters generated by this heuristic. It can be observed that all clusters have at least a record which is far from the average. Although this could be good for a classic micro-aggregation algorithm, in this case, this behaviour could lead to the generation of a numerous set of unassigned records located close to the average record. This set of unassigned records must be assigned before the algorithm finishes (lines 17:20 of Algorithm 1) and due to the fact that they are all located far from the previously created cluster, the information loss associated to their assignment could be important.

Algorithm 2 Selection of initial points based on MDAV

Require: Q: the set of key attributes.
Require: UR: the set of records $\mathbf{X}$ which have not been assigned to any group yet.
Require: $x_j(Q)$: the projection of record x_j on its key attributes.

1: $\overline{x}(Q) = \text{Average}(x_1(Q), \ldots, x_n(Q), \forall x_i \in UR)$
2: $x_r = \text{ElementWithMaximumDistance}(UR, \overline{x}(Q))$
3: **return** x_r

3.2 p-Sensitive k-anonymity with Random Seeds

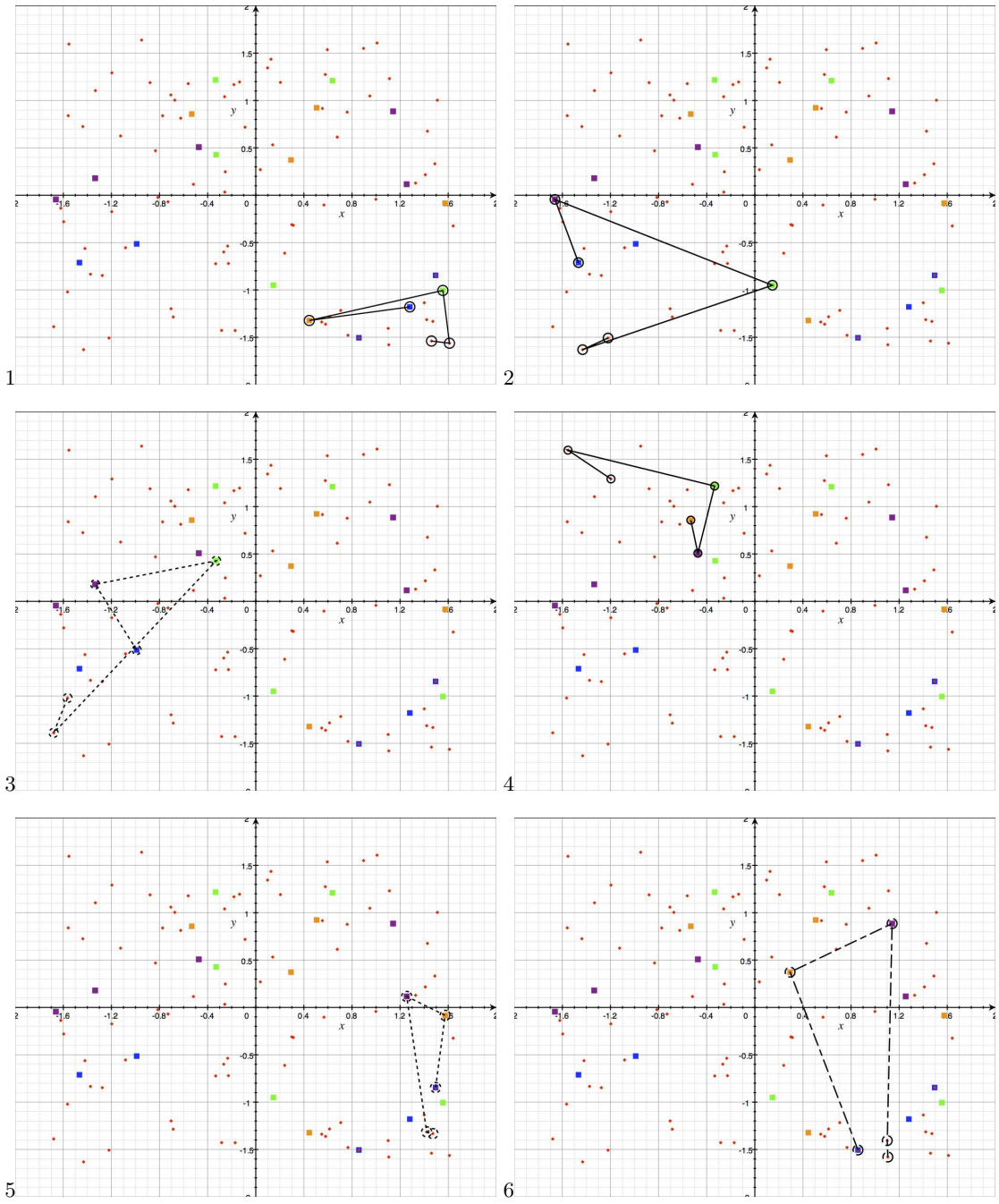

Figure 3: Initial clusters created by the MDAV-based heuristic. Graph 1 shows the first cluster, graph 2 shows the second cluster, and so on.

Algorithm 1 p-Sensitive k-anonymity micro-aggregation-based heuristic

Require: $x_1, x_2, \ldots, x_n$: the records in the original data set $\mathbf{X}$.
Require: L: the set of confidential attributes.
Require: Q: the set of key attributes.
Require: $x_j(Q)$: the projection of record x_j on its key attributes.
Require: k: the minimum number of records per group.
Require: p: the minimum number of different values for each confidential attribute in a group.
Require: P: an initially empty partition.
Require: UR: the set of records $\mathbf{X}$ which have not been assigned to any group yet.

1: $i := 0$
2: **while** (Cardinality(UR) $\geq k$ **and** UR contains at least p different values for each attribute in L) **do**
3: $x_r :=$ SelectRecordToBuildCluster();
4: $C_i :=$ newEmptyGroup();
5: $C_i :=$ AssignRecordToGroup(C_i, x_r);
6: **while** (confidential attributes of the records in C do not satisfy p-sensitivity) **do**
7: Take $x_s \in UR$ so that $x_s(Q)$ is the nearest record to $x_r(Q)$ that contributes to the compliance of p-sensitivity
8: $C_i :=$ AssignRecordToGroup(C_i, x_s);
9: **end while**
10: **while** (Cardinality(C_i) $< k$) **do**
11: $x_s :=$ ElementWithMinimumDistance(UR, x_r));
12: $C_i :=$ AssignRecordToGroup(C_i, x_s);
13: **end while**
14: $P :=$ AddGroupToPartition(C_i,P);
15: $i := i + 1$
16: **end while**
17: **for** ($\forall x \in UR$) **do**
18: $i :=$ ClosestGroup(x, P);
19: $C_i :=$ AssignRecordToGroup(C_i, x);
20: **end for**
21: **for** ($j = 1$ to n) **do**
22: $x'_j := x_j$ with $x_j(Q)$ replaced by $Centroid(C(Q))$, where C is the group in P to which x_j has been assigned.
23: **end for**
24: **return** The micro-aggregated, p-sensitive, k-anonymous data set X' formed by records $x'_1, \cdots x'_n$.

The previous heuristic fails to properly distribute the clusters amongst the complete data set. Thus, the information loss in terms of SSE grows.

In order to overcome this limitation, we propose a different scheme to select initial records which is mainly random.

Algorithm 3.2 details the proposed scheme. Figure 4 shows the clusters created by this heuristic. It can be clearly observed that several clusters are located close to the average vector. In fact, the distribution of the clusters is uniform (by construction).

As we will see in the experimental results, this distribution of clusters helps reduce the information loss.

4. EXPERIMENTAL RESULTS

With the aim of testing the proposed heuristics, we have generated a number of synthetic data sets and we have applied our heuristics on them. In addition we have used the Census data set[2] to test our heuristics with real data.

For each data set, we have measured the information loss in terms of SSE/SST, where SSE is defined in Equation (1)

[2]`http://neon.vb.cbs.nl/casc/`

Algorithm 3 Random selection of initial points

Require: Q: the set of key attributes.
Require: UR: the set of records in $\mathbf{X}$ which have not been assigned to any group yet.
Require: $x_j(Q)$: the projection of record x_j on its key attributes.

1: **for** (i:=0 to cardinality(Q)) **do**
2: $RandomVector_i := RNDUniform(MIN_{Q_i}, MAX_{Q_i})$;
3: **end for**
4: $x_r =$ ElementWithMinimumDistance(UR, $RandomVector_i$);
5: **return** x_r

and SST is the Sum of Square Errors applied over the whole data set. Moreover, we consider the improvement of the heuristic based on random selection of initial records vs. the heuristic based on MDAV. To do so we use the next expression:

$$IMP = \frac{H1 - H2}{H1} \times 100$$

where $H1$ is the SSE/SST obtained by the MDAV-based heuristic and $H2$ is the SSE/SST obtained by the random-seed-based heuristic.

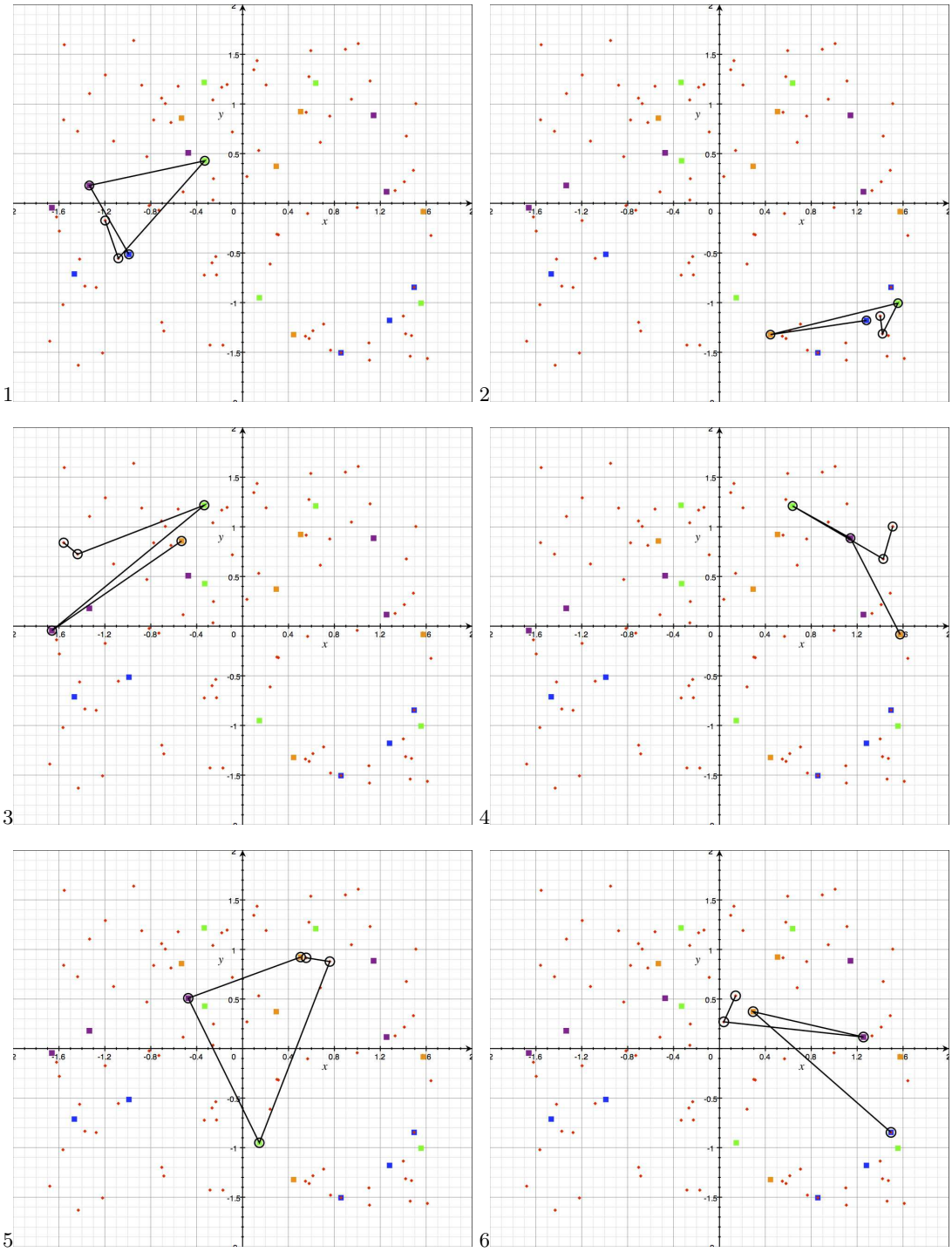

Figure 4: Initial clusters created by the Random-Seed-based heuristic. Graph 1 shows the first cluster, graph 2 shows the second cluster, and so on.

k	p	H1	H2	Improv. %
2	2	25.47	16.29	36.05
3	2	24.38	16.48	32.4
	3	30.32	22.03	27.33
4	2	20.93	17.08	18.39
	3	31.52	22.16	29.7
	4	32.72	26.19	19.95
5	2	21.59	16.5	23.55
	3	27.3	22.54	17.44
	4	34.28	26.26	23.38
	5	34.18	29.38	14.05

Table 1: Results for the Census data set

k	p	H1	H2	Improv. %
2	2	19.61	9.08	53.7
3	2	13.85	10.42	24.77
	3	31.3	17.25	44.88
4	2	12.94	9.93	23.3
	3	33.55	17.87	46.75
	4	34.26	23.57	31.2
5	2	8.35	11.06	-32.38
	3	26.62	19.11	28.2
	4	43.32	24.61	43.19
	5	38.48	34.37	10.66

Table 2: Results for the Scattered data set (100 records)

The Census data set contains 1080 records with 13 numerical attributes. 12 of these attributes have been used as key attributes and the last one has been discretised and considered as a confidential attribute. Table 1 shows the results for the Census data set.

The synthetic data sets have 100 and 1000 records. Each record has two key numerical attributes and a discretised confidential attribute (*cf.* Figure 2 for an example). They have been generated by random sampling a uniform distribution $\sim U(-10000, 10000)$. Table 2 shows the results for synthetic data set built up of 100 records, and Table 3 shows the results for the synthetic data set with 1000 records.

Due to the inherent randomness of $H2$, it cannot be assured that it always outperforms $H1$. However, from these results it can be seen that in most cases using the heuristic based on the random selection of initial records is better in terms of information loss.

5. CONCLUSIONS

p-Sensitive k-anonymity is a novel property that, when satisfied by micro-data sets, can help increase the privacy of the respondents whose data is being used.

Previous approaches to obtain micro-data sets fulfilling the p-sensitive k-anonymity property were mainly based on suppression and generalisation. In this article, we have shown how to achieve the same property by means of micro-aggregation. Specifically, we have presented two heuristics to deal with this problem.

Thanks to this novel approach, the shortcomings related to generalisation and suppressions are overcome whilst the information loss remains low.

k	p	H1	H2	Improv. %
2	2	11.21	7.29	34.95
3	2	12.49	7.43	40.53
	3	14.4	14.98	-4.07
4	2	13.85	8.31	39.99
	3	17.94	14.42	19.64
	4	31.99	24.2	24.35
5	2	14.65	7.25	50.51
	3	18.51	15.71	15.14
	4	22.46	26.37	-17.42
	5	33.2	28.14	15.24

Table 3: Results for the Scattered data set (1000 records)

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 generalisations, so the micro-aggregation approach proposed in this paper would be a novelty in all of them.

Acknowledgements

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organisation. 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.

6. REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity: Algorithms and hardness. Technical report, Stanford University, 2004.

[2] C. Boyens, R. Krishnan, and R. Padman. On privacy-preserving access to distributed heterogeneous healthcare information. In I. C. Society, editor, *Proceedings of the 37th Hawaii International Conference on System Sciences HICSS-37*, Big Island, HI., 2004.

[3] A. Campan, T. M. Truta, J. Miller, and R. Sinca. A clustering approach for achieving data privacy. In R. Stahlbock, S. F. Crone, and S. Lessmann, editors, *DMIN*, pages 321–330. CSREA Press, 2007.

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

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

[6] J. Domingo-Ferrer and V. Torra. A critique of the sensitivity rules usually employed for statistical table protection. *International Journal of Uncertainty,*

Fuzziness and Knowledge-Based Systems, 10(5):545–556, 2002.

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

[8] HIPAA. Health insurance portability and accountability act, 2004. `http://www.hhs.gov/ocr/hipaa/`.

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

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In F. Ězcan, editor, *SIGMOD Conference*, pages 49–60. ACM, 2005.

[11] A. Meyerson and R. Williams. 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, 2004. ACM.

[12] A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18(4):345–354, 2001.

[13] E. Parliament. DIRECTIVE 2002/58/EC of the European Parliament and Council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), 2002. `http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf`.

[14] C. Privacy. Canadian privacy regulations, 2005. `http://www.media-awareness.ca/english/issues/privacy/canadian_legislation_privacy.cfm`.

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

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

[17] A. Solanas, J. Domingo-Ferrer, A. Martinez-Balleste, and V. Daza. A Distributed Architecture for Scalable RFID Identification. *Computer Networks*, 51, 2007.

[18] A. Solanas and A. Martinez-Balleste. Privacy protection in location-based services through a public-key privacy homomorphism. In J. Lopez, P. Samarati, and J. Ferrer, editors, *4th European PKI Workshop. EuroPKI'07*, volume 4582, pages 362 – 368. Springer, June 2007.

[19] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.

[20] L. Sweeney. k-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[21] T. M. Truta, A. Campan, and P. Meyer. 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, 2007.

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

[23] USPrivacy. U.S. privacy regulations, 2005. `http://www.media-awareness.ca/english/issues/privacy/us_legislation_privacy.cfm`.