

# Microaggregation: Achieving $k$ -Anonymity with Quasi-Optimal Data Quality

Josep Domingo-Ferrer

April 2006

## Introduction

- ♣ Several types of attributes can be distinguished in a set of microdata (records corresponding to individuals or companies):
  - **Identifiers.** These unambiguously identify the respondent to whom the record corresponds. Identifiers are usually suppressed in microdata sets released by NSIs.
  - **Key attributes.** Those are attributes that, in combination, can be linked with external information to re-identify (some of) the respondents corresponding to (some of) the records.
  - **Confidential outcome attributes.** They contain sensitive information about the respondent (income, religion, health condition, etc.).
- ♣ NSIs must release microdata sets so as to prevent respondent re-identification or attribute disclosure.

## The SDC problem

- ♠ Statistical disclosure control (SDC) techniques are used to thwart re-identification/disclosure.
- ♠ SDC techniques either **mask key attributes or confidential outcome attributes**.
- ♠ Masking can be non-perturbative (partial suppression, generalization/recoding) or perturbative (functionally equivalent to noise addition).
- ♠ The **SDC problem** is to optimize the tradeoff between information loss and disclosure protection: the more masking, the more protection, but the greater utility loss.

## $k$ -Anonymity

- ♣  $k$ -anonymity (Samarati and Sweeney, 1998; Samarati, 2001; Sweeney, 2002) is a useful concept to solve the tension between data utility and respondent privacy.
- ♣ A protected data set satisfies  $k$ -anonymity if, for any combination of values of **key attributes** (e.g. address, age, gender, etc.), at least  $k$  records exist in the dataset sharing that combination.
- ♣ If  $V'$  is a  $k$ -anonymous data set, an intruder trying to link  $V'$  with an external non-anonymous  $V$  finds at least  $k$  matching records in  $V'$  whatever the key attributes used for linkage.
- ♣ If for a certain  $k$ ,  $k$ -anonymity is assumed to be enough protection, one can concentrate on minimizing information loss with the only constraint that  $k$ -anonymity should be satisfied.
- ♣ **Clean way to solve the tension between protection and utility!**

## Plan of this presentation

- A critique of the generalization/suppression approach to  $k$ -anonymity
- Microaggregation for  $k$ -anonymity
- An approximation heuristic to optimal microaggregation
- An example of  $k$ -anonymity through microaggregation
- Conclusion

## A critique of generalization/suppression for $k$ -anonymity

- $k$ -anonymity with minimal generalization and local suppression has been shown to be NP-hard (Meyerson et al., 2004; Aggarwal et al., 2005).
- Even how to optimally combine generalization and local suppression is an open issue (careless combination may greatly diminish utility).
- Furthermore, generalization poses several practical problems:
  1. Cost of finding the optimal recoding: for an attribute with  $c$  categories, there are  $2^c - c - 1$  possible generalizations.
  2. Determining the subset of appropriate generalizations: which are the new categories and which is the appropriate recoding between old and new categories.
  3. Example: When generalizing ZIP codes, recoding 08201 and 08205 into 0820\* makes sense only if 0820\* is meaningful as a location. For the same reason, recoding 08201 and 08205 into 0\*201 probably lacks any geographical significance. So, automatic generalization is thorny.

## Drawbacks of generalization

- Given a particular generalization rule  $c_i \rightarrow C$ , the literature diverges on which records containing  $c_i$  are recoded:
  - Global recoding** All occurrences of  $c_i$  are recoded ( $\mu$ -Argus).
  - Local recoding** Only some of the occurrences are recoded (Sweeney, 2002; Samarati 2001).
- Drawbacks of global recoding:
  1. It implies greater information loss.
  2. The recoding suitable for a set of records may be unsuitable for another set.
- Drawbacks of local recoding:
  1. It is difficult to automate.
  2. It complicates data analysis as old and new categories co-exist, and an old category can be recoded into more than one new category.

## Example: different best recodings for different records

r	×	×	×		×
s					×
t	×				×
u	×				
v	×		×	×	×
	a	b	c	d	e

## Drawbacks of local suppression

- Unknown how to optimally combine with generalization.
- Use of suppression diverges in the literature:
  - Tuple vs attribute suppression** In Sweeney (2002) entire tuples are suppressed. In  $\mu$ -Argus only particular attributes are suppressed for some records.
  - Blanking vs averaging** A suppressed value can be blanked or replaced by a locally neutral value (some sort of average).
- Whatever the suppression type, user analysis of partially suppressed data is difficult and probably requires highly specific software (dealing with censored data).

## Generalization/suppression and data types

- For categorical (nominal or ordinal) data, generalization/suppression can still be used, even though it is far from perfect for the above reasons.
- For continuous (numerical) data, generalization/suppression is unsuitable: it causes continuous data to become categorical and lose their numerical semantics.

## $k$ -Anonymity through microaggregation

- Multivariate microaggregation stands out as a natural approach to achieve  $k$ -anonymity (Domingo-Ferrer and Torra, 2005).
- Microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002) consists of two steps:
  - Partition:** The set of original records is partitioned into several groups in such a way that records in the same group are **similar** to each other and so that the number of records per group is at least  $k$ . Such a partition is called a  **$k$ -partition**.
  - Aggregation:** An aggregation operator (for example the mean for continuous data or the median for categorical data) is computed for each group and is used to replace the original records. That is, each record in a group is replaced by the group's prototype.
- In the above two steps, projections of the records on a particular set of attributes are normally used, rather than the entire records.
- A typical similarity measure to form groups is minimization of the within-groups sum of squares  $SSE$ .

## Microaggregating for *k*-anonymity

- To *k*-anonymize a dataset, microaggregate records by projecting them on the subset of key attributes.
- This is a **unified approach** to *k*-anonymity, unlike the dual generalization and suppression.
- Even if optimal microaggregation (with minimum *SSE*) is NP-hard (Oganian and Domingo-Ferrer, 2001), **efficient near-optimal heuristics exist**, unlike for generalization and suppression.
- Microaggregation **does not complicate data analysis** by adding new categories or suppressing data.
- Microaggregation is perfectly **suited to protect all data types**: numerical, ordinal and nominal.

## An approximation heuristic to optimal microaggregation

- Univariate optimal microaggregation is computable in polynomial time (Hansen and Mukherjee, 2003).
- Several heuristics have been proposed for the NP-hard problem of multivariate microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002, Sande, 2002; Laszlo and Mukherjee, 2005; Domingo-Ferrer and Torra, 2005,  $\mu$ -Argus).
- In Domingo-Ferrer, Sebé and Solanas (2005), the first approximation heuristic for multivariate microaggregation is proposed.
- Using this approximation, the  $SSE$  of the heuristic  $k$ -partition verifies

$$SSE \leq 2(2k - 1) \frac{\max(2k - 1, 3k - 5)}{2} \max(2k - 1, 3k - 5) SSE_{opt}$$

where  $SSE_{opt}$  corresponds to the optimal  $k$ -partition (with minimum  $SSE$ ).

## Rationale of the approximation heuristic

- Our approximation heuristic to optimal multivariate microaggregation is obtained by adapting the Aggarwal *et al.* (2005) approximation to optimal  $k$ -anonymization via suppression.
- The idea is to create a directed forest such that:
  1. Records are vertices;
  2. Each vertex has at most one outgoing edge;
  3.  $(u, v)$  is an edge only if  $v$  is one of the  $k - 1$  nearest neighbors of  $u$  (according to the distance function);
  4. The size of every tree in the forest, *i.e.* the number of vertices in the tree, is between  $k$  and  $2k - 1$ .

An example

## An example of $k$ -anonymity through microaggregation

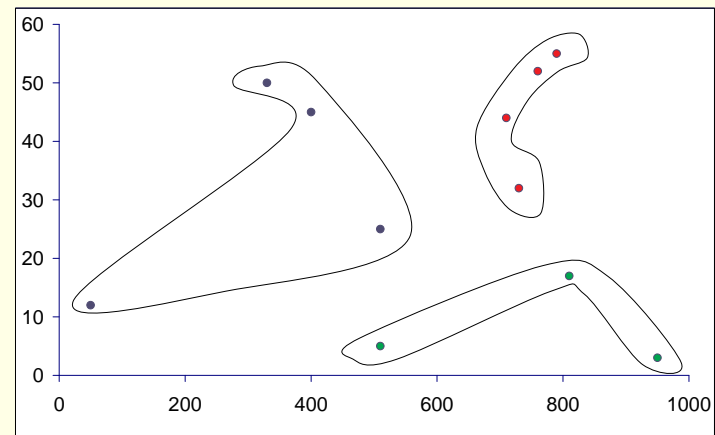
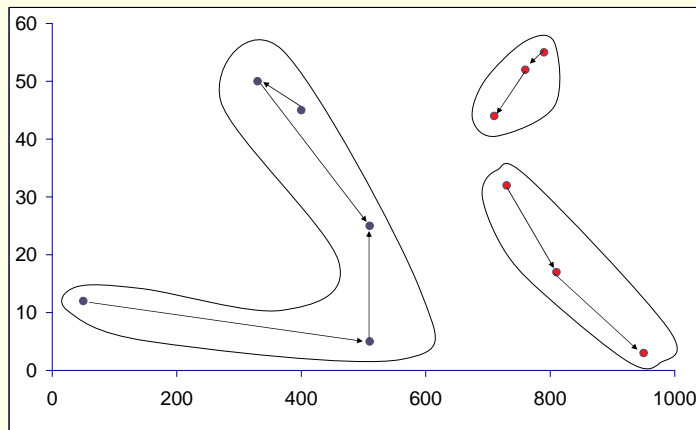
<i>Company name</i>	Surface (m <sup>2</sup> )	No. employees	Turnover (Euros)	Net profit (Euros)
A&A Ltd	790	55	3212334	313250
B&B SpA	710	44	2283340	299876
C&C Inc	730	32	1989233	200213
D&D BV	810	17	984983	143211
E&E SL	950	3	194232	51233
F&F GmbH	510	25	119332	20333
G&G AG	400	45	3012444	501233
H&H SA	330	50	4233312	777882
I&I LLC	510	5	159999	60388
J&J Co	760	52	5333442	1001233
K&K Sarl	50	12	645223	333010

An example

### 3-anonymized dataset via microaggregation with the approximation heuristic

Surface (m <sup>2</sup> )	No. employees	Turnover (Euros)	Net profit (Euros)
747.5	46	3212334	313250
747.5	46	2283340	299876
747.5	46	1989233	200213
756.67	8	984983	143211
756.67	8	194232	51233
322.5	33	119332	20333
322.5	33	3012444	501233
322.5	33	4233312	777882
756.67	8	159999	60388
747.5	46	5333442	1001233
322.5	33	645223	333010

## The heuristic 3-partition vs the optimal 3-partition



## Conclusions

- ♠  $k$ -Anonymity is a clean approach to microdata SDC.
- ♠ The drawbacks of the standard implementation of  $k$ -anonymity via generalization/suppression have been analyzed.
- ♠ Microaggregation stands out as a useful way to implement  $k$ -anonymity.
- ♠ We have sketched the first approximation heuristic to optimal multivariate microaggregation.
- ♠ A numerical example of 3-anonymity via microaggregation has been described.

## References

- G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu (2005), "Approximation algorithms for  $k$ -anonymity", *Journal of Privacy Technology*, paper no. 20051120001.
- J. Domingo-Ferrer and J. M. Mateo-Sanz (2002), "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Trans. on Knowledge and Data Engineering*, 14:1, 189-201.
- J. Domingo-Ferrer, F. Seb e and A. Solanas (2005), "A polynomial-time approximation to optimal multivariate microaggregation", manuscript.
- J. Domingo-Ferrer and V. Torra (2001), "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, eds. Doyle *et al.*, North-Holland, 111-134.
- J. Domingo-Ferrer and V. Torra (2005), "Ordinal, continuous and heterogeneous  $k$ -anonymity through microaggregation", *Data Mining and Knowledge Discovery*, 11:2, 195-212.
- S. L. Hansen and S. Mukherjee (2003), "A polynomial algorithm for optimal univariate microaggregation", *IEEE Trans. on Knowledge and Data Engineering*, 15:4, 1043-1044.
- M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation", *IEEE Trans. on Knowledge and Data Engineering*, 17:7, 902-911.
- A. Meyerson and R. Williams (2004), "On the complexity of optimal  $k$ -anonymity", in *PODS'2004*, Paris, France, 223-228.
- A. Oganian and J. Domingo-Ferrer (2001), "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of the UNECE*, 18:4, 345-354.
- P. Samarati and L. Sweeney (1998), "Protecting privacy when disclosing information:  $k$ -anonymity and its enforcement through generalization and suppression", SRI Intl. Tech. Rep.

## References

- P. Samarati (2001), "Protecting respondents' identities in microdata release", *IEEE Trans. on Knowledge and Data Engineering*, 13:6, 1010-1027.
- G. Sande (2002), "Exact and approximate methods for data directed microaggregation in one or more dimensions", *IJUFKS*, 10:5, 459-476.
- L. Sweeney (2002), "Achieving  $k$ -anonymity privacy protection using generalization and suppression", *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:5, 571-588.

