Calidad de Revistas Científicas Españolas
FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2011

ISSN: 1696-2281
eISSN: 2013–8830
Special Issue
September 2011

SORT

Statistics and Operations Research Transactions

Guest editors:
Josep Domingo-Ferrer
Vicenç Torra

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution
Spanish Region of the International Biometric Society

Generalitat
de Catalunya
Institut d'Estadística
de Catalunya

SORT Special issue: Privacy in statistical databases

Introduction to the special issue on Privacy in Statistical Databases

Privacy in statistical databases is more than a need nowadays. While the demand for detailed statistical information grows steadily, so does the legal and ethical obligation to protect the right of individuals about keeping their data away from the general public access. It makes sense that respondents are willing to participate in surveys only if there is a guarantee that their individual responses will not be disclosed and they may be worried about the access to data, which is easier than ever.

Both collectors, who spend time and effort to gather individual data, and respondents, who want to preserve their private records, have concerns about the need to keep to minimal confidentiality standards. However dealing with the protection and disclosure of statistical information poses a bunch of relevant and interesting methodological problems that have open up a fruitful area of research, which has also benefited from contributions in the field of computer science.

The current issue of SORT-Statistics and Operations Research Transactions is composed by a selection of papers by authors that participated to the PSD'2010-Privacy in Statistical Databases international conference held in Corfu (Greece) in 2010. PSD'2010 was a conference sponsored and organized by the UNESCO Chair in Data Privacy and the CONSOLIDER ARES project, with proceedings published by Springer-Verlag in Lecture Notes in Computer Science. Its purpose was to attract world-wide, high-level research in statistical database privacy. The conference was a successor to PSD 2008 (Istanbul, Sep. 24-26, 2008), PSD 2006 (Rome, Dec. 13-15, 2006) and PSD 2004 (Barcelona, June 9-11, 2004), all with proceedings published by Springer in LNCS 5262, LNCS 4302 and LNCS 3050, respectively. Those four PSD conferences follow a tradition of high-quality technical conferences on SDC which started with "Statistical Data Protection-SDP'98", held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published in Springer LNCS 2316.

The editorial committee of SORT-Statistics and Operations research Transactions has had the pleasure to invite Josep Domingo-Ferrer and Vicenç Torra to become guest editors for this special issue of this journal.

The selected articles are an extended version of those presented to the PSD'2010 conference and differ in more than 25% to the original content published in the proceedings. All featured articles have undergone the usual blind referee process and were handled by the invited editors of this special ssue.

The first article is by Philipp Bleninger, Jörg Drechsler and Gerd Ronning. It is a very interesting piece of research entitled “Remote Data Access and the Risk of Disclosure from Linear Regression”. Here the authors point out to the risk that an intruder who makes educated queries to a database can disclose sensitive individual information using a simple linear regression. The results are of interest to agencies who are about to implement security barriers and it can also be useful to determine which type of queries should be allowed.

“Coprivacy: An Introduction to the Theory and Applications of Co-operative Privacy” is the topic addressed by Josep Domingo-Ferrer in his article. He presents the concept of coprivacy or co-operative privacy to make privacy preservation attractive. After a brilliant discussion of the new theory, concepts are illustrated in P2P anonymous keyword search, in content privacy in social networks, in vehicular network communications and in controlled content distribution and digital oblivion enforcement.

Arnau Erola, Jordi CastellÓ-Roca, Guillermo Navarro-Arribas and Vicenç Torra present an article about “Using the Open Directory Project to protect query logs with semantic microaggregation”. In their work they focus on the anonymization of web search logs and indicate that existing classical methods can pose a problem of loss of utility of those logs. Their notable contribution is based on methods that are typical for statistical disclosure control, which improve data usefulness when compared to other alternatives.

The next contribution is made by Sarah Giessing and Jörg Höhne on “Eliminating Small Cells From Census Counts Tables: Empirical vs. Design Transition Probabilities”. These authors present a splendid analysis the software SAFE, that has been used in the State Statistical Institute Berlin-Brandenburg already for several years. The authors compare empirically observed transition probabilities that arise once the protection algorithm is implemented to transition matrices in the context of variants of microdata key based post-tabular random perturbation methods that have been proposed in the literature.

Jason Lucero, Michael Freiman, Lisa Singh, Jiashen You, Michael DePersio and Laura Zayatz present an article on “The Microdata Analysis System at the U.S. Census Bureau”. They show the features of a system that is under development, which will allow users to receive certain statistical analysis of Census Bureau data, such as cross-tabulations and regressions, without ever having access to the data themselves. Such analyses must satisfy several statistical confidentiality rules (including the requirement to remove some observations before the analysis is performed) and those that fail these rules will not be output to the user. Approaches to creating a system of this sort, evaluation of its effectiveness and some directions for future research are discussed in this exceptional contribution.

Finally, Anna Oganian also makes a remarkable contribution on “Multiplicative Noise for Masking Numerical Microdata with Constraints”. In her paper, she presents several multiplicative noise masking schemes that are applied by statistical agencies under the form of Statistical Disclosure Limitation (SDL) methods, which are applicable to microdata (i.e. collection of individual records) and are often called masking methods. The new schemes that are proposed by Anna Oganian are designed to preserve positivity and inequality constraints in the data together with means and covariance matrix.

I wish you enjoy the reading. Besides the topics represented in this selection of excellent articles, there is much more to be done. The methodological challenges and the increasing concern in our society about the need to protect privacy are obvious, and there is no doubt that our audience of practitioners and academics is waiting for deeper insights.

Let me finish with a sincere acknowledgment to the invited editors Josep Domingo-Ferrer and Vicenç Torra for producing this remarkable special issue of SORT-Statistics and Operations Research Transactions, and for giving our readers the opportunity to plunge into the knowledge of privacy in statistical databases.

Montserrat Guillén
Editor in Chief
Barcelona, March 24th, 2011

Remote data access and the risk of disclosure from linear regression

P. Bleninger[1], J. Drechsler[1,*], and G. Ronning[2]

[1]*Institute for Employment Research,* [2]*Tübingen University*

Abstract

In the endeavor of finding ways for easy data access for external researchers remote data access seems to be an attractive alternative to the current standard of data perturbation or restricted access only at designated data archives or research data centers. However, even if the microdata are not available directly, disclosure of sensitive information is still possible. We illustrate that an ill-intentioned user could use some commonly available background information to reveal sensitive information using simple linear regression. We demonstrate the real risks from this approach with an empirical evaluation based on a German establishment survey, the IAB Establishment Panel.

MSC: 62J68, 62P68

Keywords: Remote Data Access, Data Privacy, Disclosure, IAB Establishment Panel, Linear Regression, Artificial Outlier, Strategic Dummy.

1. Introduction

Data collecting agencies generally have two options if they are willing to provide access to their data for external researchers. They can release data sets to the public if they can guarantee that the dissemination will not harm the privacy of any survey respondent or they can allow external researchers on-site access to the data in research data centers (RDC) or data enclaves. Since most data have to be altered in some way to allow data dissemination, many researchers prefer the direct access to the unaltered data at the RDC, especially if the data dissemination requires perturbation of the microdata. For this reason more and more agencies deposit their data at data enclaves or set up their own research data centers. However, the use of these facilities comes at a high price both for

[*] Corresponding author: Jörg Drechsler, Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg; e-mail: joerg.drechsler@iab.de

Received: November 2010
Accepted: March 2011

the researcher and the providing agency. Researchers have to travel to the agency before they ever get in touch with the original data. Although some agencies provide dummy data sets to give the researcher an idea of the real data, these dummy data sets often are of very low quality and the researcher might not realize that the data collected by the agency is not suitable for her analysis before traveling to the agency. Furthermore, researchers can request a certain time slot at the RDC in which they expect to finish their research. It is very difficult for the researcher to anticipate how long the data preparation will take without access to the data, and unexpected problems might require more days than the admitted time slot will allow. Besides, if the researcher wants to extend her research maybe using more variables than she asked for in the original proposal, she might have to go through the complete reviewing process again before she can actually add the variables to her analysis. On the other hand, the agency has to check every output from the analysis for potential disclosure violations. Only cleared outputs may leave the RDC and may be used by the researcher for publication. At present, this output checkin is still carried out manually. With the growing popularity of the RDCs the capacity of handling all this output checking is at the limit.

Given these drawbacks remote data access seems to be the panacea for data access for external researchers. In an ideal world full remote access would enable the external researcher to connect to a host server from her desktop machine. She would see the microdata on the screen and would be allowed to manipulate them in any way but the actual data would never leave the server and it would not be possible to store the microdata on the desktop computer. Requested queries would be automatically scanned for possible confidentiality violations and only those queries that pass the confidentiality check would be answered by the server. Remote access would free the researcher from the burden of traveling to the RDC and it would render the cost intensive and time consuming manual output checking unnecessary. However, there are many obstacles with this approach making the full implementation of a remote data access more than questionable. Apart from the technical issues of guaranteeing a safe connection between the desktop computer of the external user and the microdata server at the agency, direct access to the unchanged microdata is prohibited by law in many countries. For example in Germany, the data accessible for external researchers is required to be *de facto* anonymised which means that the effort that is necessary to identify a single unit in the data set is higher than the actual benefit the potential intruder would achieve by this identification. This is still a privilege compared to the *absolute* anonymity that is required for all published results. One solution in this context could be that the researcher would only see an anonymised version of the microdata on her screen but the queries she submits to the server would actually be run on the original data. However, this would still require the server to identify all queries that might lead to a breach of confidentiality.

Some of these queries are easy to identify. For example queries that ask for the maximum or minimum of a variable should never be allowed. For tabulation queries potentially identifying small cells could be suppressed using standard rules from the cell suppression literature.[1] However, there are other analyses for which it is not that obvious that they actually might impose an increased risk of disclosure and illustrating this for a specific set of queries is the main aim of this paper.

We focus on the risks from simple linear regression analysis under the assumption that the user will never see the true microdata. Given the legal restrictions in many countries (see discussion above), we believe that even under remote access the user will only see an anonymised version of the true microdata. In this sense our notion of remote access is located somewhere in the middle between the dream of a full remote access and the idea of a remote analysis server that can only answer specified queries without providing access to any microdata at all. We note that our findings are also relevant in the context of a plain remote analysis server.

Often regression analysis is considered as safe in the sense that it is assumed that no output checking is required. Following the discussion in Gomatam *et al.* (2005) we illustrate that an intruder with background knowledge on some of the variables contained in the data set can get accurate estimates for any sensitive variable she is interested in using only the results from a linear regression analysis. We use the IAB Establishment Panel to demonstrate empirically that at least for business data very limited and easily available background information can be sufficient to allow the intruder to obtain sensitive information with this approach.

The remainder of the paper is organized as follows. Section 2 recapitulates the basic concept that allows the intruder to retrieve sensitive information for a single respondent based on the background information she has about that respondent. In this section we follow the outline described in Gomatam *et al.* (2005). In Section 3 we briefly introduce the data set we used for the empirical simulations: the IAB Establishment Panel. This data set is used in Section 4 to illustrate that only very limited background information is required to learn sensitive information about a survey respondent in this setting. The paper concludes with some final remarks.

2. The formal approach

In the following we assume that the intruder has at least approximate knowledge about some of the variables contained in the survey for a certain survey respondent m. It is important to note that this knowledge may refer to any set of variables in the data set, no matter if the variables are sensitive or not. For example in a business survey, the external information available to the intruder might be the energy consumption or the total production time. The intruder would then use these variables for obtaining information on sensitive variables such as investment, sales, or research expenditures.

[1] Even cell suppression can quickly become problematic, if we allow dynamic queries. In this case, the server would have to keep track of all earlier queries and would have to guarantee that requests submitted at a later point in time would not allow the calculation of cell entries that are being suppressed now.

In the following we denote the variable for which information is at hand by x and the true value for this variable provided by the survey respondent m by x_m^0. Let $\hat{x}_m$ be the external information the intruder obtained about the survey respondent m for this variable. Finally, let y_m be the reported value for respondent m for the sensitive variable of interest y.

Gomatam *et al.* (2005) pointed out that the knowledge of $\hat{x}_m$ may be used to obtain information for any other variable contained in the microdata set for this respondent by making the variable of interest the dependent variable in a simple linear regression analysis. The authors propose two approaches: (i) The intruder could generate an "artificial outlier" obtained by transformation. (ii) Alternatively, the intruder could employ a "strategic dummy variable" which uses the background information for identifying the respondent m.

2.1. Artificial outliers

For the artificial outlier approach we assume the intruder knows the exact reported value for x_m, that is $\hat{x}_m = x_m^0$. She defines a new regressor variable

$$z = \frac{1}{|x - \hat{x}_m| + \varepsilon} \tag{1}$$

where ε is arbitrarily small. If we include this regressor variable in a linear regression with the variable of interest specified as the dependent variable, the regressor z will become extremely large for the respondent m and therefore generates a leverage point such that the predicted value of the dependent variable tends towards the true value y_m^0 for this respondent. A formal proof that

$$\lim_{z_m \to \infty} \hat{y}_m \quad = \quad y_m$$

holds, is given in Appendix A1. It is important to note that this is true only if no other respondent reports a value for x that is equal to x_m^0. If other respondents report the same value, y_m will generally not be predicted exactly (see Appendix A1 for details).

2.2. Strategic dummies

Alternatively, the intruder could define a dummy that exploits the knowledge regarding the variable x. In case of exact knowledge of the reported value the dummy would be given by

$$\mathfrak{I}_{x=x_m} \quad = \begin{cases} 1 & \text{if } x = \hat{x}_m \\ 0 & \text{else.} \end{cases} \tag{2}$$

In other situations only vague information might be available represented by an interval in which the true value x_m^0 must fall. This range might be formulated in additive or multiplicative terms, that is

$$x_m^0 - \gamma < \hat{x}_m < x_m^0 + \gamma \quad \text{or} \quad (1-\delta)x_m^0 < \hat{x}_m < (1+\delta)x_m^0.$$

Thus, assuming only approximate knowledge one would create a strategic dummy according to

$$\Im_{x \simeq x_m} \quad = \begin{cases} 1 & \text{if } x - \gamma < \hat{x}_m < x + \gamma \\ 0 & \text{else} \end{cases} \tag{3}$$

or the corresponding multiplicative specification mentioned above.
It is shown in Appendix A.2 that a simple regression which uses just this dummy variable and any variable of interest as the dependent variable will result in

$$\hat{y}_m \; = \; y_m^0.$$

The result remains valid if other regressors are added to the model (see Appendix A.2).

However, the proof again is based on the assumption that only a single respondent is identified using the knowledge regarding x. If x is a categorical variable, this is an unrealistic assumption and even for continuous variables more than one respondent may report the same value. Still, with the dummy variable approach the constructed dummy can easily be based on more than one variable exploiting all the information the intruder has about the survey respondent. In our business survey example this could mean that the intruder uses her information about the industry, an approximate number of employees, and regional information about the establishment she is looking for. In this case we could define an indicator dummy for each variable for which the intruder has background information.

Let $x_1, \ldots, x_p$ be the variables for which background information is available and let $\Im_1, \ldots \Im_p$ be the corresponding indicators defined as in (2) or (3). Now the final indicator can be defined as follows:

$$\Im \quad = \begin{cases} 1 & \text{if } \Im_1 = 1 \wedge \Im_2 = 1 \wedge \cdots \wedge \Im_p = 1 \\ 0 & \text{else.} \end{cases} \tag{4}$$

It is important to note that both the artificial and the strategic dummy approach critically rely on the assumption that a single record can be identified with the external information the intruder has about m. However, the artificial outlier approach requires that the intruder knows x_m^0 exactly. This is often unrealistic in reality. With the dummy variable approach it can be sufficient to have a rough estimate of x_m^0.

3. The IAB Establishment Panel

Since our empirical evaluations in the next section are based on the wave 2007 of the IAB Establishment Panel a short introduction of the data set should prelude our illustrations. The IAB Establishment Panel is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only include employees covered by social security – civil servants and unpaid family workers for example are not included – approx. 80% of the German workforce are represented. However, the degree of coverage varies considerably across the occupations and the industries.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry.

These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually – since 1996 with over 4,700 establishments in East Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. For a detailed description of the data set we refer to Fischer *et al.* (2008) or Kölling (2000). For the simulations we use one data set with all missing values imputed. We treat all imputed values like originally observed values for simplicity. See Drechsler (2010) for a description of the multiple imputation of the missing values in the survey.

4. Empirical evidence

For our empirical evaluations, we use the wave 2007 of the establishment survey and treat the turnover of an establishment as the sensitive variable to be disclosed. Thus, we exclude all entities from the survey that do not report turnover such as non-industrial organizations, regional and local authorities and administrations, financial institutions, and insurance companies. The remaining data set includes 12,814 completely observed establishments. We analyze different subsets of the dataset defined by quantiles of the establishment size to illustrate the increased risk for larger establishments.

Table 1: *Disclosure risk evaluations using an artificial outlier generated from establishment size.*

	quantile	N	prop. uniqu. identified	Δ all identified	Δ uniqu. identified
size	all	12814	0.034	13773.795	0.001
	0.5	6516	0.066	26936.156	0.001
	0.75	3217	0.134	51952.053	0.001
	0.9	1282	0.335	1.957	0.001
	0.99	129	0.969	0.011	0.0001

4.1. Empirical evidence for artificial outliers

Using the number of employees as the available background information we construct a variable z according to (1) setting $\epsilon = 0.0001$. To evaluate the risks for the complete data set we successively treat each record in the data set as the target m for which background information is available. Table 1 summarizes the results of the artificial outlier regressions for different subsets of the data. The first column defines the subset of the data. For example, the results for the 90% quantile represent only the largest 10% of establishments. The second column provides the number of records that are contained in the subset. Column 3 contains the percentage of records that are uniquely identified based on an artificial outlier derived from the establishment size, i. e. it contains the percentage of unique high leverage points regarding the number of employees. If there is more than one high leverage point, additional establishments reduce the prediction accuracy for the target's turnover (see the proof in Appendix A.1). Column 4 presents the average absolute relative error between the predicted and the observed value for turnover for the target record m, i.e.

$$\Delta = \frac{1}{N}\sum_{j=1}^{N}\frac{|\hat{y}_{m=j} - y_{m=j}|}{y_{m=j}} \tag{5}$$

for all records in the subset. Finally, column 5 presents the same quantity only for the records that are uniquely identified and therefore generate a unique high leverage point for z.

As expected the disclosure risk clearly increases with establishment size. Under the assumption that the intruder would know the exact reported establishment size, we observe a substantial increase in the risk when going from the largest 10% of the establishments (33.5% correctly identified) to the largest 1% of establishments (96.9% correctly identified). Below these thresholds identification risks are relatively low since establishment size alone will not uniquely identify a single record. The results in column 4 illustrate that generally risks are low as long as a unique identification is not possible. The average absolute relative error is very large (often far more than 100%) indicating that the predicted value on average differs substantially from the reported value. Finally, all the values close to zero in the last column are by no means surprising. This is a direct result of the proof given in Appendix A.1. We only include these results to emphasize that once a record is uniquely identified, the intruder does not have to have direct access to the microdata. Instead she can use the artificial outlier approach (or the dummy variable approach discussed below) to exactly reveal any sensitive information about the identified record.

Often the intruder will have more background information on the target than just one variable. Generally she can use this information to generate more artificial outliers and also include them in the regression. For brevity we omit the proof that an exact prediction is possible with more than one outlier variable. A detailed proof can be found in Ronning *et al.* (2010).

However, simply using two outlier variables in the regression will not necessarily increase the number of uniquely identified records. The proof only holds if both outliers individually identify the same single record uniquely. This means that in general there is no benefit from adding a second artificial outlier to the regression since the dependent variable will only be predicted correctly for those units for which one of the background variables alone already uniquely identifies the target. The same results would be achievable if the intruder would run two separate regressions using one outlier at a time. To fully utilize the additional background knowledge the intruder should interact the background variables and apply the artificial outlier approach to the interaction term. If the joint background information identifies a record uniquely, the value of the interaction term will also be a unique value in the data set.

We illustrate the increased risks if the intruder has information on more than one variable in Table 2. We assume the intruder knows the exact number of employees and the German Federal State in which the establishment is located and uses the interaction of the two variables to generate the artificial outlier.

As expected the disclosure risks increase considerably. For example 15.9% (87.2%) of the establishments in the complete data set (of the largest 10% of the establishments) are identified uniquely compared to only 3.4% (33.5%) if the establishment size is used alone to identify the target. In theory the intruder could further improve her results if more background information is available. The more variables are interacted to generate the artificial outlier the higher is the chance of a unique identification and thus a perfect

Table 2: *Disclosure risk evaluations using artificial outliers generated from the interaction term of establishment size and German Federal State.*

	quantile	N	prop. uniqu.	Δ all	Δ uniqu.
			identified	records	identified
size*fed. state	all	12814	0.159	17820.512	0.035
	0.5	6516	0.312	33908.516	0.036
	0.75	3217	0.596	63907.225	0.006
	0.9	1282	0.872	0.338	0.0003
	0.99	129	1	$1.86 * 10^{-5}$	$1.86 * 10^{-5}$

prediction. However, a regression using three-way, four-way or even higher interaction terms will look very suspicious or might not be allowed in a remote access setting.

4.2. Empirical evidence for strategic dummies

For the strategic dummy approach we evaluate for each record if a unique identification is possible using a varying amount of background information. For the background information we chose four variables that we believe are easy to obtain for an intruder from public records, namely the (approximate) size of the establishment, i.e. its (approximate) total number of employees, the German Federal State the establishment is located in, its legal form and its industrial sector (recorded in 40 categories). We evaluate the increase in risk if these variables are added successively to the strategic dummy. The results are summarized in the Table 3. Not surprisingly the same percentage of records as in Ta-

Table 3: *Disclosure risk evaluations using the strategic dummy approach.*

quantile	N	indicators $\Im_k$	prop. uniqu. identified	Δ all records	Δ uniqu. identified
all	12814	exact size	0.034	13801.825	0
		approx. size	0.0009	11025.450	0
		+ federal state	0.023	11739.574	0
		+ legal form	0.116	13633.345	0
		+ branch	0.658	1.478	0
0.5	6516	exact size	0.066	26985.871	0
		approx. size	0.002	21526.008	0
		+ federal state	0.046	21945.190	0
		+ legal form	0.200	26774.728	0
		+ branch	0.846	0.323	0
0.75	3217	exact size	0.134	52023.417	0
		approx. size	0.003	40965.983	0
		+ federal state	0.085	39651.427	0
		+ legal form	0.228	48390.483	0
		+ branch	0.868	0.147	0
0.9	1282	exact size	0.335	1.956	0
		approx. size	0.009	4.296	0
		+ federal state	0.186	1.944	0
		+ legal form	0.352	1.499	0
		+ branch	0.895	0.070	0
0.99	129	exact size	0.969	0.011	0
		approx. size	0.085	1.311	0
		+ federal state	0.682	0.136	0
		+ legal form	0.806	0.055	0
		+ branch	0.953	0.021	0

ble 1 are identified, if the exact establishment size is used as a dummy. Relaxing the unrealistic assumption of exactly knowing the size of the establishment we use an indicator for the approximate total number of employees that identifies all records that lie within ±2.5% of the reported establishment size. This information alone almost never uniquely identifies a record in the data set. Even for the top 0.1% of establishments only 31% are uniquely identified. However, adding more information significantly increases the risk. When all four background variables are used, more than 65% of the establishments are identified uniquely in the entire data set. Since arguably intruders will only be interested in the larger establishments and not in small family businesses, the fact that almost 90% of the records can be uniquely identified for the largest 10% of the establishments based on very little background information is an alarming result. Again, we only include the results in the last column of the table to emphasize that once a record is uniquely identified all information in the data set for that record can be revealed easily without access to the actual microdata.

This leads to the question how the intruder will know that she has indeed uniquely identified the mth respondent. Of course, the natural way would be to check the residuals of the regression for zeroes. However, residuals usually are not reported in remote access. Alternatively, for the dummy variable approach the intruder could check the mean of the generated dummy variable which should be $1/n$ in case of unique identification. If the agency decides to suppress means for binary variables with few positive (or negative) outcomes, the intruder could compute the variance of the dummy variable. Given a unique identification it should be equal to $Var(\mathfrak{I}) = 1/n + 1/n^2$. Both approaches are of course not possible when generating an artificial outlier since z would just be a new continuous variable with unknown mean and variance. In this case, the intruder might check, if a unique maximum exists for z. Only if the maximum is unique, a single record has been identified. However, such requests will likely be suppressed by the remote server. This can be seen as an additional argument in favor of the strategic dummy approach.

5. Conclusion

It is obvious that agencies – once they are aware of the risks described in the previous sections – can easily prevent this type of disclosure, e.g. by prohibiting regressions that contain dichotomous regressors with less than say 3 positive outcomes or by allowing only certain transformations for the variables. But it is important that the agency must be aware of the problem to prevent it. The point that we are trying to make is that there are many constellations that might lead to a risk of disclosure. Some are obvious whereas others are more difficult to detect in advance. Full remote access without any intervention of the agency would require that all possible constellations are considered and ruled out before data access is provided. The risk from linear regressions that is the main topic of this paper is only one example of a disclosure risk that might not be obvious at first glance. We believe there are many other situations that might be equally harmful. For example it is well known that saturated models can reveal the exact information for small cell table entries that would have been protected by cell suppression or any other statistical disclosure limitation technique if the table would have been requested directly. We believe that more research in the area is needed to detect other user queries that might impose a risk of disclosure. Whether it will be possible to rule out all potential disclosure risks in advance remains an open question.

Acknowledgments

This research was supported by a grant from the German Federal Ministry of Education and Research.

References

Drechsler, J. (2010). Multiple imputation in practice – a case study using a complex German establishment survey. *Advances in Statistical Analysis (online first)*.

Fischer, G., Janik, F., Müller, D. and Schmucker, A. (2008). The IAB Establishment Panel – from sample to survey to projection. Technical report, FDZ-Methodenreport, No. 1.

Gomatam, S., Karr, A. F., Reiter, J. P. and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science*, 20, 163–177.

Hoaglin, D. and Welsh, R. (1978). The hat matrix in regression and anova. *The American Statistician*, 32, 17–22.

Kölling, A. (2000). The IAB-Establishment Panel. *Journal of Applied Social Science Studies*, 120, 291–300.

Ronning, G., Bleninger, P., Drechsler, J. and Gürke, C. (2010). Remote Access - Eine Welt ohne Mikrodaten? (in German). *IAW Discussion Papers*, 66.

A. Artificial outliers and strategic dummies

We consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{6}$$

where $\mathbf{y}$ and $\mathbf{u}$ are n-dimensional vectors, $\boldsymbol{\beta}$ is a K-dimensional vector and $\mathbf{X}$ a $(n \times K)$ matrix with $\boldsymbol{\iota}' = (1, 1, \ldots, 1)$ as the first column. The vector of predicted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{7}$$

where **H**, called the hat matrix, measures the “leverage” of a certain regressor (see, e.g. Hoaglin and Welsh (1978)).

A.1 Artificial outliers

In the following we assume that the observations are ordered such that observations for survey respondent m are in the first row of the data matrix. Therefore z_1 contains the artificial outlier which tends towards infinity; compare the definition (1) of artificial outliers in the main text.

Unique identification

In the special case of a simple regression ($K = 2$) with

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \mathbf{z} \end{pmatrix}$$

the elements of the hat matrix are given by

$$h_{jk} = \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \left(\sum_{i=1}^{n} z_i^2 - z_j \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_j z_k \right),$$

with $j = 1, \ldots, n$ and $k = 1, \ldots, n$. Therefore the jth element of the vector of predicted values $\hat{\mathbf{y}}$ is given by

$$\hat{y}_j = \sum_{k=1}^{n} h_{jk} y_k = \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \sum_{k=1}^{n} \left(\sum_{i=1}^{n} z_i^2 - z_j \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_j z_k \right) y_k$$

and in particular for $j = 1$ we have

$$\begin{aligned}
\hat{y}_1 &= \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \sum_{k=1}^{n} \left[\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_1 z_k \right] y_k \\
&= \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \left[\left(\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i \right) \sum_{k=1}^{n} y_k - \left(\sum_{i=1}^{n} z_i - n z_1 \right) \sum_{k=1}^{n} z_k y_k \right] \\
&= \frac{\left(\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i\right) \sum_{k=1}^{n} y_k}{n\sum z_i^2 - (\sum z_i)^2} - \frac{\left(\sum_{i=1}^{n} z_i - n z_1\right) \sum_{k=1}^{n} z_k y_k}{n\sum z_i^2 - (\sum z_i)^2} \\
&= \frac{\left(z_1^2 + \sum_{i>1} z_i^2 - z_1 (z_1 + \sum_{i>1} z_i)\right) \sum_{k=1}^{n} y_k}{n(z_1^2 + \sum_{i>1} z_i^2) - (z_1 + \sum_{i>1} z_i)^2} - \frac{(z_1 + \sum_{i>1} z_i - n z_1)(z_1 y_1 + \sum_{k>1} z_k y_k)}{n(z_1^2 + \sum_{i>1} z_i^2) - (z_1 + \sum_{i>1} z_i)^2} \\
&= A - B.
\end{aligned}$$

In order to obtain results for $z_1 \to \infty$ we write the two terms as follows:

$$A = \frac{\left[\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)\right] \sum_{k=1}^{n} y_k}{n\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

and

$$B = \frac{\left(1+\frac{\sum_{i>1} z_i}{z_1} - n\right)\left(y_1 + \frac{\sum_{k>1} z_k y_k}{z_1}\right)}{n\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

from which we obtain

$$\lim_{z_1\to\infty} \hat{y}_1 \;=\; \lim_{z_1\to\infty} (A-B) \;=\; \frac{0}{n-1} - \frac{(1-n)y_1}{n-1} \;=\; y_1. \tag{8}$$

Therefore for a sufficiently large z_1 we can approximate y_1 by its predicted value $\hat{y}_1$.

Non-unique identification

To this point we assumed that the target is uniquely identified by the background information resp. the transformed outlier generating variable (see (1) in the main text). Now consider the case where more than a single subject is identified by x_m resp. z. In this case the matrix containing the outlier is given by

$$\mathbf{X}_2 = \begin{pmatrix} z_1 \boldsymbol{\iota}_q \\ \mathbf{z}_2 \end{pmatrix}.$$

We assume q subjects are identified, i.e. have the exact same value for the background variable as the target record. These q subjects are transformed to artificial outliers. Without loss of generality let them be the first q observations in the dataset. $\boldsymbol{\iota}_q$ is a q-vector of ones and $\mathbf{0}$ a $(n-q)$-vector of zeros so that

$$\mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' = \frac{1}{qz_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}_2' \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}_q' & \mathbf{z}_2\mathbf{z}_2' \end{pmatrix}$$

and

$$\mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \frac{1}{qz_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}_2' \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}_q' & \mathbf{z}_2\mathbf{z}_2' \end{pmatrix} \left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right).$$

The predicted values are given by

$$\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \frac{1}{qz_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}_2' \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}_q' & \mathbf{z}_2\mathbf{z}_2' \end{pmatrix} \left(\mathbf{y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

respectively

$$\begin{pmatrix} \hat{\mathbf{y}}_q \\ \hat{\mathbf{y}}_{n-q} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \\ z_1 \mathbf{z}_2 \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\}$$

For the first q elements of the vector $\hat{\mathbf{y}}$ of predicted values we get

$$\begin{aligned}
\hat{\mathbf{y}}_q = \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 &+ \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\} \\
&= \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \\
&+ \quad \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \left(\mathbf{y}_q - \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1\right) \\
&+ \quad \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \left(\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q} \hat{\boldsymbol{\beta}}_1\right) \\
&= \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \\
&+ \quad \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}} \; \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \left(\mathbf{y}_q - \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1\right) \\
&+ \quad \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}} \; \frac{1}{z_1} \boldsymbol{\iota}_q \mathbf{z}_2' \left(\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q} \hat{\boldsymbol{\beta}}_1\right) \qquad (9)
\end{aligned}$$

If z_1 becomes infinitely large the limit of the predicted values is

$$\lim_{z_1 \to \infty} \hat{\mathbf{y}}_q = \frac{1}{q} \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \; \mathbf{y}_q + \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \; - \; \frac{1}{q} \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \qquad (10)$$

resulting in

$$\hat{y}_i = \overline{y}_q \; + \left(\; 1, \quad x_{i2} - \overline{x}_q^{(2)}, \quad \ldots, \quad x_{iK} - \overline{x}_q^{(K)} \; \right) \hat{\boldsymbol{\beta}}_1 \;, i = 1, 2, \ldots, q. \qquad (11)$$

Here we use

$$\overline{y}_q = \frac{1}{q} \sum_{i=1}^{q} y_i \quad \text{and} \quad \overline{x}_q^{(k)} = \frac{1}{q} \sum_{i=1}^{q} x_{ik} \,, k = 2, \ldots, K \,.$$

Both (10) and (11) show that

- if $q = 1$, i.e. unique identification, the result reduces to (8) because $\bar{y}_q = y_1$ and $\bar{x}_q^{(k)} = x_{1k}$ for all regressors.
- If only the artificial outlier generating z is used in a simple linear regression it holds that

  $$\lim_{z_1 \to \infty} \hat{y}_i \quad = \quad \bar{y}_q \,, \quad i = 1, \ldots, q,$$

  for all q subjects selected.
- In general however, under non-unique identification no clear-cut statement regarding the difference between $\hat{y}_i$ and y_i , $i = 1, 2, \ldots, q$, can be made.

A.2 Strategic dummy variables

Simple regression

In case of unique identification by (2), (3) or (4) in the main text the regressor matrix is given by

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \mathbf{e}_1 \end{pmatrix},$$

where $\mathbf{e}_1$ is an n-dimensional vector with 1 as the first element and 0 for the remaining $n-1$ elements. Therefore

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{(n-1)} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix}$$

and

$$\mathbf{H} = \frac{1}{n-1} \begin{pmatrix} n-1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\iota}_{n-1}\boldsymbol{\iota}'_{n-1} \end{pmatrix},$$

where $\mathbf{0}$ is the $(n-1)$-dimensional null vector and $\boldsymbol{\iota}_{n-1}$ a $(n-1)$-dimensional vector of ones. Note that $h_{11} = 1$ and $h_{1j} = 0,\ j > 1,$ so that the predicted value for y_1 is given by

$$\hat{y}_1 = \sum_{k=1}^{n} h_{1k} y_k = \frac{1}{n-1} \left((n-1) y_1 + \sum_{k>1} 0 \cdot y_k \right) = y_1.$$

The case of additional regressors

We now consider the case that other regressors are added to the regression which might be motivated by the idea that the use of a strategic dummy is not so easily detected by the agency if other regressors are also included in the model. We write the model in partitioned form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \begin{pmatrix} \mathbf{X}_1 \; \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}.$$

with

$$\mathbf{X}_2 = \mathbf{e}_1$$

so that this submatrix contains only the information regarding the strategic dummy. Then the vector of predicted values can be written as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 \; + \; \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 \; + \; \mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) \end{aligned} \tag{12}$$

Since

$$\mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) \\ &= \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1, \end{aligned}$$

we obtain for the vector of predicted values in (12):

$$\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 \; + \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \tag{13}$$

and in particular for the first element we get

$$\hat{y}_1 = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + y_1 - \begin{pmatrix} 1 & x_{12} & \ldots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 = y_1. \tag{14}$$

Non-unique identification

The empirical example in Section 4 shows that y_m and $\hat{y}_m$ may differ substantially if more than one respondent is identified using the background information available for x_m. In this section we evaluate the fitted value $\hat{y}_m$ in this case.

If more than one respondent is picked by the strategic dummy the submatrix $\mathbf{X}_2$ (which actually is a vector) has the form

$$\mathbf{X}_2 = \begin{pmatrix} \boldsymbol{\iota}_q \\ \mathbf{0} \end{pmatrix}$$

where we assume that q units in the data set have the same reported value for the available background information as the target record x_m and that they are placed in the first q rows of the data matrix. $\boldsymbol{\iota}_q$ is a vector of ones and $\mathbf{0}$ denotes a $n-q$ dimensional vector of zeroes. Moreover, we have

$$\mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' = \frac{1}{q}\begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{X}_2\,(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) &= \frac{1}{q}\begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) \\ &= \begin{pmatrix} \bar{y}_q \\ \bar{y}_q \\ \vdots \\ \bar{y}_q \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & \bar{x}_q^{(2)} & \ldots & \bar{x}_q^{(K)} \\ 1 & \bar{x}_q^{(2)} & \ldots & \bar{x}_q^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x}_q^{(2)} & \ldots & \bar{x}_q^{(K)} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1, \end{aligned}$$

where we use

$$\bar{y}_q = \frac{1}{q}\sum_{i=1}^{q} y_i \quad \text{and} \quad \bar{x}_q^{(k)} = \frac{1}{q}\sum_{i=1}^{q} x_{ik}\,, k = 2, \ldots, K.$$

Comparing this with (9) we note that for the first q elements of the vector $\hat{\mathbf{y}}$ we obtain

$$\hat{y}_i = \bar{y}_q + \left(\begin{array}{cccc} 1, & x_{i2} - \bar{x}_q^{(2)}, & \ldots, & x_{iK} - \bar{x}_q^{(K)} \end{array} \right) \hat{\boldsymbol{\beta}}_1 \, , i = 1, 2, \ldots, q. \tag{15}$$

which implies the following: (i) If $q = 1$ and therefore a single unit is identified, the above result is equivalent with (14) because then $\bar{y}_q = y_1$ and for all regressors $\bar{x}_q^{(k)} = x_{1k}$. (ii) If the strategic dummy is used as a single regressor then for all q units

$$\hat{y}_i = \bar{y}_q$$

holds, that is, the estimated value of y equals the arithmetic mean of all q units. (iii) If more regressors are added to the model, no clear-cut statement regarding the difference between y_m and $\hat{y}_m$ can be made.

Coprivacy: an introduction to the theory and applications of co-operative privacy

Josep Domingo-Ferrer[*]

Universitat Rovira i Virgili

Abstract

We introduce the novel concept of coprivacy or co-operative privacy to make privacy preservation attractive. A protocol is coprivate if the best option for a player to preserve her privacy is to help another player in preserving his privacy. Coprivacy makes an individual's privacy preservation a goal that rationally interests other individuals: it is a matter of helping oneself by helping someone else. We formally define coprivacy in terms of Nash equilibria. We then extend the concept to: i) general coprivacy, where a helping player's utility (*i.e.* interest) may include earning functionality and security in addition to privacy; ii) mixed coprivacy, where mixed strategies and mixed Nash equilibria are allowed with some restrictions; iii) correlated coprivacy, in which Nash equilibria are replaced by correlated equilibria. Coprivacy can be applied to any peer-to-peer (P2P) protocol. We illustrate coprivacy in P2P anonymous keyword search, in content privacy in social networks, in vehicular network communications and in controlled content distribution and digital oblivion enforcement.

MSC: 91A26, 91A40, 68P20, 94A60

Keywords: Data privacy, game theory, anonymous keyword search, content privacy in social networks, vehicular networks, content distribution, digital oblivion.

1. Introduction

The motivation of the coprivacy concept and its incipient theory presented in this paper is one of double sustainability in the information society:

1. *Privacy preservation is essential to make the information society sustainable just as environment preservation is essential to make the physical world sustainable*.

[*] Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. josep.domingo@urv.cat

Received: November 2010
Accepted: March 2011

This idea, which we already introduced in Domingo-Ferrer (2009) and in the conference paper Domingo-Ferrer (2010) which this article extends, should lead to clean information and communications technologies (ICT) offering functionality with minimum invasion of the privacy of individuals. Such an invasion can be regarded as a virtual pollution as harmful in the long run to the moral welfare of individuals as physical pollution is to their physical welfare. A parallel of climate change is an information society with dwindling privacy, where everyone is scared of using any service at all. Just as people's views on environment preservation have changed (they now care about environment, they require and pay for green products, etc.) and this has forced companies to change to green, the same change is happening now regarding privacy.

2. *Privacy preservation itself should be sustainable, and be achieved as effortlessly as possible as the result of rational co-operation rather than as an expensive legal requirement*. Indeed, even if privacy was acclaimed as a fundamental right by the United Nations in article 12 of the Universal Declaration of Human Rights (1948), relying on worldwide legal enforcement of privacy is nowadays quite unrealistic and is likely to stay so in the next decades. However, unlike law, technology is global and can enforce privacy worldwide, provided that privacy is achieved as the result of rational cooperation. This is the objective of the coprivacy concept and theory presented in this paper.

Two major pollutants of privacy are privacy-unfriendly security and privacy-unaware functionality. *Privacy-unfriendly security* refers to the tendency of sacrificing privacy with the excuse of security. This is partly justified by the global threat of international terrorism. With that argument, Western states have adopted shock measures on information security. Beyond the sheer technological challenge of mass-scale communications security and analysis, a new, subtler and unaddressed challenge arises: security must be increased with minimum privacy loss for the citizens. The current trend is to sacrifice privacy for alleged security: disputably, governments track phone calls, e-mails and, as seen in the Wikileaks case, social media interactions. In the private sector, privacy-unfriendly security is also present: more and more often, biometrics is enforced on customers with the argument of fighting identity theft. *Privacy-unaware* (let alone privacy-unfriendly) *functionality* is illustrated by search engines (Google, Yahoo, etc.), social networks, Web 2.0 services (*e.g.* Google Calendar, Streetview, Latitude) and so on, which concentrate on offering enticing functionality for users while completely disregarding their privacy. At most, privacy vs third parties is mentioned, but not privacy of the user vs the service provider itself, who becomes a big brother in the purest Orwellian sense.

1.1. Contribution and plan of this paper

The environmental analogy above can be pushed further by drawing inspiration on the three "R" of environment: reducing, reusing and recycling.

Reducing Re-identifiable information must be reduced. This is the idea behind database anonymization: *e.g.* *k*-anonymization (Samarati 2001) by means of microdata masking methods (*e.g.*, Domingo-Ferrer, Sebé and Solanas (2008)) reduces the informational content of quasi-identifiers. Reduction is also the idea behind ring and group signatures (Chaum and Van Heyst 2006, Groth 2007), which attempt to conciliate message authentication with signer privacy by reducing signer identifiability: the larger the group, the more private is the signer. Just as in the environment there are physical limits to the amount of waste reduction, in the privacy scenario there are functionality and security limits to reduction: completely eliminating quasi-identifiers dramatically reduces the utility of a data set (functionality problem); deleting the signature in a message suppresses authentication (security problem). A useful lesson that can be extracted from reduction is *privacy graduality*: privacy preservation is not all-or-nothing, it is a continuous magnitude from no privacy to full privacy preservation.

Reusing The idea of reusing is certainly in the mind of impersonators mounting replay attacks, but it can also be used by data protectors to gain privacy. Such is the case of re-sampling techniques for database privacy: an original data set with N records is re-sampled M times with replacement (where M can be even greater than N) and the resulting data set with M records is released instead of the original one. This is the idea behind synthetic data generation via multiple imputation (Rubin 1993). Re-sampling is also the idea of the tabular protection method in (Domingo-Ferrer and Mateo-Sanz (1999). However, as it happened for reduction there are functionality limitations to data reuse: the more reuse, the less data utility.

Recycling The idea of recycling is probably more intriguing and far less explored than reducing and reusing. Adapted to the privacy context, recycling can be regarded as leveraging other people's efforts to preserve their privacy to preserve one's own privacy. The environmental analog would be to share a car with other people: we leverage the other people's wish to save fuel to save fuel ourselves. Of course, whether in the privacy or the environment scenario, there is a functionality toll to this kind of recycling: one must adjust to the needs of other people. Nonetheless, we believe that *recycling has an enormous potential in privacy preservation, as it renders privacy an attractive and shared goal, thereby making it easier to achieve and thus more sustainable*. In this spirit, we next introduce a new recycling concept, called *coprivacy*, around which this proposal is centered.

Section 2 gives some background on game theory. Section 3 gives a game-theoretic definition of coprivacy and some of its generalizations. Section 4 illustrates coprivacy in the context of peer-to-peer (P2P) anonymous keyword search. Section 5 illustrates correlated coprivacy applied to content disclosure in social networks. Section 6 shows how general coprivacy applies to vehicular networks. Section 7 sketches how coprivacy can help enforcing controlled content distribution and digital oblivion. Section 8 summarizes conclusions and open research issues. A preliminary conference version of this paper appeared in Domingo-Ferrer (2010).

2. Basics of game theory

A game is a protocol between a set of N *players*, $\{1,\dots,N\}$. Each player i has her own *set of possible strategies*, say S_i. To play the game, each player i selects a strategy $s_i \in S_i$. We will use $s = (s_1,\dots,s_N)$ to denote the vector of strategies selected by the players and $S = \Pi_i S_i$ to denote the set of all possible ways in which players can pick strategies.

The vector of strategies $s \in S$ selected by the players determines the outcome for each player, which can be a payoff or a cost. In general, the outcome will be different for different players. To specify the game, we need to give, for each player, a preference ordering on these outcomes by giving a complete, transitive, reflexive binary relation on the set of all strategy vectors S. The simplest way to assign preferences is by assigning, for each player, a value for each outcome representing the payoff of the outcome (a negative payoff can be used to represent a cost). A function whereby player i assigns a payoff to each outcome is called a utility function and is denoted by $u_i : S \longrightarrow \mathbb{R}$.

For a strategy vector $s \in S$, we use s_i to denote the strategy played by player i and s_{-i} to denote the $(n-1)$-dimensional vector of the strategies played by all other players. With this notation, the utility $u_i(s)$ can also be expressed as $u_i(s_i, s_{-i})$.

A strategy vector $s \in S$ is a *dominant strategy solution* if, for each player i and each alternate strategy vector $s' \in S$, it holds that

$$u_i(s_i, s'_{-i}) \geq u_i(s'_i, s'_{-i}) \tag{1}$$

In plain words, a dominant strategy s is the best strategy for each player i, independently of the strategies played by all other players.

A strategy vector $s \in S$ is said to be a *Nash equilibrium* (Nash 1951) if, for all players i and each alternate strategy $s'_i \in S_i$, it holds that

$$u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$$

In plain words, no player i can change her chosen strategy from s_i to s'_i and thereby improve her payoff, assuming that all other players stick to the strategies they have chosen in s. A Nash equilibrium is self-enforcing in the sense that once the players are playing such a solution, it is in every player's best interest to stick to her strategy. Clearly, a dominant strategy solution is a Nash equilibrium. Moreover, if the solution is strictly dominant (*i.e.* when the inequality in Expression (1) is strict), it is also the unique Nash equilibrium. See Nisan, Roughgarden, Tardos and Vazirani (2007) for further background on game theory.

3. Coprivacy and its generalizations

We introduce in this section the novel concept of coprivacy in a community of peers, whereby one peer recycles to her privacy's benefit the efforts of other peers to maintain their own privacy. Informally, there is coprivacy when the best option for a peer to preserve her privacy is to help another peer in preserving his privacy. The great advantage is that *coprivacy makes privacy preservation of each specific individual a goal that interests other individuals*: therefore, privacy preservation becomes *more attractive* and hence *easier to achieve and more sustainable*. A game-theoretic formalization of coprivacy follows.

Definition 1 (Coprivacy) *Let Π be a game with self-interested, rational peer players $P^1, \ldots, P^N$, and an optional system player P^0. Each player may have leaked a different amount of private information to the rest of players before the game starts. The game is as follows: i) P^1 selects one player P^k with $k \in \{0\} \cup \{2, \cdots, N\}$ and submits a request to P^k; ii) If $k = 0$, P^0 always processes P^1's request; if $k > 1$, P^k decides whether to process P^1's request (which may involve accessing the system player on P^1's behalf) or reject it. The players' strategies are $S^0 = \{s^0_1\}$ (process P^1's request); $S^1 = \{s^1_0, s^1_2, \cdots, s^1_N\}$, where s^1_j means that P^1 selects P^j; for $i > 1$, $S^i = \{s^i_1, s^i_2\}$, where s^i_1 means processing P^1's request and s^i_2 rejecting it. Game Π is said to be* coprivate *with respect to the set $U = (u_1, \cdots, u_N)$ of privacy utility functions if, for some $k > 1$, a peer P^k exists such that (s^1_k, s^k_1) is a pure strategy Nash equilibrium between P^1 and P^k, that is, if the best strategy for P^1 is to request help to P^k and the best strategy for P^k is to provide the requested help.*

Note that the notions of privacy utility function and therefore of coprivacy are based on the aforementioned privacy graduality: one can have a varying degree of privacy preservation, hence it makes sense to trade it off. In the environmental analogy, coprivacy is a recycling concept which involves trading off waste reduction among players. A quantification of coprivacy follows:

Definition 2 (δ-Coprivacy) *Given $\delta \in [0, 1]$, the game of Definition 1 is said to be δ-coprivate with respect to the set $U = (u_1, \cdots, u_N)$ of privacy utility functions if the probability of it being coprivate for U is at least δ.*

The following extensions of coprivacy are conceivable:

- **General coprivacy** can be defined by replacing the set U of privacy utility functions in Definition 1 with a set $\mathscr{U}$ of general utility functions for peer players P^k combining privacy preservation with security and/or functionality. In general coprivacy, the interests of peers include, in addition to privacy, functionality and/or security.
- **General $\boldsymbol{\delta}$-coprivacy** can be defined by replacing U with $\mathscr{U}$ in Definition 2.
- **Mixed coprivacy** results if one allows mixed strategies for players and replaces the requirement of pure strategy Nash equilibrium in Definition 1 by a mixed strategy Nash equilibrium. The good point of mixed coprivacy is that a theorem by Nash (Nash 1951) guarantees that any game with a finite set of players and a finite set of strategies has a mixed strategy Nash equilibrium, and is therefore *mixedly coprivate*.
- **Correlated coprivacy** results if one replaces the requirement of pure Nash equilibrium in Definition 1 by a correlated equilibrium. Indeed, the outcome of independent rational behavior by users, provided by Nash equilibria, can be inferior to a centrally designed outcome. Correlated equilibria resulting from coordination of strategies may give a higher outcome. We will illustrate this in Section 5 below. In correlated equilibria, players do not have any incentive to deviate from their corresponding equilibrium strategies. An approximation to correlated equilibria are ε-correlated equilibria, in which players have at most an incentive $\varepsilon > 0$ to deviate from their corresponding equilibrium strategies. The advantage of ε-correlated equilibria is that they can always be reached by distributed heuristics run by a set of autonomous players without centrally designed strategies.
- The above extensions can be combined to yield **mixed general coprivacy** and **correlated general coprivacy**. Since mixed coprivacy is always achievable if any mixed strategy is valid for any player, **mixed $\boldsymbol{\delta}$-coprivacy** and **mixed general $\boldsymbol{\delta}$-coprivacy** only make sense when players have boundary conditions that define a subset of feasible mixed strategies.

A *coprivate protocol* is a protocol based on a coprivate game. If a privacy preservation problem can be solved by a coprivate protocol, the advantage is that it is in a player's rational privacy interest to help other players to preserve their privacy. We next give an example to show that the coprivacy concept is latent in existing protocols. More examples of the potential of coprivacy follow in the next sections.

Example 1 (Coprivacy in anonymous communication) *The success of the well-known system Tor (`http://www.torproject.org`) for anonymous communication, made even more famous by Wikileaks, can be explained by coprivacy. As hinted in the Tor website, "each new user and relay provides additional diversity, enhancing Tor's ability to put control over your security and privacy back into your hands". Therefore, using Tor is not only good for one's own privacy, but for other people's privacy as well.*

4. Coprivacy in P2P anonymous keyword search

Private information retrieval (PIR) is normally modeled as a game between two players: a user and a database. The user retrieves some item from the database without the latter learning which item was retrieved. Most PIR protocols are ill-suited to provide PIR from a search engine or large database, not only because their computational complexity is linear in the size of the database, but also because they (unrealistically) assume active cooperation by the database in the PIR protocol.

Pragmatic approaches to guarantee some query privacy have therefore been based so far on two relaxations of PIR: standalone and peer-to-peer (P2P). In the standalone approach, a program running locally in the user's computer either keeps submitting fake queries to cover the user's real queries (TrackMeNot, Howe and Nissenbaum 2009)) or masks the real query keywords with additional fake keywords (GooPIR, Domingo-Ferrer, Solanas and Castellà-Roca 2009)). In the P2P approach, a user gets her queries submitted by other users in the P2P community; in this way, the database still learns which item is being retrieved, but it cannot obtain the real query histories of users, which become diffused among the peer users, thereby achieving anonymous keyword search. We first proposed a P2P anonymous keyword search system in Domingo-Ferrer, Bras-Amorós, Wu and Manjón (2009).

Consider a system with N peers P^1 to P^N, who are interested in querying a database DB playing the role of system player P^0. If any P^i originates a query for submission to DB, she can send the query directly to DB or ask some other peer to submit the query on P^i's behalf and return the query results.

More formally, the strategies available for a requesting P^i are:

Sii**:** P^i submits her query directly to DB;
Sij**:** P^i forwards her query to P^j, for some $j \neq i$, and requests P^j to submit the query on P^i's behalf.

When receiving P^i's query, P^j has two possible strategies:

Tji**:** P^j submits P^i's query to DB and returns the answer to P^i;
Tjj**:** P^j ignores P^i's query and does nothing.

Let $X^i(t)$ be the set of queries originated by P^i up to time t. Let $Y^i(t)$ be the set of queries submitted to DB by P^i up to time t. For each query x^i_r in $X^i(t)$, define $F^i(x^i_r, t)$ as the set of players to whom P^i has forwarded x^i_r for submission up to time t. The players in $F^i(x^i_r, t)$ can be associated relative frequencies as follows: for $j = 1$ to N with $j \neq i$, let $f^{ij}(x^i_r, t)$ be the relative frequency with which P^i has forwarded x^i_r to player P^j, up to time t.

The privacy utility function for P^i should reflect the following intuitions: (i) the more homogeneous the relative frequencies of queries in $Y^i(t)$, the more private stay the interests of P^i vs DB; (ii) the more homogeneous the relative frequencies of peers in $F^i(x^i_r, t)$ for every $x^i_r \in X^i(t)$, the more private stay the interests of P^i vs the other peers.

Given a random variable Z taking values $z_1, z_2, \ldots, z_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, Shannon's entropy (Shannon 1948) is a measure of uncertainty defined as

$$H(Z) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

The more homogeneous the p_i, the higher is $H(Z)$: the rationale is that the outcome of Z becomes more uncertain as the p_i become more homogeneous. The maximum $H(Z)$ is reached when $p_1 = \cdots = p_n = 1/n$.

By assimilating $Y^i(t)$ and $F^i(x^i_r, t)$ to random variables and relative frequencies to probabilities, intuition (i) above can be expressed as maximizing $H(Y^i(t))$ and intuition (ii) as maximizing $H(F^i(x^i_r, t))$ for all $x^i_r \in X^i(t)$. Hence, those Shannon entropies are reasonable privacy utility functions for P^i. When P^i generates a query x^i_r at time $t+1$:

- P^i chooses Sii (direct submission) if $H(Y^i(t+1)) \geq H(Y^i(t))$, where $Y^i(t+1) = Y^i(t) \cup \{x^i_r\}$;
- Otherwise P^i chooses Sik (forwarding the query to P^k), where

$$k = \arg \max_{j \in \{1,\cdot,N\} \setminus \{i\}} H(F^i(x^i_r, t+1)). \tag{2}$$

In plain words, if direct submission decreases privacy vs DB, the query is forwarded to the player P^k vs whom the privacy loss is minimum. Note that if P^i forwards her query to a player P^j, P^i always incurs some privacy loss vs P^j, because P^j knows the query has been generated by P^i. Therefore, the best policy is to distribute the successive submissions of a certain query x^i_r as evenly as possible among the various peers. This is what the choice of k in Expression (2) attempts.

When P^k receives x^i_r, it proceeds as follows:

- P^k chooses Tki (submitting x^i_r) if $H(Y^k(t+1)) > H(Y^k(t))$, where $Y^k(t+1) = Y^k(t) \cup \{x^i_r\}$;
- Otherwise P^k chooses Tkk (ignoring x^i_r).

In plain words, P^k submits x^i_r only if doing so increases her privacy vs the DB. If P^k ignores x^i_r, then P^i will have to look for a second best player to submit x^i_r (and a third best if the second best ignores x^i_r, and so on). If, after a number of attempts to be decided by P^i, no peer is found who is willing to help, then P^i must submit x^i_r herself.

If P^i's best strategy is Sik and P^k best strategy is Tki, then (Sik, Tki) is a pure-strategy Nash equilibrium between P^i and P^k and *there is coprivacy* between P^i and P^k.

We give a detailed formalization and empirical results for the N-player P2P anonymous keyword search game in the manuscript Domingo-Ferrer and González-Nicolás (2011).

5. Correlated coprivacy in social networks

Social networks (SNs) have become an important web service with a broad range of applications: collaborative work, collaborative service rating, resource sharing, friend search, etc. Facebook, MySpace, Xing, etc., are well-known examples. In an SN, a user publishes and shares information and services.

There are two types of privacy in SNs:

- *Content privacy.* The information a user publishes clearly affects her privacy. Recently, a privacy risk score (Liu and Terzi 2009) has been proposed for the user to evaluate the privacy risk caused by the publication of a certain information. Let the information attributes published by the users in an SN be labeled from 1 to n. Then the privacy score risk of user j is

$$PR(j) = \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} \times V(i, j, k)$$

  where $V(i, j, k)$ is the visibility of user j's value for attribute i to users which are at most k links away from j and β_{ik} is the sensitivity of attribute i vs those users.
- *Relationship privacy.* In some SNs, the user can specify how much it trusts other users, by assigning them a trust level. It is also possible to establish several types of relationships among users (like “colleague of”, “friend of”, etc.). The trust level and the relationship type are used to decide whether access is granted to resources and services being offered (*access rule*). The availability of information on relationships (trust level, relationship type) has increased with the advent of the Semantic Web and raises privacy concerns: knowing who is trusted by whom and to what extent discloses a lot about the user's thoughts and feelings. For a list of related abuses see Barnes (2006). In Domingo-Ferrer, Viejo, Sebé and González-Nicolás (2008), we described a new protocol offering private relationships in an SN while allowing resource access through indirect relationships without requiring a mediating trusted third party.

We focus here on content privacy in SNs. A possible privacy-functionality score for user j reflecting the utility the user derives from participating in an SN is

$$\begin{aligned} PRF(j) &= \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i, j', k) I(j, j', k)}{1 + PR(j)} \\ &= \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i, j', k) I(j, j', k)}{1 + \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i, j, k)} \end{aligned}$$

where $I(j, j', k)$ is 1 if j and j' are k links away from each other, and it is 0 otherwise.

Note that:

- $PRF(j)$ decreases as the privacy score $PR(j)$ in its denominator increases, that is, as user j discloses more of her attributes.
- $PRF(j)$ increases as its numerator increases; this numerator adds up the components of privacy scores of users $j' \neq j$ due to those users disclosing attribute values to j.

The dichotomous version of the above privacy-functionality score, for the case where an attribute is simply either made public or kept secret, is:

$$PRF_2(j) = \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,\, j')}{1 + PR(j)}$$

$$= \frac{\sum_{j'=1, j' \neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,\, j')}{1 + \sum_{i=1}^{n} \beta_i V(i,\, j)} \qquad (3)$$

If we regard $PRF(j)$ as a game-theoretic utility function (Tardos and Vazirani 2007), the higher $PRF(j)$, the higher the utility for user j.

For instance, take a strategy vector $s = (s_1, \ldots, s_N)$ formed by the strategies *independently and selfishly* chosen by all users and consider the dichotomous case, that is, let the utility incurred by user j under strategy s be $u_j(s) = PRF_2(j)$. It is easy to see (and it is formally shown in Domingo-Ferrer (2010b) that rational and independent choice of strategies leads to a Nash equilibrium where no user offers any information on the SN, which results in an SN collapse. See Example 2 below.

A similar pessimistic result is known for the P2P file sharing game, in which the system goal is to leverage the upload bandwidth of the downloading peers: the dominant strategy is for all peers to attempt "free-riding", that is, to refuse to upload (Babaioff, Chuang and Feldman 2007), which causes the system to collapse.

Example 2 *The simplest version of the above game is one with two users having each one attribute, which they may decide to keep hidden (a strategy denoted by H, which implies visibility 0 for the attribute) or publish (a strategy denoted by P, which implies visibility 1). Assuming a sensitivity $\beta = 1$ for that attribute and using $u_j(s) = PRF_2(j)$, the user utilities for each possible strategy vector are as follows:*

$$u_1(H,H) = 0; u_1(H,P) = 1; u_1(P,H) = 0; u_1(P,P) = 1/2$$
$$u_2(H,H) = 0; u_2(H,P) = 0; u_2(P,H) = 1; u_1(P,P) = 1/2$$

This simple game can be expressed in matrix form:

User 2	*H*	*P*
User 1 *H*	*0* *0*	*0* *1*
P	*1* *0*	*1/2* *1/2*

The above matrix corresponds to the Prisoner's Dilemma (Tardos and Vazirani 2007), perhaps the best-known and best-studied game. Consistently with our argument for the general case, it turns out that (H,H) *is a dominant strategy, because:*

$$u_1(H,P) = 1 \geq u_1(P,P) = 1/2; u_1(H,H) = 0 \geq u_1(P,H) = 0$$

$$u_2(P,H) = 1 \geq u_1(P,P) = 1/2; u_2(H,H) = 0 \geq u_2(H,P) = 0$$

The second and fourth equations above guarantee that (H,H) *is a Nash equilibrium (in fact, the only one). The Prisoner's Dilemma with* $N > 2$ *users is known as the Pollution Game (Tardos and Vazirani 2007) and corresponds to the dichotomous SN game considered above.*

The outcome of independent rational behavior by users, provided by Nash equilibria and dominant strategies, can be inferior to a centrally designed outcome. This is clearly seen in Example 2: the strategy (P, P) would give more utility than (H, H) to *both* users. However, usually no trusted third-party accepted by all users is available to enforce correlated strategies; in that situation, the problem is how User 1 (resp. User 2) can guess whether User 2 (resp. User 1) will choose P.

Using a solution based on cryptographic protocols for bitwise fair exchange of secrets would be an option, but it seems impractical in current social networks, as it would require a cryptographic infrastructure, unavailable in most SNs.

A more practical solution to this problem may be based on direct reciprocity (*i.e.* tit-for-tat) or reputation, two approaches largely used in the context of P2P file-sharing systems. We describe in Domingo-Ferrer (2010b) two correlated (actually ε-correlated) equilibrium heuristic protocols based on tit-for-tat and reputation, respectively. They are intended as "assistants" to the human user of the SN in deciding whether to disclose an attribute to another user; however, the ultimate decision belongs to the human, who may quit and renounce to reach the equilibrium.

Those heuristic protocols offer *ε-correlated general coprivacy*, referred to a utility combining privacy and functionality.

6. General coprivacy in vehicular networks

Vehicular *ad hoc* networks permitting car-to-car communication are expected to be available in cars manufactured in the near future. Several standards for VANET communication are under way both in the United States (DSRC, Dedicated Short Range Communications, IEEE 802.11p) and Europe (C2C Consortium). We argue that VANETs must provide functionality, security and privacy and are therefore an application where general coprivacy can be used:

Functionality. The main *raison d'être* of VANETs is to allow vehicles to timely disseminate announcement messages about *current* road conditions (*e.g.* icy road, traffic jam, etc.) to other vehicles, in order to improve traffic safety and efficiency.

Security. Announcement messages must be trustworthy, because false messages could seriously disrupt traffic, cause accidents and/or cause certain areas to become deserted and thus an easy prey for criminals. A posteriori security consists of punishing vehicles that have been proven to have originated false messages (*e.g.* Lin, Sun, Ho and Shen (2007)); hence, means are required to identify malicious vehicles, for example digital signatures. *A priori* security is an alternative or a complement whereby one attempts to prevent the generation of false messages (*e.g.* Raya, Aziz and Hubaux (2006)): a message is given credit only if it is has been endorsed by a number of nearby vehicles greater than a certain threshold.

Privacy. It would not be very fair if the collaboration of a driver to improve traffic safety and efficiency (functionality) by generating or endorsing announcements forced her to disclose her identity and location. Note that knowing someone's mobility pattern reveals a lot of private information: the driving style leaks information about an individual's character (nervous, calm), her whereabouts tell about her work and social habits, etc. Privacy can be added to *a posteriori* security by using pseudonyms or advanced cryptography like group signatures. Adding privacy to *a priori* security may imply vulnerability against the Sybil attack, whereby a vehicle generates a false message and takes advantage of anonymity to compute itself as many endorsements as required. We have proposed in Daza, Domingo-Ferrer, Sebé and Viejo (2009) a private *a priori* scheme based on threshold signatures which is resistant against the Sybil attack and provides irrevocable anonymity to cars generating or endorsing messages.

Security is a *must* in VANETs and cannot be traded off. Therefore *the general coprivacy that applies in vehicular networks involves a utility function combining functionality and privacy*. General coprivacy is applicable to VANETs in the following sense:

- The more privacy players allow to another player, the more announcements (functionality) they can expect from that player.

- Conversely, the more announcements players originate, the more privacy for other announcing players: indeed, the more cars originate an announcement “icy road near longitude X latitude Y”, the more private the originators stay (this is the “reusing” principle mentioned above).

7. Controlled content distribution and digital oblivion

In conventional multicast transmission one sender sends the same content to a set of receivers. This precludes fingerprinting the copy obtained by each receiver (in view of redistribution control and other applications). A straightforward alternative is for the sender to separately fingerprint and send in unicast one copy of the content for each receiver. This approach is not scalable and may implode the sender.

Distributed multicast of fingerprinted content can be modeled as a coprivate protocol. Indeed, mechanism design can be used to craft a protocol such that content receivers rationally co-operate in fingerprinting and further spreading the content in a tree-like fashion. If fingerprinting at each forwarding step is anonymous Pfitzmann and Waidner 1997, Bo, Piyuan and Wenzheng 2007, Domingo-Ferrer 1999), honest receivers will stay anonymous and free from false accusation, but unlawful redistributors will be traceable.

A related problem is the lack of digital forgetting in the information society. Digital storage allows perfect and unlimited remembering. However, the right of an individual to enforce oblivion for pieces of information about her is part of her fundamental right to privacy. Enforcing expiration dates for content has been championed as a solution in Mayer-Schönberger (2009), but in a way that depends on trusted storage devices deleting the content after its expiration date. Alternative hardware approaches based on employing smart cards on the user side to process encrypted content (Domingo-Ferrer 1997) could also be envisioned, whereby the smart card would not decrypt the content after its expiration date. However, such devices do not currently exist; worse yet, placing trust in the hardware (storage devices, smart cards, etc.) to implement information protection policies has proven to be a flawed approach: *e.g.*, hardware copy prevention mechanisms for CDs and DVDs were easily bypassed.

Digital oblivion via expiration date enforcement can be reached through a coprivate protocol (Domingo-Ferrer 2011). The idea is just to fingerprint expiration dates in the content. This allows identifying and punishing whoever spreads or uses content past the expiration date. If fingerprinting is asymmetric and/or anonymous, it will not be possible to falsely accuse honest content receivers. The problem then reduces to distributed multicast of asymmetrically/anonymously fingerprinted content, which is approachable a coprivate protocols, as hinted above. With anonymous fingerprinting, honest players preserve their privacy. Therefore, the receivers must honestly contribute to the content source’s privacy preservation (oblivion enforcement by anonymously fingerprinting any forwarded content with its expiration date) to preserve their own privacy. Hence, the solution is a coprivate protocol.

8. Conclusions and research directions

We have introduced in this paper the novel concept of coprivacy, as well as an incipient generalization theory on it. The main contribution of coprivacy is to make data privacy an attractive feature, especially in peer-to-peer applications:

- In many situations, players can better preserve their own privacy if they help other players in preserving theirs. We say that those situations can be handled by so-called coprivate protocols.
- In other situations, the utility of players consists of a combination of privacy plus security and/or functionality. If they can increase their own utility by helping others in increasing theirs, the situation can be handled by a generally coprivate protocol.

We have sketched the potential of coprivate protocols in very diverse areas: P2P anonymous keyword search, content disclosure in social networks, communication in vehicular networks, controlled content distribution and digital oblivion implementation.

Future research directions include developing the theory of coprivacy in the following non-exhaustive directions:

- Develop a theory of coprivacy which, given a privacy preservation problem and a parameter $\delta \in [0,1]$, can answer under which conditions a δ-coprivate game (*i.e.* protocol) that solves the problem exists.
- Elaborate a theory of general coprivacy which also takes security and functionality into account. In this generalization, the Nash or the correlated equilibrium that characterizes coprivacy is to be reached by considering utilities which combine the privacy with the security and/or the functionality obtained by the players.
- Elaborate a theory of mixed coprivacy to characterize when mixed strategies and therefore mixed coprivacy make sense for utilities about privacy, security and functionality.
- Create new cryptographic protocols to implement the privacy graduality needed in coprivacy. Specifically, *ad hoc* broadcast encryption and anonymous *ad hoc* broadcast encryption inspired in Wu, Mu, Susilo, Qin and Domingo-Ferrer (2009), (n,N)-anonymity signatures and some multiparty computation protocols for social networks are needed.

Acknowledgments and disclaimer

This work was partly funded by the Spanish Government through projects TSI2007-65406-C03-01 “E-AEGIS” and CONSOLIDER INGENIO 2010 CSD2007-0004 “ARES”, and by the Government of Catalonia through grant 2009 SGR 1135. The author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia. He holds

the UNESCO Chair in Data Privacy, but the views expressed in this paper are his own and do not commit UNESCO.

References

Babaioff, M., Chuang. J. and Feldman, M. (2007). Incentives in peer-to-peer systems, in N. Nisan, T. Roughgarden, É. Tardos and V. V. Vazirani (eds.), *Algorithmic Game Theory*, Cambridge University Press, 593–611.

Barnes, S. B. (2006). A privacy paradox: social networking in the United States, *First Monday*, 11.

Bo, Y., Piyuan, L. and Wenzheng, Z. (2007). An efficient anonymous fingerprinting protocol, in *Computational Intelligence and Security*, Springer, LNCS 4456, 824–832.

Chaum, D. and van Heyst, E. (2006). Group signatures, in *Advances in Cryptology-Eurocrypt'91*, Springer, LNCS 547, 257–265.

Daza, V., Domingo-Ferrer, J., Sebé, F. and Viejo, A. (2009). Trustworthy privacy-preserving car-generated announcements in vehicular ad hoc networks, *IEEE Transactions on Vehicular Technology*, 58, 1876–1886.

Domingo-Ferrer, J. (1997). Multi-application smart cards and encrypted data processing, *Future Generation Computer Systems*, 13, 65–74.

Domingo-Ferrer, J. (1999). Anonymous fingerprinting based on committed oblivious transfer, in *Public Key Cryptography-PKC 99*, Springer, LNCS 1560, 43–52.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1999). On resampling for statistical confidentiality in contingency tables, *Computers & Mathematics with Applications*, 38, 13–32.

Domingo-Ferrer, J., Sebé, F. and Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation, *Computers & Mathematics with Applications*, 55, 717–732.

Domingo-Ferrer, J., Viejo, A., Sebé, F. and González-Nicolás, Ú. (2008). Privacy homomorphisms for social networks with private relationships, *Computer Networks*, 52, 3007–3016.

Domingo-Ferrer, J. (2009). The functionality-security-privacy game, in *Modeling Decisions for Artificial Intelligence-MDAI 2009*, Springer, LNCS 5861, 92–101.

Domingo-Ferrer, J., Solanas, A. and Castellà-Roca, J. (2009). $h(k)$-Private information retrieval from privacy-uncooperative queryable databases, *Online Information Review*, 33, 720–744.

Domingo-Ferrer, J., Bras-Amorós, M., Wu, Q. and Manjón, J. (2009). User-private information retrieval based on a peer-to-peer community, *Data and Knowledge Engineering*, 68, 1237–1252.

Domingo-Ferrer, J. (2010). Coprivacy: towards a theory of sustainable privacy, in *Privacy in Statistical Databases-PSD 2010*, Springer, LNCS 6344, 258–268.

Domingo-Ferrer, J. and González-Nicolás, Ú. (2011). Rational behaviour in peer-to-peer anonymous keyword search, manuscript.

Domingo-Ferrer, J. (2010). Rational privacy disclosure in social networks, in *Modeling Decisions for Artificial Intelligence-MDAI 2010*, Springer, LNCS 6408, 255–265.

Domingo-Ferrer, J. (2011). Rational enforcement of digital oblivion, in *4th International Workshop on Privacy and Anonymity in the Information Society (PAIS 2011)*, ACM Digital Library (to appear).

Groth, J. (2007). Fully anonymous group signatures without random oracles, in *Proc. of ASIACRYPT 2007*, LNCS 4833, 164–180.

Howe, D. C. and Nissenbaum, H. (2009). TrackMeNot: Resisting surveillance in web search, in *Lessons from the Identity Trail*, Oxford University Press, 409–428.

Lin, X., Sun, X., Ho, P.-H. and Shen, X. (2007). GSIS: A secure and privacy-preserving protocol for vehicular communications, *IEEE Transactions on Vehicular Communications*, 56, 3442–3456.

Liu, K. and Terzi, E. (2009). A framework for computing the privacy scores of users in online social networks, in *Proc. of ICDM 2009-The 9th IEEE International Conference on Data Mining*, 288–297.

Mayer-Schönberger, V. (2009). *The Virtue of Forgetting in the Digital Age*, Princeton University Press.

Nash, J. (1951). Non-cooperative games, *Annals of Mathematics*, 54, 289–295.

Nisan, N., Roughgarden, T., Tardos, É. and Vazirani, V. V. eds. (2007). *Algorithmic Game Theory*, Cambridge University Press.

Pfitzmann, B. and Waidner, M. (1997). Anonymous fingerprinting, in *Advances in Cryptology-EUROCRYPT 1997*, Springer, LNCS 1233, 88–102.

Raya, M., Aziz, A. and Hubaux, J.-P. (2006). Efficient secure aggregation in VANETs, in *Proc. of 3rd Intl. Workshop on Vehicular Ad Hoc Networks-VANET*, 67–75.

Rubin, D. B. (1993). Discussion on statistical disclosure limitation, *Journal of Official Statistics*, 9, 461–468.

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

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal*, 27, 379–423 and 623–656.

Tardos, É. and Vazirani, V. V. (2007). Basic solution concepts and computational issues, in N. Nisan, T. Roughgarden, É. Tardos and V. V. Vazirani (eds.), *Algorithmic Game Theory*, Cambridge University Press, 3–28.

Wu, Q., Mu, Y., Susilo, W., Qin, B. and Domingo-Ferrer, J. (2009). Asymmetric group key agreement, in *Advances in Cryptology-EUROCRYPT 2009*, Springer, LNCS 5479, 153–170.

SORT Special issue: Privacy in statistical databases, 2011, 41-58

Semantic microaggregation for the anonymization of query logs using the open directory project

Arnau Erola[1], Jordi Castellà-Roca[1], Guillermo Navarro-Arribas[2]
and Vicenç Torra[3]

Abstract

Web search engines gather information from the queries performed by the user in the form of query logs. These logs are extremely useful for research, marketing, or profiling, but at the same time they are a great threat to the user's privacy. We provide a novel approach to anonymize query logs so they ensure user *k*-anonymity, by extending a common method used in statistical disclosure control: microaggregation. Furthermore, our microaggregation approach takes into account the semantics of the queries by relying on the Open Directory Project. We have tested our proposal with real data from AOL query logs.

MSC: 68T35

Keywords: Privacy, web search engines, query logs, *k*-anonymity, microaggregation, semantic.

1. Introduction

Web Search Engines play a decisive role in the Internet nowadays. For instance, there is an estimate of over 113 billion searches conducted globally on the Internet during July 2009, which is up by 41% percent compared to July 2008 (SearchEngineWatch, 2009). These numbers give some insight on the relevance and growth rate use of Web search engines (WSE). Major WSE such as Google, Yahoo!, Baidu, or Microsoft's Bing serve most of the searches in the global Internet with respective shares of 67.5%, 7.8%, 7.0%,

[1] Departament d'Enginyeria Informàtica i Matemàtiques, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Spain. arnau.erola, jordi.castella@urv.cat

[2] Department of Information and Communications Engineering, Universitat Autónoma de Barcelona, 08193 Bellaterra (Catalonia, Spain). gnavarro@deic.uab.cat

[3] IIIA, Institut d'Investigació en Intel·ligència Artificial, CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain). vtorra@iiia.csic.es

Received: November 2010
Accepted: March 2011

and 2.9% in 2008. This share is more proportional if we look for example at US figures, where in September 2010 the share of searches was 65.4% (Google), 17.4% (Yahoo), and 11.1% (Microsoft) (SearchEngineWatch, 2010). Web search is not only important in the global Internet, as most sites, corporate intranets, or community portals provide local WSEs.

The information gathered by a WSE is stored and can be used to provide personalized search results (Gauch and Speretta, 2004), to conduct marketing research (Hansell, 2006), or provide personalized advertisement. These data, normally referred to as *search* or *query logs*, are a great economic source for the WSE, for instance, Google had a revenue of 21 128.5 million dollars in 2008 from advertisements (Google, 2008), which is strongly based in the information gathered by their search engine. WSEs also charge law enforcement agencies for access to user or group profiles (Summers, 2009; Zetter, 2009).

The detailed information that can be obtained from query logs, make these data an important threat to the privacy of the users. For instance, in 2006, AOL Research, in an attempt to help the information retrieval research community, released over 21 million queries from over 650, 000 subscribers over a 3 month period. Although the data were previously anonymized, they still carried enough information to be an important threat to the subscribers' privacy. Journalists from the New York Times were able to locate an individual (Barbaro and Zeller, 2006) from the query logs, and several other sensitive information was exposed. The case ended up not only with an important damage to AOL users' privacy, but also with a major damage to AOL itself, with several class action suits and complaints against the company (EFF, 2009; Mills, 2006).

In this paper, we address the privacy problem exposed by the WSE query logs when they are made publicly available, transferred to third parties, or stored for future analysis. The main objective is to preserve the utility of the data without risking the privacy of their users. To that end, we follow the same ideas found in statistical disclosure control (SDC), proposing a novel microaggregation method to anonymize query logs. This approach ensures a high degree of privacy, providing k-anonymity at user level, while preserving some of the data usefulness. Moreover, and unlike most of the previous work, our approach takes into account the semantics of the queries made by the user in the anonymization process making use of information obtained from the Open Directory Project (2010).

The paper is organized as follows. Section 2 introduces microaggregation and our motivation and approach for the semantic anonymization of query logs. In Section 3 we detail our proposal, and Section 4 presents our results in terms of protection and utility. Section 5 discusses the related work, and finally, Section 6 concludes the paper.

1.1. Privacy Problems

The privacy problem of query logs is given by the fact that they can contain personal information (Soghoian, 2007). For instance, a user may have searched for her city, a local team, a disease suffered by herself, adult content, or she can make a vanity query, for which the user searches for her own name (Kumar *et al.*, 2007; Soghoian, 2007). This information, either by itself or with help of more information can allow to re-identify the user (Frankowski *et al.*, 2006). So the main threat exposed by a query log is to be able to link user queries with a real identity. The anonymization process can remove a lot of information to provide a high level of privacy to the user, but the resulting log might not be very useful. On the other hand, a more useful log can be obtained if less information is removed. So, there is a privacy-utility tradeoff (Adar, 2007). Query logs should be properly protected with an anonymization process and data should remain useful.

Accordingly, any release of query logs must ensure two requirements:

- **Anonymity**: queries alone or with external information cannot be used to re-identify any user.
- **Usefulness**: queries must contain enough true information to bear likeness to the reality and to be minimally useful. If the information is very damaged, it loses its reliability and value.

To make the personal information retrieval difficult, queries are usually combined with other ones that obfuscate them. Microaggregation (Defays and Nanopoulos, 1993) is a popular statistical disclosure control technique, which provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

Microaggregation provides privacy comparable with k-anonymity (Samarati, 2001; Sweeney, 2002), i.e., a query of a certain user cannot be distinguished from at least $k-1$ queries generated by other users. So, the identification of a user must be imprecise. In terms of usefulness, the larger k is, the less achieved usability because the microaggregated log keeps less information of each user (see Section 4.1).

2. Towards a semantic microaggregation for query logs

In this paper, we propose a novel microaggregation method for query logs taking into account the semantics of the queries made by the users. In this section, we overview microaggregation and discuss the motivations of our proposal.

2.1. Microaggregation

In microaggregation, privacy is ensured because all clusters have at least a predefined number of elements, and therefore, there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data.

Microaggregation was originally defined for numerical attributes (Defays and Nanopoulos, 1993), but later extended to other domains, for example, to categorical data in Torra (2004) (see also Domingo-Ferrer and Torra, 2005), and in constrained domains in Torra (2008).

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least k records.
- **Aggregation.** For each of the clusters, a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. Let v_i be the representative of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$\begin{aligned}
&\text{Minimize} \quad SSE = \textstyle\sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij} (d(x_j, v_i))^2 \\
&\text{Subject to} \;\; \textstyle\sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\
&\qquad\qquad 2k \geq \textstyle\sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\
&\qquad\qquad u_{ij} \in \{0, 1\}
\end{aligned}$$

For numerical data, it is usual to require that $d(x, v)$ is the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors, and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i is defined as the arithmetic mean of the records in the cluster, that is, $v_i = \sum_{j=1}^{n} u_{ij} x_i / \sum_{j=1}^{n} u_{ij}$. As the solution of this problem is NP-Hard (Oganian and Domingo-Ferrer, 2001) when we consider more than one variable at a time (multivariate microaggregation), heuristic methods have been developed. One such method is MDAV (*Maximum Distance to Average Vector*) (Domingo-Ferrer and Mateo-Sanz, 2002).

Note that when all variables are considered at once, microaggregation is a way to implement k-anonymity (Samarati, 2001; Sweeney, 2002).

2.2. Motivations of our proposal

In order to ensure the privacy of the users, we provide k-anonymity at user level in the protected query logs. That is, in the protected logs there will be at least k indistinguishable users.

```
Open Directory Categories  (1-5 of 5)
   1. Sports: Soccer: UEFA: Spain: Clubs: Barcelona   (11 matches)
   2. World: Polski: Sport: Sporty pilki i siatki: Pilka nozna: Kluby: Hiszpan’skie: (...)
   3. World: Español: Regional: Europa: España: Deportes y tiempo libre: Deportes: (...)
   4. World: Deutsch: Sport: Ballsport: Fuball: Vereine: Spanien   (3)
   5. World: Français: Sports: Balles et ballons: Football: Regional: Europe: Espagne   (3)
```

Figure 1: *Example of ODP query result.*

A key point, thus, for the microaggregation of search logs is to determine how the users are clustered. If the users in the same cluster do not share any interest, the protected query logs can be useless, that is, the resulting search logs are too much distorted and we cannot obtain useful information from them.

For example, we can consider two soccer supporters, and two anti-sports users. If we create a cluster of size two with a soccer supporter and an anti-sports user, we can obtain non-valid results. The entries of the protected query logs are confusing. On the other hand, if the two soccer supporters are in the same cluster, the protected logs provide more reliable results.

Thus, we should create the groups of users taking into consideration their interests. The users with common interests between them should be grouped in the same cluster. In order to do so, we should be able to determine if their interests are closer, that is, we need a tool to compute the semantic distance of two queries.

In this work, we use the Open Directory Project (ODP) (ODP, 2010) to compute the semantic distances between users. The ODP is the most widely distributed database of Web content classified by humans. ODP data powers the core directory services for some of the most popular portals and search engines on the Web, including AOL Search, Netscape Search, Google, Lycos, and HotBot, and hundreds of others. Thus, a query result using them is hardly influenced by the ODP classification. ODP uses a hierarchical ontology structure to classify sites according to their themes. For example, when we search for *Barcelona FC*, ODP returns a list of categories to which the query belongs (Figure 1). Each result starts with a root category followed by deeper categories in the ODP tree.

Our proposal groups users with common interests using the ODP classification. We consider that the users with common interest are those who have more terms in the same categories.

2.3. An ODP similarity measure

In order to be able to microaggregate users from the query logs, we have to define a distance or similarity measure between users. We introduce a similarity coefficient based on the common categories shared between queries from each user. We also introduce some notation here to formalize the process.

We consider the set of n users $U = \{u_1, \dots, u_n\}$ from the query log, and their respective set of queries $Q = \{Q_1, \dots, Q_n\}$, where $Q_i = \{q^i_1, \dots, q^i_{m_i}\}$ are the queries of the user u_i. Each query q^i_j has several terms $q^i_j = \{t_1, \dots, t_{r_j}\}$.

Given a term t_s, we can obtain its classification in the ODP at a given depth. When querying the ODP, the returned categories can be divided in depth levels. Let l be the parameter that identifies the depth level in the ODP hierarchy. For example, if we have the classification $Sports:Soccer:UEFA:Spain:Clubs:Barcelona$ and $l=1$, we only work with the root category $Sports$; when $l=2$ we work with $Sports:Soccer$; and so on. We will consider a maximum depth L to restrict the search space, so $l \in \{1,\ldots,L\}$.

We denote as $C_l = \{c^l_1,\ldots,c^l_{p_l}\}$ the set of possible categories at level l in the ODP. Given a user u_i we can obtain all the categories at level l from all queries of the user. We denote as $C_l(u_i)$ the set of categories for user u_i at level l. Note that considering all queries of user u_i, $Q_i = \{q^i_1,\ldots,q^i_{m_i}\}$, and their respective sets of terms $q^i_j = \{t_1,\ldots,t_{r_j}\}$ for $j = 1\ldots m_i$, the number of categories for user u_i at level l is given by $|C_l(u_i)| = r_1+\ldots+r_{m_i}$.

We can then define a similarity coefficient ODP_{sim} between two given users u_i and u_j as:

$$OPD_{sim}(u_i,u_j) = \sum_{l=1}^{L}\{|c_l| : c_l \in \{C_l(u_i) \cup C_l(u_j)\}\} \tag{1}$$

This similarity coefficient between two users computes the common categories between them for all considered levels, that is levels up to L. Note that OPD_{sim} is symmetric and ranges from 0 (there is no similarity between the users) to $\sum_{l=1}^{L}|C_l|$ (maximum similarity between two users).

3. ODP-based microaggregation of query logs

The method we propose to protect the query logs is a microaggregation that follows the outline of Section 2 with an extra step of data preparation. That is, our approach consists of the following steps:

1. Data preparation.
2. Partition.
3. Aggregation.

These steps are described in detail in the following sections.

3.1. Data preparation

To easy the computation of the protected data, the data is prepared by pre-querying the ODP to classify the user queries. Following the notation introduced in Section 2.3, for every term t_s, we can obtain its classification for all levels $l \in \{1,\ldots,L\}$ using the ODP.

This allows us to obtain all the categories associated to all the users in all levels, that is $C_l(u_i)$ for all user $u_i \in U$, and all considered levels. Next, we create a *classification matrix* that contains the number of queries for each user and category at level l, $M_{U\times C_l}$. Please, note that, we obtain one matrix for every level $l \in \{1,\ldots,L\}$. So, $M_{U\times C_l}(i,j)$ is the number of times that category c^l_j is found in the queries of user u_i.

Finally, we use the $M_{U\times C_l}$ matrices in order to compute the *incidence matrix* that contains the semantic similiarity of the users $M_{U\times U}$. Given the incidence matrix $M_{U\times U}$, $M_{U\times U}(i,j)$ is the number of common categories between users u_i, and u_j for all depth levels $l \in \{1,\ldots,L\}$. Moreover note that the incidence matrix corresponds to the similarity coefficient described in Section 2.3, that is, $M_{U\times U}(i,j) = ODP_{sim}(u_i,u_j)$.

The process works as follows:

1. Obtain the classification matrices $M_{U\times C_l}$ using Algorithm 1.
2. Obtain the incidence matrix $M_{U\times U}$ using Algorithm 2, i.e. the similarity coefficient between users.

Algorithm 1 Algorithm for computing the classification matrices $M^L_{U\times C}$ where $L = \{1,\ldots,l\}$

```
Require: the maximum depth $L$ for the ODP categories
Require: the set of users $U = \{u_i,\ldots,u_n\}$
Require: the set of queries $Q_i = \{q^i_1,\ldots,q^i_{m_i}\}$ of each user $u_i$
Require: the set of terms $\{t_1,\ldots,t_{r_j}\}$ of each query $q_j$
Ensure: $\{M_{U\times C_1},\ldots,M_{U\times C_L}\}$, i.e. for every level $l$, the matrix $M_{U\times C_l}$ with the number of queries for each
  category and user in the depth $l$
  for $l \in \{1,\ldots,L\}$ do
    for $u_i \in \{u_1,\ldots,u_n\}$ do
      for $q^i_j \in Q_i = \{q^i_1,\ldots,q^i_{m_i}\}$ do
        for $t_s \in q^i_j = \{t_1,\ldots,t_{r_j}\}$ do
          obtain the categories $c_t$ at depth $l$ for the term $t_s$ using ODP;
          for each $c_t$ do
            if $c_t \in M_{U\times C_l}$ then
              $M_{U\times C_l}(u_i,c_t) = M_{U\times C_l}(u_i,c_t)+1$;
            else
              add the column $c_t$ to $M_{U\times C_l}$;
              $M_{U\times C_l}(u_i,c_t) = 1$;
            end if
          end for
        end for
      end for
    end for
  end for
  return $\{M_{U\times C_1},\ldots,M_{U\times C_L}\}$.
```

Algorithm 2 Algorithm for computing the incidence matrix $M_{U\times U}$

```
Require: the classification matrices $\{M_{U\times C_1},\dots,M_{U\times C_L}\}$
Ensure: $M_{U\times U}$
  Initialize $M_{U\times U}(i,j) \leftarrow 0$ for all $i,j = 1\dots n$;
  for $M_{U\times C_l} \in \{M_{U\times C_1},\dots,M_{U\times C_L}\}$ do
    for each column $c_j \in M_{U\times C_l}$ do
      for each row $u_i \in M_{U\times C_l}$ do
        for each row $u_\rho \in M_{U\times C_l}$ do
          $M_{U\times U}(u_i,u_\rho) \leftarrow M_{U\times U}(u_i,u_\rho) + \min(M_{U\times C_l}(u_i,c_j), M_{U\times C_l}(u_\rho,c_j))$;
        end for
      end for
    end for
  end for
  return $M_{U\times U}$.
```

3.2. Partition

The partition step creates groups of k users with similar interests using Algorithm 3.

Let us assume that u_i and u_ρ are the most similar users in the set. We calculate the users' similarity ODP_{sim} using the incidence matrix $M_{U\times U}$, (see Section 3.1). The most similar users are those that have the highest similarity coefficient in the matrix. Next, we include u_i and u_ρ to the cluster. If the group size k is two, we delete u_i and u_ρ records from the incidence matrix and we repeat the process to obtain a new cluster. When the group size is bigger than two, we merge the columns and rows of u_i and u_ρ creating a new user u'. u' is the addition of both users, u_i and u_ρ. Let us assume, that u_ξ is the most similar user with u'. Next, we include u_ξ to the cluster with u_i and u_ρ. The method executes this process $k-2$ times.

Algorithm 3 Algorithm for computing the clusters $Z = \{z_1,\dots,z_\gamma\}$ of users

```
Require: the set of users $U = \{u_1,\dots,u_n\}$
Require: the incidence matrix $M_{U\times U}$
Require: the clusters size $k$
Ensure: the clusters $Z = \{z_1,\dots,z_\gamma\}$ of users for $\gamma = \lceil n/k \rceil$
  $M'_{U\times U} \leftarrow M_{U\times U}$;
  $U' \leftarrow U$;
  while $|U'| \leq k$ do
    obtain the cluster $z$ of $k$ users using the Algorithm 4 and $M'_{U\times U}$;
    remove the users $u_i \in z$ form $U'$;
    remove the columns and the rows of the users $u_i \in z$ form $M'_{U\times U}$;
    add $z$ to the set $Z$;
  end while
  return $Z = \{z_1,\dots,z_\gamma\}$.
```

Algorithm 4 Algorithm for computing a cluster z of k users

```
Require: a incidence matrix $M'_{U\times U}$
Require: the clusters size $k$
Ensure: a cluster $z$ of $k$ users
  $z \leftarrow \emptyset$;
  obtain the two most similar users $(u_i, u_\rho)$, i.e. the cell of $M'_{U\times U}$ with the highest value;
  add $(u_i, u_\rho)$ to the set $z$;
  while $(|z| < k)$ and $(columns(M'_{U\times U}) > 0)$ do
    for each column $c_s \in M'_{U\times U}$ do
      $M'_{U\times U}(c_s, u_\rho) = M'_{U\times U}(c_s, u_\rho) + M'_{U\times U}(c_s, u_i)$;
    end for
    for each row $r_s \in M'_{U\times U}$ do
      $M'_{U\times U}(u_i, r_s) = M'_{U\times U}(u_i, r_s) + M'_{U\times U}(u_\rho, r_s)$;
    end for
    delete the column $u_\rho$ of matrix $M'_{U\times U}$;
    delete the row $u_\rho$ of matrix $M'_{U\times U}$;
    obtain the new $u_i$'s most similar user $u_\rho$, i.e. the cell of the user $u_i$ with the highest value;
    add $u_\rho$ to the set $z$;
  end while
  return $z$.
```

3.3. Aggregation

For every cluster z_j formed in the partition step, we compute its aggregation by selecting specific queries from each user in the group. That is, given the cluster of users $z_j = \{u_1, \ldots, u_k\}$, we obtain a new user u_{z_j} as the representative (or centroid) of the cluster, which summarizes the queries of all the users of the cluster. The selection of queries is based on the following principles:

1. We give priority to queries semantically close between them.
2. The number of queries a user contributes to the cluster representative is proportional to the number of queries of the user.

The first principle is considered in the partition step described in Section 3.2, since clusters are composed of users with semantically similar queries. The second principle is formalized defining some indexes as described below.

First, the number of queries of the centroid is the average of the number of queries of each user u_i of the cluster z_j. Then, the contribution of a user u_i ($Contrib_i$) to the centroid of a cluster with k users, depends on her number of queries $|Qi|$. This contribution is as follows:

$$Contrib_i = \frac{|Qi|}{\sum_{i=1}^{k} |Qi|} \tag{2}$$

Algorithm 5 Algorithm to aggregate the k users of the cluster z

```
Require: a cluster $z$ of $k$ users
Require: the quota $Quota_i$ of each user of the cluster $z$
Require: the contribution $Contrib_i$ of each user of the cluster $z$
Require: the set of queries $Q_i$ of each user of the cluster $z$
Require: the queries list $SL$
Require: the microagregged log $ML$
Ensure: the centroid of the cluster $z$
  $ML \leftarrow \emptyset$
  for each user $u_i \in z$ do
    $SL \leftarrow$ sort $Q_i = \{q_1^i, \ldots, q_{m_i}^i\}$ by query repetitions.
    while not reach $Quota_i$ do
      Add the first query $q_1^i$ with a probability $Contrib_i \times \#q_1^i\_repetitions$ to $ML$.
      Delete $q_1^i$ of $SL$.
    end while
  end for
  return $ML$.
```

Thus, the quota of each user u_i in the new centroid u_{z_j} can be computed as:

$$Quota_i = \frac{|Qi|}{k} \tag{3}$$

More formally, the aggregation method runs the Algorithm 5 for each cluster. First, it sorts logs from all users descending by query repetitions. Then, for each user u_i of the cluster and while not reaching $Quota_i$ do:

1. Add the first query of her sorted list with a probability $Contrib_i \times \#q_j\_repetitions$. For example, if u_i has a query repeated 3 times, and $Contrib_i$ is 0.4, as $3 \cdot 0.4 = 1.2$, the method adds one query to the new log and then randomly chooses to add it again or not according to the presence probability 0.2.
2. Delete the first query of the list.

4. Evaluation

We have tested our microaggregation method using real data from the AOL logs released in 2006, which correspond to the queries performed by 650 000 users over three months. We randomly select 1 000 users, which correspond to 55 666 lines of query logs. The usefulness evaluation and the results are presented below.

4.1. Usefulness evaluation method

For each user we have her original set of queries and the corresponding protected ones by means of our microaggregation method. All queries can be classified in categories, that is, each query is classified in the L first depth levels of the ODP.

In order to verify that our method preserves the usefulness of the data (i.e., does not introduce too much perturbation), we count the number of queries of each category, for a given level l, that are in the original log as well as in the centroid, ρ. This number is divided by the number of original queries in l, χ, obtaining a *semantic remain percentage* (*SRP*) in the level.

$$SRP = \frac{\rho}{\chi} \tag{4}$$

To summarize, our evaluation method does not only match two equal terms in both logs, but also a term in the protected log that replaces one with closest semantic in the original log. Using a random partition algorithm, users of each cluster might not be semantically close.

Consider, as an example of the worst case, a cluster of k users $\{u_1, \ldots, u_k\}$ with respective queries $Q = \{Q_1, \ldots, Q_k\}$, such that $Q_i \cap Q_j = \emptyset$ for all $i \neq j$. Thus, only the queries of a single user in a specific topic will appear in the centroid.

In this case, the number of queries of u_i that appear in the centroid can be calculated using formula 3 and it is known that the sum of all quotas is χ. Therefore, in the worst case when no common interests between users exists, we can calculate the average *SRP* as:

$$\frac{\sum_{i=1}^{k} \frac{|Q_i|}{\chi}}{k} = \frac{1}{k} \tag{5}$$

4.2. Results

As discussed in Section 2.2, ODP returns a list of categories for every term (or query), and each category is composed of various hierarchical levels. In our method, one or all categories can be used and, for each category, either all hierarchical levels or some of them can be considered. Intuitively, the more categories and levels (deeper levels) that are used, the higher the computational cost should be, and, perhaps, a better SRP can be achieved. Thus, we want to study how these parameters influence the SRP and the computational cost:

- ODP levels: every term has a categorization up to a hierarchical level, and the deepest level can be different for every term. The deeper the level is, the less terms that have information in this level there are. We want to know the deepest level that gives information for a majority of terms.

- SRP vs. ODP-categories: we want to know the SRP value when we use more or less categories; that is, if we use more categories, the SRP can be either higher, or have approximately the same SRP.
- Computational cost vs. ODP-categories: supposing that more categories are used, the higher the computational cost will be, but the extra cost should be known. If the extra cost is not significant and a better SRP is obtained, more categories can be used.

4.2.1. ODP levels

In the ODP, not all terms rank up to a certain level. For example, our working set of queries has terms with two levels (minimum) and others with twelve levels (maximum). In the study of the above mentioned relations (SRP vs. ODP-categories and computational-cost vs. ODP-levels), levels that do not have a ranking for the majority of terms can be ignored because such levels only give information to improve the SRP for a reduced number of terms. Thus, we consider a level if it has information for, at least, the 50% of the terms (queries).

In this sense, we have calculated for every level the percentage of terms that have a result for the level, and Figure 2 shows the percentage of queries (our working set of queries) that can be classified up to a certain depth level in the ODP tree. It can be observed that only 57% of queries can be classified up to the level 5. So, we only run tests up to this level.

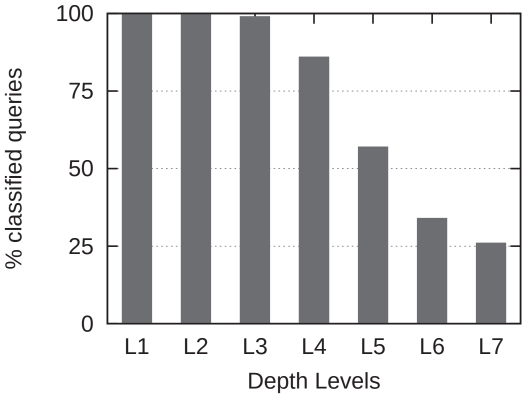

Figure 2: *Percentage of queries that can be classified up to a certain level in ODP.*

4.2.2. SRP vs. ODP-categories

Besides some initial tests (Erola *et al.*, 2010), we have calculated the percentage of semantically similar queries as the accumulation of the levels; that is, we add the coincidences of level 1 and 2 to calculate the percentage of semantically similar queries at level 2. In this current work, we have changed the evaluation method because we think that

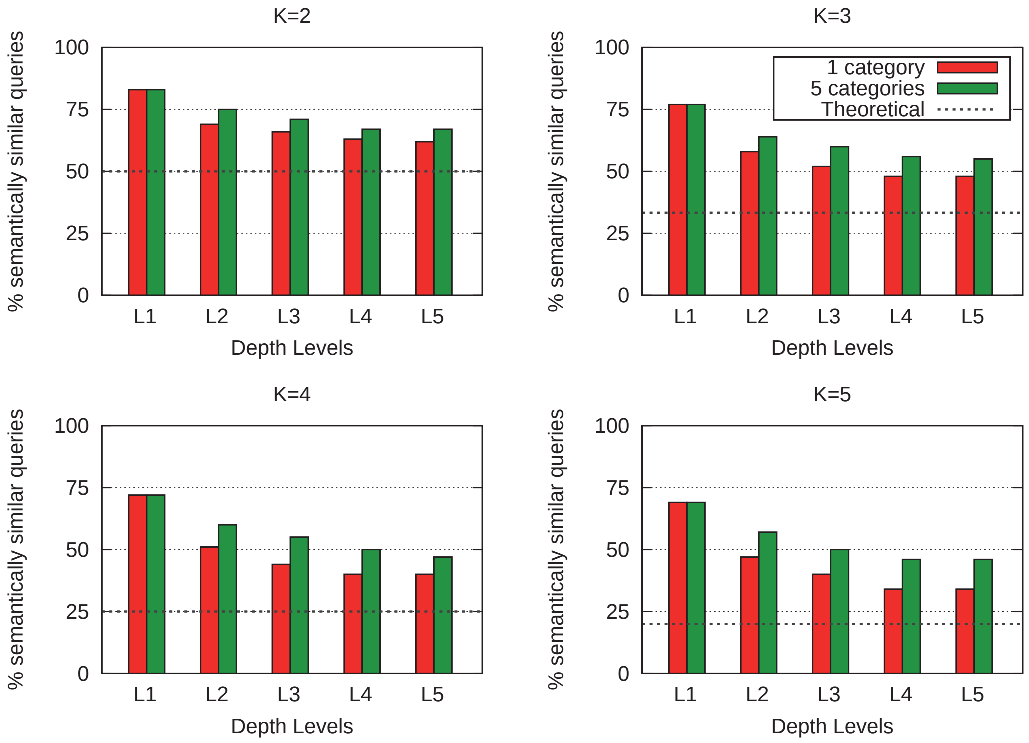

Figure 3: *Semantic similarity percentage of microaggregated logs using either the first category or the five first categories returned by the ODP.*

to evaluate each level separately is better to understand the remaining similarity of the queries in that level.

We have compared the results obtained (SRP) by either using the first five categories returned by the ODP or using only the first one. The range is enough in order to evaluate the SRP behaviour when we use more categories. Note that the first category that gives ODP is the most significative for the introduced term. Figure 3 shows, for cluster sizes 2, 3, 4 and 5, the average *SRP* that users obtain for various levels L. The red colour represents the obtained results using the first category returned by the ODP and the green colour represents the obtained results using the first five categories. It can be observed that both tests improve the theoretical *SRP* (see Section 4) with all depth levels. Using more categories in the ODP classification we achieve less similarity loss for deeper levels and larger cluster sizes. For instance, when $L = 1$, the same gain is obtained in all cases, but when $L = 5$ and $k = 5$, the difference gain is approximately 10% using the first five categories instead of only the first one.

4.2.3. Computational cost vs. ODP-categories

The computation cost is larger when more categories are used. Figure 4 shows the average time required to microaggregate logs for cluster sizes $k = 2, \ldots, 5$ for various levels. It can be determined that using the first five ODP categories, the average time is three times larger than using only the first one.

Tests were run on a Pentium Core 2 Duo 2.2Ghz without source code parallelization. Figure 4 demonstrates that the required time increases linearly with the number of user queries. Nonetheless, the program could be parallelized as follows:

- **Data preparation:** as each user has her queries, the classification matrices $M_{U\times C}$ can be computed simultaneously. Then, each cell of the incidence matrix $M_{U\times U}$ can be calculated independently, since we have available the classification matrix of each user.
- **Partition:** the partition process is linear and cannot be parallelized, but it is a negligible part of the whole process. The time required for its calculation is less than one percent of the total time.
- **Aggregation:** as users are divided into k groups, the logs' aggregation of each group can be run simultaneously.

Thus, the program parallelization could make the proposal scalable for very large systems.

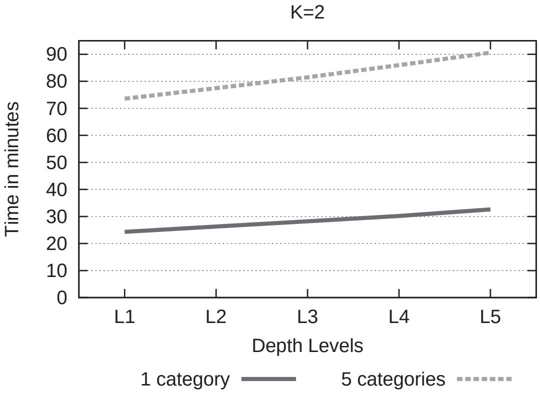

Figure 4: Average required time to microaggregate logs using our method for various ODP levels.

4.2.4. Considerations

It should be taken into consideration that we have repeated the tests of the previous initial work (Erola *et al.*, 2010) and we have observed that the results have improved because the ODP is constantly getting better. It now classifies more words. Furthermore, notice that we are working with a set of 1 000 users, randomly selected from the AOL files. We expect to achieve greater *SRP* values working with a larger set, because more similar users may be grouped.

It is important to remark that our proposal achieves k-anonymity (Samarati, 2001; Sweeney, 2002) at user level, which guarantees that at least k users are indistinguishable in the protected version. This guarantees a high degree of privacy, preventing the famous privacy leaks of the AOL logs.

To some readers our proposal might resemble agglomerative hierarchical clustering methods such as the well known Ward method (Ward, 1963). This method has been also adapted to perform microaggregation, although in another context, in Domingo-Ferrer and Mateo-Sanz (2002).

5. Related work

There are several approaches to anonymize query logs in the literature (Cooper, 2008), but they are normally reduced to the deletion of specific queries or logs. For instance, in (Adar, 2007) the authors propose a technique to remove infrequent queries, while in Poblete *et al.* (2008) a more sophisticated technique is introduced to remove selected queries to preserve an acceptable degree of privacy, or in the case of Korolova *et al.* (2009) to choose the publishable queries. Common techniques used in statistical disclosure control (SDC) have not been applied to this specific problem until very recently (Navarro-Arribas and Torra, 2009; Hong *et al.*, 2009; Navarro-Arribas *et al.*, in press, 2011). Moreover, these systems use spelling similarities to link users; that is, two users would be grouped if they had submitted syntactic similar queries. Therefore, they cannot distinguish different senses of a term, if it has more than one.

The use of supporting semantic taxonomies to anonymize query logs was considered in He and Naughton (2009) where the authors anonymize the set of queries made by a user by generalizing the queries using WordNet (Miller, 2009). WordNet is a generic lexical database of the English language, where concepts are interlinked by means of conceptual-semantic and lexical relations. The problem of relying on WordNet when facing the anonymization of query logs is that the query introduced by the user, despite the fact that they might not be in English, can be meaningless in a generic dictionary. We think that better results can be obtained for query logs by gathering semantic information from the Open Directory Project (ODP), which its main purpose is precisely to serve as a catalogue of the Web by providing a content-based categorization or classification of Web pages. This will be the case in general for data which is composed of uncommon words, which could not be found in WordNet. Note that if all words in the query logs were present in WordNet, the use of the WordNet framework will presumably give good results as well. Nevertheless, we need to introduce novel approaches to make the information obtained from the ODP useful. Unlike WordNet, which already has lots of published and tested distances functions, or aggregation operations, ODP lacks this extensive previous work.

6. Conclusions

The existing microaggregation techniques for query logs do not usually take into account the semantic proximity between users, which is negatively reflected in the usefulness of the resulting data. This paper presents a new microaggregation method for query logs based on a semantic clustering algorithm. We use ODP to classify the queries of all users and then aggregate the most semantically close logs. As we have seen, the resulting logs achieves higher usefulness while preserving *k*-anonymity.

We have tested our proposal with real query logs from AOL, showing some good results. Both in terms of information loss and in terms of protection, which is guaranteed because our method ensures *k*-anonymity at user level. As future work, new evaluation methods such as as Domingo-Ferrer and Solanas (2009), will be tested to better assess the quality of the results obtained using our system.

Acknowledgment

Partial support by the Spanish MICINN (projects eAEGIS TSI2007-65406-C03-02, TSI2007-65406-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process PT-430000-2010-31and N-KHRONOUS TIN2010-15764), the Spanish Ministry of Industry, Commerce and Tourism (projects eVerification TSI-020100-2009-720 and SeCloud TSI-020302-2010-153), and the Government of Catalonia (grant 2009 SGR 1135) is acknowledged. G. Navarro-Arribas enjoyed a Juan de la Cierva grant (JCI-2008-3162) from the Spanish MICINN.

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.

References

Adar, E. (2007). User 4xxxxx9: Anonymizing query logs. In *Query Logs workshop*.

Barbaro, M. and Zeller, T. (2006). A face is exposed for AOL searcher no. 4417749. The New York Times.

Cooper, A. (2008). A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Transactions on the Web*, 2.

Defays, D. and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada*, 195–204.

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, 189–201.

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

Domingo-Ferrer, J. and Solanas, A. (2009). Erratum: Erratum to “a measure of variance for hierarchical nominal attributes”. *Information Sciences*, 179, 3732. Elsevier Science Inc. New York.
http://dx.doi.org/10.1016/j.ins.2009.06.019.

EFF. (2009). AOL's massive data leak. Electronic Frontier Foundation. http://w2.eff.org/Privacy/AOL/.

Erola, A., Castellà-Roca, J., Navarro-Arribas, G. and Torra, V. (2010). Semantic microaggregation for the anonymization of query logs. In *Proceedings Privacy in Statistical Databases (PSD 2010)*, 6344 of LNCS, 127–137.

Frankowski, D., Cosley, D., Sen, S., Terveen, L. and Riedl, J. (2006). You are what you say: privacy risks of public mentions. In *Annual ACM Conference on Research and Development in Information Retrieval*, 565–572, Seattle Washington.

Gauch, S. and Speretta, M. (2004). Personalized search based on user search histories. In *Proceedings of International Conference of Knowledge Management-CIKM'04*, 622–628.

Google (2008). 2008 annual report. http://investor.google.com/order.html.

Hansell, S. (2006). Increasingly, Internet's data trail leads to court. The New York Times.

He, Y. and Naughton, J. (2009). Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, 2, 934–945.

Hong, Y., He, X., Vaidya, J., Adam, N. and Atluri, V. (2009). Effective anonymization of query logs. In *CIKM'09: Proceedings of the 18th ACM conference on Information and knowledge management*, 1465–1468.

Korolova, A., Kenthapadi, K., Mishra, N. and Ntoulas, A. (2009). Releasing search queries and clicks privately. In *WWW'09: Proceedings of the 18th international conference on World wide web*, 171–180.

Kumar, R., Novak, J., Pang, B. and Tomkins, A. (2007). On anonymizing query logs via token-based hashing. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, 629–638.

Miller, G. (2009). WordNet-about us. WordNet. Princeton University. http://wordnet.princeton.edu.

Mills, E. (2006). AOL sued over web search data release. CNET News. http://news.cnet.com/8301-10784_3-6119218-7.html.

Navarro-Arribas, G. and Torra, V. (2009). Tree-based microaggregation for the anonymization of search logs. In *WI-IAT'09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 155–158.

Navarro-Arribas, G., Torra, V., Erola, A. and Castellà-Roca, J. (in press, 2011). User *k*-anonymity for privacy preserving data mining of query logs. *Information Processing & Management*.
DOI:10.1016/j.ipm.2011.01.004

ODP. (2010). Open directory project. http://www.dmoz.org.

Oganian, A. and Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commision for Europe*, 18, 345–353.

Poblete, B., Spiliopoulou, M. and Baeza-Yates, R. (2008). Website privacy preservation for query log publishing. In *First International Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007)*, 80–96.

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

SearchEngineWatch. (2009). Global search market share, july 2009 vs. july 2008.
http://searchenginewatch.com/3634922.

SearchEngineWatch. (2010). Top search providers for september 2010.
http://searchenginewatch.com/3641456.

Soghoian, C. (2007). The problem of anonymous vanity searches. *I/S: A Journal of Law and Policy for the Information Society*, 3.

Summers, N. (2009). Walking the cyberbeat. Newsweek. http://www.newsweek.com/id/195621.

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

Torra, V. (2004). Microaggregation for categorical variables: a median based approach. In *Proceedings Privacy in Statistical Databases (PSD 2004)*, 3050 of LNCS, 162–174.

Torra, V. (2008). Constrained microaggregation: adding constraints for data editing. *Transactions on Data Privacy*, 1, 86–104.

Ward, J.H. (1963). Hierarchical Grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

Zetter, K. (2009). Yahoo issues takedown notice for spying price list. Wired. http://www.wired.com/threatlevel/2009/12/yahoo-spy-prices/#more-11725.

SORT Special issue: Privacy in statistical databases, 2011, 59-76

Eliminating small cells from census counts tables: empirical vs. design transition probabilities

Sarah Giessing[1] and Jörg Höhne[2]

Abstract

The software SAFE has been developed at the State Statistical Institute Berlin-Brandenburg and has been in regular use there for several years now. It involves an algorithm that yields a controlled cell frequency perturbation. When a microdata set has been protected by this method, any table which can be computed on the basis of this microdata set will not contain any small cells, e.g. cells with frequency counts 1 or 2. We compare empirically observed transition probabilities resulting from this pre-tabular method to transition matrices in the context of variants of microdata key based post-tabular random perturbation methods suggested in the literature, e.g. Shlomo, N., Young, C. (2008) and Fraser, B.,Wooton, J. (2006).

MSC: 62Q05 ”Statistical tables”

Keywords: Tabular data protection, Census Frequency Tables, SAFE, Post-tabular protection methods.

1. Introduction

In preparation for the German census 2011 we have started a comparative study of several perturbation methods for census frequency counts. The German Census will partly be register based, and partly be the outcome of a sample survey. This leads of course to limitations in the amount of detail of tables that can sensibly be released, as compared to a full census. Nevertheless, a huge amount of tabular output is going to be published. Publication of tables will to a major extent be pre-planned, but there will also be some flexible, user demand driven release of tabular data.

[1] Federal Statistical Office of Germany, 65180 Wiesbaden.

[2] State Statistical Institute Berlin-Brandenburg, 14467 Potsdam, Dortustraße 46.

Received: November 2010
Accepted: March 2011

Given the size of the publication, and other complexities (like non-nested hierarchies that are foreseen for some classification variables like “age”) non-perturbative methods like cell suppression do not seem to be a good choice: one of the issues to be raised here is that with cell suppression, there would be a considerable disclosure risk due to incomplete coordination of cell suppression patterns across tables. Perturbation methods also have the advantage that they introduce ambiguity into the zero cells which helps to avoid attribute disclosure when (nearly) all members of a population group score on only one (sensitive) category of a variable.

In this paper, we investigate into basically three alternative methods. The software SAFE (c.f. Höhne, J. (2003a), Höhne, J. (2003b)) is in regular use at the State Statistical Institute Berlin-Brandenburg. SAFE is an implementation of an algorithm that yields a controlled cell frequency perturbation. When a microdata set has been protected by this method, any table which can be computed on the basis of this microdata set will not contain any small cells, e.g. cells with frequency counts one or two. These small frequencies are the main concern for disclosure risk in Census counts tables, since they give information on the uniqueness or rareness of certain attributes or attribute combinations of individuals. Because SAFE is a pre-tabular method, all tables computed from the perturbed microdata set protected by SAFE are fully consistent and additive.

In comparison to SAFE, we study two post-tabular perturbation methods which both are based on the use of microdata keys. This technique can ensure full, or at least approximate, consistency of perturbations across different tables. Across table consistency has two aspects: On one hand, inconsistencies may be irritating to users. More severe from the disclosure control point of view is that inconsistency may lead to disclosure risk. For example, an average taken over eventually inconsistently perturbed values of logically identical cells (taken from different tables) should not be an unbiased estimate of the original cell value.

Each of the two post-tabular methods involves two steps. The first step yields fully or approximately consistently perturbed, but non-additive tables. Non-additivity is a potential nuisance for users, and may also be a source of disclosure risk. Therefore, in a second step, table additivity should be restored. This can be achieved by statistical methods such as the iterative proportional fitting algorithm. In this paper we discuss using linear optimization techniques for this step.

In order to avoid a perception of disclosure risk, and to provide a “visible” kind of protection, we require both methods to provide, like SAFE, perturbed data without small cells (i.e. without counts of one and two). Note that we imagine a rather naïve use of the data here, keeping in mind that for researchers there will be other options of accessing the data.

This paper reports on findings of the first phase of the study when implementing the methodologies. It is organized in eight sections: In Section 2, we outline the methodological approach of SAFE. Technical issues of constructing suitable probability transition matrices for random perturbation methods are discussed in Sections 3 and 4. In Section 5, we suggest an optimization technique to restore table-additivity, e.g.

the CTA method of Castro, J., González, J.A. (2009). Some test results are presented in Section 6, and a measure of information loss on the cell level for SAFE results is proposed in Section 7. We conclude the paper with a brief summary Section 7.

2. Methodological background of SAFE

In this section we briefly describe the methodological approach of SAFE, as far as it is relevant for an application to protect tabulations of population Census counts data. Starting point for the method is a microdata file where all variables are recoded to give the highest degree of detail foreseen for any publication. Imagine a variable like age, where perhaps data are collected so that for each person age could be deduced down to the level of age in months, but publications should offer data at most by age in years. Then the variable would be recoded to the level of age in years. We also assume here that the data set consists of categorical variables only.

The basic idea of the method is to turn this data set (with, say, N variables at n_i $(i = 1, \ldots, N)$ categories) into a data set, in which either none of the records, or at least three records score on each of the $n_1 * n_2 * \cdots * n_N$ theoretical combinations of categories.

With respect to data quality, the method aims to preserve as far as possible cell counts in a pre-defined set of ‘controlled’ tables. For those tables, the method yields results that are in some sense ‘optimal’. If any other table is derived from the perturbed data set, it will be safe (i.e. it will not contain any ones or twos), but differences between original counts and those computed on basis of the perturbed data set can be much larger than they arise for the controlled tables. The experience is that the method is usually able to achieve a maximum deviation between 4 and 8 for a sensibly defined set of controlled tables.

2.1. The SAFE mathematical model

The algorithm computes a heuristic solution for the problem of minimizing the maximum absolute deviation between true and perturbed cell values in the controlled tables.

Instances are defined by the following parameters:

- A set of linear relations $Ay = a$ defining the table cells of the controlled tables as sums of cells of an elementary table consisting of all combinations of categories of all variables in the microdata set.
- Vector $a, a = (a_i,\, i \in I)$ denote the original frequencies presented in the controlled tables, and vector $y, y = (y_j,\, j = 1, \ldots, N)$ the entries (e.g. frequencies of category combinations) of the cells of the elementary table. In a valid solution, vector y does not contain any entries of 1 or 2.

- Vector w of weights associated to perturbations of the table cells of the controlled tables. For example, we may want to allow larger perturbations for larger cells, or avoid them for cells that are rated "highly important".

The objective of the model is to minimize the maximum entry of vector $d = (d_i,\, i \in I)$, $d_i \in \mathbb{Z}$, denoting the deviations of original and perturbed cell counts in the controlled tables. With these definitions, broadly, the model is as follows:

Solve the problem

$$\begin{array}{l} \min_y \quad \max_{i \in I} \left(|d_i| + w_i\right) \\ \textit{subject to} \quad Ay = a + d \\ \qquad y_j \in \{0,3,4,5,\dots\} \quad j = 1,\dots,N \end{array} \tag{1}$$

This statement of the problem resembles a huge non-linear integer optimization problem which is computationally intractable[1]. Therefore, an efficient heuristic algorithm has been developed that gives near optimal solutions at reasonable expense of computer resources.

Beginning with the (infeasible) initial solution given by $d = 0$, i.e. where cell values are kept at their original value, a first feasible solution is obtained. This solution is optimized later on.

A first feasible solution

In addition to the above parameters, we define now

- Vector $b = (b_i,\, i \in I)$, $b_i \in B$ of bounds for maximum allowed deviations. In practice, B consists of two values only, one stating the maximum deviation to be allowed for cells defined by only one variable, the other one stating the maximum allowed deviation for the other cells, e.g. cells defined as cross-combination of categories of two or more variables,
- Vector $x = (x_j,\, j = 1,\dots,N)$, $x_j \in \{0,1\}$ is 1, if elementary table cell j is "unsafe", e.g. if $y_j \in \{1,2\}$ and 0 otherwise.

The problem to be solved is

$$\begin{array}{l} \min_y \quad \sum_{j=1,\dots,N} x_j \\ \textit{subject to} \quad |Ay - a| < w + b \\ \quad y_j \in \{0,1,2,3,\dots\} \\ x_j = 1;\ \textit{if}\, y_j \in \{1,2\} \quad x_j = 0; if\ y_j \notin \{1,2\} \end{array} \tag{2}$$

[1] Note, in our test setting which is still substantially smaller than the real setting will be, the size of vector y (and thus the number of columns of matrix A) is approximately $N = (2*4*7*8*111*10000) \sim 5*10^8$.

A feasible solution is obtained when the objective function is zero.

Minimizing the number of "unsafe" frequencies, using a heuristic, the algorithm step by step changes critical frequencies of 1 and 2 into uncritical frequencies 0,3,4,... If the process stagnates, the statement of the problem is modified automatically by increasing the vector of bounds b, e.g. $b = b + 1$.

Optimizing the solution

Once a feasible solution has been obtained, the method will seek to improve the solution by reducing the maximum allowed perturbation, e.g. b and eventually w. Usually, the number of cells where the deviation is identical or near-identical to the respective bound is relatively small. In the optimization step, after changing (decreasing) b or w, some of the constraints in model (2) will be violated. Accordingly, we define now

- Vector $z = (z_i, i \in I)$, $z_i \in \{0,1\}$ is 1, if for controlled tables cell i the bound constraint of model (2) is violated and 0 otherwise.

The algorithm derives a heuristic solution to

$$\begin{aligned} &\min_y \quad \sum_{i \in I} z_i \\ &\textit{subject to} \quad |Ay - a| - (w + b) < z \\ &\quad y_j \in \{0, 3, 4 \ldots\} \end{aligned} \tag{3}$$

If a solution is obtained where $\sum_{i \in I} z_i = 0$, the constraints will be tightened further (e.g. decrease b or w), and model (3) will be solved again. This step is repeated until either an expected level of optimality (in the bounds) is reached, or further attempts seem to be rather unpromising.

3. Generating random noise for frequency tables

The Australian Bureau of Statistics Fraser, B., Wooton, J. (2006), Leaver, V. (2009) has developed a concept for a cell perturbation method. They propose that the random noise should have zero-mean and a fixed variance. An alternative cell perturbation method referred to as "Invariant Post-tabular SDL" method was suggested in Shlomo, N., Young, C. (2008). In the following two subsections we briefly outline the two alternative concepts and discuss the technical construction of suitable probability transition matrices for a random perturbation eliminating all small frequency counts.

3.1. How to create zero-mean/fixed variance cell perturbations?

Fraser, B., Wooton, J. (2006) propose to generate for each cell c with non-zero cell count i_c an independent integer value perturbation d_c satisfying the following two criteria:

(a) mean of zero

(b) fixed variance V for all cells c and all frequency counts i

A third criterion, in order to meet the requirement that perturbed cells do not have a count of one or two, would be

(c) $i_c + d_c \notin \{1, 2\}$ f.a. i_c, d_c

This means we look for a $L \times L$ transition matrix $\mathbf{P}$[2] containing conditional probabilities: $p_{ij} = p$ (perturbed cell value is $j \mid$ original cell value is $i)$ with the following properties:

(1) $p_i v_i = 0$

(2) $p_i (v_i)^2 = V$

(3) $p_{ij} = 0$ for $jin\{1,2\}$

(4) $\sum_j p_{ij} = 1$

(5) $p_{ij} = 0$; if $j < i - D$ or $j > i + D$,

(6) $p_{00} = 1$ and $p_{0j} = 0$ for $j > 0$, and of course

(7) $0 \leq p_{ij} \leq 1$

where p_i denote the ith row-vector of matrix $\boldsymbol{P}$ and v_i a column vector of the noise which is added, if an original value of i is turned into a value of j. I.e. the j^{th} entry of v_i is $(j - i)$. For example $v_i = (-1, 0, 1, 2, 3, \ldots, L - 2)$. (1) is equivalent to (a) and expresses the requirement that the expected value of the noise should be zero. Similarly, (2) is equivalent to (b), expressing the requirement of a constant variance, and (3) relates to (c). (4) and (7) are of course necessary for any Transition matrix, (5) states a maximum allowed absolute perturbation of some pre-defined constant D and (6) states that zero frequencies must not change. Note for all rows after row $D+2$, condition (3) is always satisfied, when (5) holds. Hence we can facilitate the task of computing suitable transition probabilities by adding a symmetry requirement for all rows after row $D+2$:

(8) $p_{i,i-k} = p_{i,i+k}$ for $k = 1, \ldots, D$, if $i > D + 2$

[2] As index j may take a value of zero (when a cell value is changed to zero), in the following we start counting matrix and vector indices at 0, enumerating rows and columns of the $L \times L$ matrix by $0, 1, 2, \ldots, L - 1$.

With (8), condition (1) is always satisfied because the negative and positive deviations balance each other. (2) simplifies into

$$\text{(2a)}\quad 2 \sum_{j=1,\ldots,D} p_{ij}\, j^2 = V.$$

For simplicity, in the following we therefore assume $L-1 = D+3$, applying the perturbation probabilities given by row $(D+3)$ of matrix $\boldsymbol{P}$ to all cell counts $\geq D+3$.

For every row (or cell count) i $(i = 1,\ldots,D+2)$ conditions (1) to (5) can be rewritten as system of three linear equations

$$\text{(9)}\quad \boldsymbol{A}_{iD}\, x = b,$$

where $\boldsymbol{A}_{iD}$ is a $(3 \times (\min(i,D)+1+D-k))$ [3] coefficient matrix and $b = (1,0,V)'$. The elements of x correspond to the entries of row i in P which are not zero anyway by definition (because of (3) or (5)). The first row of $\boldsymbol{A}_{iD}$ corresponds to condition (4), the second row to (1) (e.g. unbiasedness) and the third row to (2) (fixed variance V).

Consider for example $\boldsymbol{A}_{13} = \left\{ \begin{array}{rrr} 1 & 1 & 1 \\ -1 & 2 & 3 \\ 1 & 4 & 9 \end{array} \right\}$. In this simple case, the coefficient matrix is invertible. The last row of the inverse is $(-1/2, -1/4, 1/4)$. Hence, in order for p_{13} to be positive, $(-1/2, -1/4, 1/4) \cdot b = (-1/2 + V/4)$ must be positive, and hence V must be at least 2. In this case (9) has a unique solution, depending on the choice of V only. If V is exactly 2, p_{13} is zero.

In general, $\boldsymbol{A}_{iD}$ has more columns than rows. So usually, there is no unique solution for (9). But we can use (9) to derive feasibility intervals for x (e.g. for the p_{ij}). A practical approach is to fix V to $2+\varepsilon$ with a small positive value for ε (increasing ε and hence the variance of the perturbation leads to an unnecessary loss of information). The system (9) can be further strengthened by additional constraints, for example to express desirable monotony properties like $p_{ij} \geq p_{i,j+1}$ for $j > i$, or to improve symmetry by bounding the difference between $p_{i,i-1}$ and $p_{i,i+1}$.

We have experimented with $D = 3$, 4 and 5. One of the findings was that for small D and i, the linear programming problem derived from (9) (eventually together with the additional constraints) gives quite small intervals for x. For larger D and i the intervals for x are wider. In those cases we first fixed a value (like 70 %) for the centre of the distribution, p_{ii}. Afterwards we fitted each tail of the distribution p_{ij}, $j > i$ and p_{ij}, $j < i$ to the tails of a normal distribution using a simple heuristic approach:

At first, provisionally fix one (say, the left-hand) tail of the distribution. This gives a target total probability and target total variance for the right-hand tail (through subtracting the corresponding left hand tail values from differencing one (V, resp.)).

[3] k is the number of elements in $\{1,2\} \cap [i-D;\, i+D]$.

Then approximate p_{ij} $(j > i)$ by $F_{k+0.5+i} - F_{k-0.5+i}$, where F_x denote the Normal distribution with zero expectation and suitable Variance σ^2 at x, and k denote the starting point of the distribution tail. The starting point k should be selected as to achieve that the approximate p_i, D_{+i} is about zero. See Gießing, S., Höhne, J., (2010) (Appendix), for further details, and how to obtain a suitable variance parameter.

The corrected approximate $p_{i,i+j}$ distribution can then be used to derive the target values for a corrected total probability and variance of the left-hand tail. Carry out the procedure described for the right hand tail for the left hand tail now. Finally, feed back the corrected approximate p_{ij} into the system (9) and (by minimizing or maximizing one of the variables) obtain a final distribution which meets the requirements of (9) with sufficient precision.

Table 1 in the appendix shows the final probability matrices for $D = 3$, 4 and 5, e.g. the design transition probabilities and compares them to the transition probabilities observed empirically for the cells of the set of controlled tables after protecting the data by SAFE. Obviously, the SAFE method results in much smaller probabilities that cell values change by less than three.

3.2. Combination of invariance and a "no-small-cells" requirement?

The idea of the "Invariant Post-tabular SDL" method Shlomo, N., Young, C. (2008) is to preserve the frequency distribution of the cell counts. But in our setting we require the frequency of perturbed small counts (ones and twos) to be zero. So for the small counts these are aims that clearly exclude each other. A possible way out would be to relax the goal of invariance. E.g. only seek to preserve the frequency distribution of cell counts above three and the total frequency of all cell counts below four. This can be achieved as follows:

As shown in Shlomo, N., Young, C. (2008), an invariant matrix $\boldsymbol{R}$ is obtained by multiplying some pre-defined initial transition matrix $\boldsymbol{P}$ (for an example see Shlomo, N., Young, C. (2008)) with a suitable matrix $\boldsymbol{Q}$. $\boldsymbol{Q}$ is obtained by transposing matrix $\boldsymbol{P}$, multiplying each column j by the relative frequency of count j and then normalizing its rows so that the sum of each row equals one. Finally the diagonal elements of this matrix are increased by the following transformation $\boldsymbol{R*} = \alpha R + (1 - \alpha) I$, where I is the identity matrix of the appropriate size.

Gießing, S., Höhne, J., (2010) explain how to adapt this procedure to the "no-small-cells" requirement. In a first stage, an invariant matrix $\boldsymbol{R*}$ is computed such that the first row gives the joint transition probabilities of all counts under four, and the first column gives the probabilities for changing a given count into a count smaller than four. The procedure to obtain $\boldsymbol{R*}$ is the same as in Shlomo, N., Young, C. (2008), except that here we use a vector of relative frequencies, where the entries corresponding to the ones, twos and threes are added up to one joint entry v_{1-3}. We also replace the first row of the initial transition matrix by a column vector where all entries except for the first two are zero. See Gießing, S., Höhne, J., (2010) for details of how to compute the first two entries of this vector, and on how to compute separate transition probabilities for counts under four. Finally, we replace the first line of ***R**** by the separate transition probabilities for counts under four (and attach three columns of zeros to the other lines). This way we get a transition matrix ***R*****, which is almost invariant, except that for counts under four only their total frequency is preserved. For illustration, in the following we present an example using real data of a table of the last West German census of 1987.

Example 1:

For a census table with frequencies $(V_1, V_2, V_3, V_4, V_5, \ldots) = (96, 32, 20, 16, 15, \ldots)$ observed for counts (1,2,3,4,5,…), we computed an initial invariant matrix ***R**** (with $D = 2$). Table 2 shows the first four rows and six columns of the matrix of expected frequencies obtained from $(V_{1-3}, V_4, V_5, V_6, V_7, \ldots) \cdot \boldsymbol{R^*}$

Table 2: *Expected frequencies $n_{i,j}$ of counts of i perturbed into counts of j.*

	0-3	4	5	6	7	8
0-3	144.62403	2.9690406	0.4069246	0	0	0
4	2.9690406	11.81552	1.0893785	0.1260606	0	0
5	0.4069246	1.0893785	12.211631	1.196397	0.0956693	
6	0	0.1260606	1.196397	10.676843	0.9239783	0.0767213

Table 3 below shows the first six rows and six columns of the matrix of expected frequencies computed as $(V_1, V_2, V_3, V_4, V_5, \ldots) \cdot \boldsymbol{R^{**}}$. The sum of the first two column totals in Table 3 (regarding $j = 0.3$) is 148, e.g. the total observed frequency of the counts under 4 $(= 96 + 32 + 20)$ is exactly preserved.

Table 3: *Expected frequencies $n_{i,j}$ of counts of i perturbed into counts of j for example 1.*

	0	3	4	5	6
1	60.531974	35.468026			
2	9.510658	22.489342			
3	4.6240348	12	2.9690406	0.4069246	0
4		2.9690406	11.81552	1.0893785	0.1260606
5		0.4069246	1.0893785	12.211631	1.196397
6		0	0.1260606	1.196397	10.676843

Note that apart from this introductive example, we did not carry out further testing of this method. The concept of preserving frequencies for each individual cell count is not too convincing when the expected use of the data is a rather naïve one[4]. An exception would be a situation where the frequencies for individual cell counts are a statistic of interest for the user. Such an application is outlined in Section 7.

[4] Note that there will be options for researchers to access the original data via research data centres.

4. Selection of random noise

The random mechanism proposed in Fraser, B., Wooton, J., (2006) can be implemented very easily: For our experiments, we used the SAS random number generator which produces pseudo random numbers distributed uniformly over $[0;2^{31}-1]$. We assign such a random key to each record in the microdata file. When computing the tables, also the random keys are aggregated. The result is then transformed back into a random number on this interval by applying the modulo function, e.g. $\mathrm{mod}_{2^{31}-1}$. If the same group of respondents is aggregated into a cell, the resulting random key will always be the same. Cells which are logically identical thus have identical random keys.

Then we simply use a transition matrix computed to give zero-mean / fixed variance noise (as explained in 3.1), compute cumulated probabilities (for each row) and multiply the resulting matrix by $2^{31}-1$. Denoting the entries of this matrix by $\boldsymbol{M}_{ij}$ we change a cell count of i of some cell c into j, if the random key of cell c is between $\boldsymbol{M}_{i,j-1}$ and $\boldsymbol{M}_{ij}$. This will guarantee that the expected values of the perturbed counts are identical to the original counts (unbiasedness) and lead to consistently perturbed data. However, for a given table, the mean perturbation of cells of a given frequency count i is not necessarily zero. This mean will depend on the actual distribution of the corresponding record keys. For the data of example 1 above the observed difference between a true cell count and the mean of the corresponding perturbed counts varies between -0.82 and 0.78.

See Section 4.1 of Gießing, S., Höhne, J., (2010) for some special issues regarding an appropriate selection procedure in the context of the invariant post tabular method.

5. How to restore table-additivity?

Non-additivity is a potential nuisance for users, and may also be source of some disclosure risk. As simple example, assume random noise with a maximum perturbation of two has been applied. Assume two cells with original count one are perturbed to count three, and the original total of two is perturbed to zero. Users are informed on the maximum perturbation. Hence they know that both inner cells must have original count one at least. But if any of them were greater then one, the original total would be at least three and could not have turned into a perturbed value of zero.

This kind of disclosure risk typically arises, when all inner cells are all perturbed in the same direction, each with the maximum possible perturbation, and the total cell is perturbed in the other direction, also with the maximum possible deviation. With perturbations based on transition matrices like the ones discussed in Section 3 with usually small probabilities on the tails these events will be relatively rare. However, we should also bear in mind, that this is only the simplest kind of attack. A systematic analysis based on linear optimization techniques and taking into account the aggregate structure of a perturbed non-additive multidimensional table with a published maximum perturbation might eventually break other perturbation patterns as well.

Restoring table additivity, as suggested in Fraser, B., Wooton, J. (2006) and Shlomo, N., Young, C., (2008) is considered there an integral part of the method. Leaver, V. (2009) and Shlomo, N., Young, C., (2008) point out that restoring additivity can be achieved by iterative methods. As an alternative, we suggest to consider a linear programming based method like Controlled Tabular Adjustement (see f.i. Dandekar, R.H., Cox, L. (2002), Castro, J. (2006)).

For a first experiment, we use the CTA implementation of Castro, J., González, J.A. (2009)[5]. The algorithm restores additivity to a table, minimizing an overall distance to the table provided as input. The distance function implemented is a weighted sum of absolute per-cell-distances. Weights are provided by the user of the software. The user can define for each cell upper and lower bounds on the deviations, and can define a set of cells labeled as '*sensitive cells*'. Sensitive cells are forced to change their values. For each sensitive cell, the user defines a '*protection interval*'. The adjusted cell value is not allowed to take a value within the protection interval.

Computational complexity of the problem depends strongly on the number of sensitive cells. In a first experiment, we therefore use a two stage approach: in a first CTA run, we only restore additivity to the table. Although in this step we assign cell weights which will avoid to some extent that the algorithm adjusts cell counts of zero[6] or three, we will usually get an adjusted table with some small cell counts (e.g. ones and twos). In a refinement run, we define these ones and twos as sensitive, and define the corresponding protection interval as the interval (0;3). At the same time, for all cells with counts greater or equal to three we defined a lower bound of at least three. For all cells with zero count, the upper bound is zero. This way, however, we run a certain risk of defining an infeasible problem, especially if we define at the same time rather narrow bounds for the non-sensitive cells. See Section 6 for a test result.

Because the adjustment cannot simultaneously take into account all tables ever to be released[7], it introduces inconsistencies in the perturbation. Identical cells, even if they received the same perturbation by the random process, may become adjusted to different values. This fact leads to some risk that some perturbations might be undone, if intruders run an LP-based analysis taking into account the aggregate structure across several tables. But this is not such an easy task, on one hand, and on the other hand, it may not be very successful, because it may happen that only original frequencies can be broken that do not cause disclosure risk.

Of course one might consider using the adjustment methodology without previous random perturbation, only to 'remove' cells with small counts from the table. But as long as this does not – unlike the SAFE method – yield a fully consistent data base, there is then a risk that by averaging cell values over a number of tables a user can recover the

[5] See Castro (2011) for an extension of the methodology.

[6] Note that we do not allow original zero cell counts to be adjusted.

[7] (This would be a problem similar to the on solved by SAFE, c.f. 2 (in particular in size) and too huge for today computational resources).

original data. With a previous random perturbation, such an approach will only recover the underlying perturbed table, as pointed out in Leaver, V., (2009).

6. Some test results

Table 1 in the appendix shows the probability matrices we computed for the zero mean/fixed variance noise approach when the maximum allowed deviations are $D =$ 3, 4 and 5 respectively, and compares them to the transition probabilities observed empirically for the cells of the set of controlled tables after protecting the data by SAFE. Obviously, the SAFE method results in much smaller probabilities that cell values change by less than three.

For all counts after $D+3$, in our implementation of the stochastic noise, transition probabilities are defined identical to those obtained for $D+3$. Figure 1a below shows the empirical SAFE probabilities for counts i to change by d for counts between 9 and 16 in the set of controlled tables, compared to the transition probabilities of the stochastic noise obtained for $D = 5$. Figure 1b shows those probabilities for counts grouped into count size classes observed for cells that are not in the controlled tables. For our experiment we defined as control tables only tables defined by cross-combination of at most 3 variables. The results presented in Figure 1b on the other hand relate to cells defined by cross-combination of 4 variables.

As can be seen in Figure 1a, the SAFE probabilities become approximately normal when the cell count increases. It is also very clear that the SAFE perturbation is stronger than that of our stochastic noise implementation: Compare f.i. the probability of no change (i.e. at $d = 0$) which is about 70 % for the stochastic noise, but between 13 % and 26 % for SAFE. However, the difference matters mainly for the small perturbations, and hence will matter more for smaller counts.

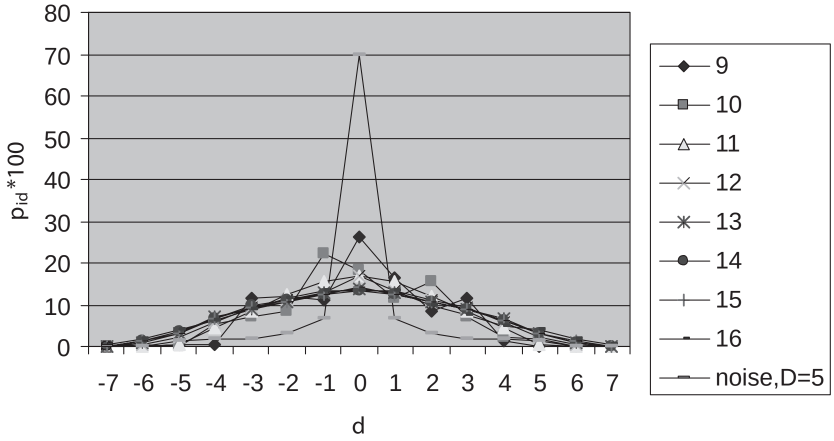

Figure 1a: *SAFE vs. stochastic noise transition probabilities.*

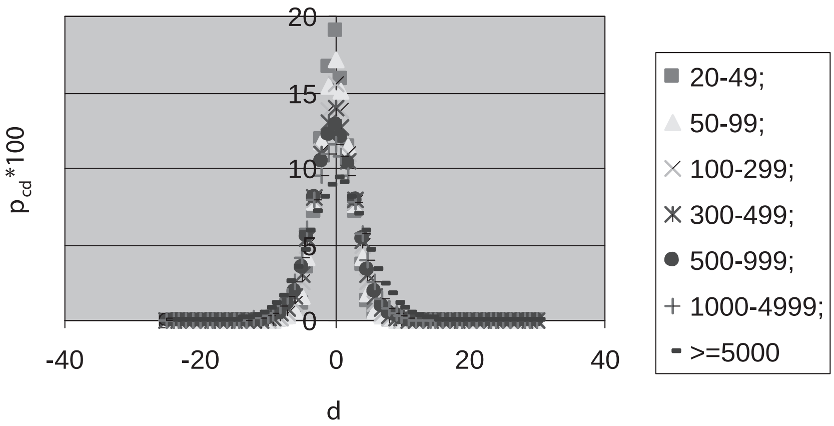

Figure 1b: *SAFE, transition probabilities for banded counts in non-control tables. probabilities.*

Figure 1b shows that also for cells that are not contained in the controlled tables, the deviations resulting from SAFE are still normally distributed, but the tails of the distribution are longer. While we got a maximum deviation between true and perturbed count of 7 for controlled tables cells, deviations of up to 30 occurred in the set of 4-dimensional cells, as can be seen in table 4 presenting the maximum observed deviations for the cell count size classes of Figure 1b.

Table 4: *SAFE, maximum observed deviations D for non-control table by cell count size class.*

Counts	**20-49**	**50-99**	**100-299**	**300-499**	**500-999**	**1000-4999**	**≥ 5000**
D	16	19	25	21	30	24	26

Considering that table additivity is a very important issue, it makes sense to compare SAFE transition probabilities not only to the design transition probabilities of stochastic noise, but also to the noisy tables after restoring additivity. We have applied the approach of Section 5 for restoring table additivity using CTA to a 3-dimensional test table. The table has been perturbed using the design transition probabilities displayed in Table 1 (appendix). For this instance we obtained adjusted tables where the maximum perturbation of cell counts is identical before and after the adjustment. This is certainly encouraging, but it seems unlikely that it is a general result. Table 5 compares the noisy

Table 5: *Distribution of 22670 non-zero test table cells by absolute deviations to true cell values.*

Abs. Dev.	**SAFE**	**noiseD3 adj.**	**noiseD3**	**noiseD4 adj.**	**noiseD4**	**noiseD5 adj.**	**noiseD5**
0	12.88	23.17	29.44	29.44	38.99	30.62	40.18
1	44.62	44.05	40.23	39.87	34.08	40.05	34.63
2	27.68	23.31	22.31	21.01	19.20	20.41	18.48
3	11.09	9.47	8.01	5.86	4.31	5.07	3.33
4	3.16			3.82	3.42	2.38	1.92
5	0.48					1.47	1.46
6	0.10						

tables before and after restoring additivity to those computed with SAFE protected data. It presents the frequency distribution of the 22670 non-zero cells of the example by absolute deviations between true and perturbed values.

For this example, we observed a mean-deviation between true and perturbed values of 1.49 for SAFE. For stochastic noise at $D = 3$, 4 and 5 we got mean deviations of 1.09, 0.99 and 0.97, resp., and after restoring additivity 1.19 ($D = 3$), 1.15 ($D = 4$) and 1.13 ($D = 5$). Obviously, in this example, even after restoring additivity, stochastic noise outperforms SAFE. On the other hand, the experiment also shows that – at least when we use the methodology of Section 5 not allowing that new small cells appear in the adjusted tables – restoring additivity tends to increase deviations (for example the mean deviation for $D = 5$-noise from 0.97 to 1.13)[8]. It has to be expected that this effect increases with increasing size of the tables where additivity has to be restored. The computationally expensive second CTA step[9] required between about 6 and 24 minutes. As it is intended that table generation for the Census results should be an OnLine process, this is certainly too long. Even, if this issue could be solved, before such an approach could be put into practice, a lot of experimentation would be necessary, for example to determine "sustainable" parameters for the initial random perturbation in the sense that the adjustment process can preserve to some extent the properties of the random perturbation (like f.i. the maximum perturbation).

7. Data utility – a cell level measure of information loss

Probably, many users of census counts data do not use them for complex statistical analyses, but are merely interested in learning simple facts, like 'how many people with properties X live in area Y?'. When those counts are perturbed, they should be informed how reliable each individual cell is. This is especially important, if a perturbation method may produce fairly large perturbations, although only for a very small portion of the cells, which can f.i. be the case for SAFE for cells which do not belong to the set of controlled tables.

A simple information loss measure on the cell level could be given by publishing along with the perturbed counts the absolute value of the perturbation. However, this may be too much information, leading to disclosure risk. Instead, one might publish the absolute value of a perturbed version of the perturbation.

Usually, to inform about data utility, one publishes information on the perturbation on the table level, like the frequency distribution of the noise (c.f. Table 5). Therefore, when perturbing the perturbations, it makes sense seeking to preserve these frequencies. E.g. use an invariant matrix of transition probabilities for perturbing the perturbations

[8] Note that these findings may not apply to all additivity methods.

[9] The first step which only restores additivity to the table takes just a few seconds for this instance.

of the original counts in a table. Generating such a transition matrix is a straightforward application of Shlomo, N., Young, C. (2008). The only difference is that, unlike the original counts which are positive numbers, the perturbations take values between $-D$ and D. Table 6 shows the results of an application to table Region x Age x Country_of_Birth[10]. The observed frequencies of the perturbed SAFE-deviations (n_d*) match the frequencies of the unperturbed SAFE-deviations (n_d) nearly exactly.

Table 6: *Number of cells of a test table by deviation of the SAFE protected results: true frequencies* (n_d) *vs. frequencies after invariant perturbation of observed deviations* (n_d*).

Cells with negative deviation *d*				**Cells with positive deviation *d***			
d	n_d	n_d*	$n_d - n_d*$	***d***	n_d	n_d*	$n_d - n_d*$
−13	1	0	1	13	0	0	0
−12	5	5	0	12	7	6	1
−11	30	31	−1	11	44	45	−1
−10	110	110	0	10	108	108	0
−9	310	309	1	9	372	370	2
−8	836	837	−1	8	878	879	−1
−7	1872	1871	1	7	2141	2141	0
−6	8203	8204	−1	6	9230	9231	−1
−5	34859	34859	0	5	37674	37675	−1
−4	162116	162115	1	4	170659	170657	2
−3	369234	369234	0	3	393652	393654	−2
−2	622462	622464	−2	2	778735	778735	0
−1	1226831	1226831	0	1	783760	783758	2
0	739905	739905	0	0	739905	739905	0

8. Summary and final remarks

In preparation for a comparative study of several perturbation methods for census tabular frequency data, in this paper we have raised some practical issues regarding the implementation of two alternative approaches explained in literature. In particular, this paper has discussed in some detail how to construct zero-mean/fixed variance transition matrices required to implement the methodology of Fraser, B., Wooton, J. (2006). We also discuss an extension of an idea of an invariant transition matrix suggested in Shlomo, N., Young, C. (2008) to a situation where the perturbation procedure should eliminate small cells.

As pointed out in Fraser, B., Wooton, J. (2006) and Shlomo, N., Young, C. (2008), additivity is not preserved by the post-tabular random perturbation method, but can be restored afterwards – however, at the expense of between tables consistency. We have

[10] Note that variable Country_of_Birth has been defined here to involve one category which defines an extra-subtotal not contained in the set of cells defined by the set of controlled tables. Therefore, SAFE perturbs some cells of this table by more than the control-tables maximum of 7.

outlined and tested on a small instance an approach based on linear optimization, e.g. CTA methodology.

Leaving a larger scale empirical comparison of the post-tabular methods discussed in the paper with the pre-tabular perturbation method SAFE outlined in Section 2 for the future, the paper provides evidence that the post-tabular methods as implemented here tend to result in smaller changes to the data than SAFE. On the other hand, as a pre-tabular method, SAFE preserves additivity and consistency, is easier to implement in a flexible OnLine table generation environment, and is able to keep the maximum deviations in a set of pre-specified tables acceptably small. These are important properties and may be worth “less optimal” performance regarding data quality to some degree. While the perturbation caused by SAFE tends to be stronger than those caused by a non-additive post-tabular approach, the paper shows that they tend to be normally distributed, e.g. large deviations are relatively unlikely, also for cells that are not contained in the set of pre-specified, controlled tables.

References

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection. *European Journal of Operational Research*, 171, 39–52.

Castro, J. and González J.A. (2009). A Package for L1 Controlled Tabular Adjustment, paper presented at the *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Bilbao, 2-4 December 2009)* available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Castro, J. (2011). Extending controlled tabular adjustment for non-additive tabular data with negative protection levels. *Statistics and Operations Research Transactions*, 35.

Dandekar, R.H. and Cox, L. (2002). Synthetic Tabular Data – an Alternative to Complementary Cell Suppression, unpublished manuscript.

Fraser, B. and Wooton, J. (2006). A proposed method for confidentialising tabular output to protect against differencing, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, 299–302.

Giessing, S. andf Höhne, J. (2010). Eliminating small cells from census counts tables: some considerations on transition probabilities. In J. Domingo-Ferrer and E. Magkos, eds., *Privacy in Statistical Databases*, 52–56. New York: Springer-Verlag. LNCS 6344.

Höhne, J. (2003a). SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung Statistischer Einzelangaben, in *Berliner Statistik-Statistische Monatsschrift 3/2003*.

Höhne, J. (2003b), SAFE – a method for statistical disclosure limitation of microdata, paper presented at the *Joint ECE/Eurostat Worksession on Statistical Confidentiality* in Luxembourg, December 2007, available at www.unece.org/stats/documents/2003/04/confidentiality/wp.37.e.pdf

Leaver, V. (2009). Implementing a method for automatically protecting user-defined Census tables, paper presented at the *Joint ECE/Eurostat Worksession on Statistical Confidentiality* in Bilbao, December 2009, available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Shlomo, N. and Young, C. (2008). Invariant post-tabular protection of census frequency counts. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 77–89. New York: Springer-Verlag. LNCS 5262.

Appendix

Table 1: *Zero mean, Variance $2+\varepsilon$ probability transition matrices for maximum perturbations D of 3, 4 and 5 vs. empirically observed transition probabilities for SAFE.*

	0	3	4	5	6	7	8	9	10	11	12	13
						Random Noise, *D* = 3						
1	0.667	0.332	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.125	0.687	0.063	0.063	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.601	0.099	0.100	0.100	0.100	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.167	0.167	0.416	0.083	0.083	0.083	0.000	0.000	0.000	0.000	0.000
6	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000	0.000	0.000
7	0.000	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000	0.000
8	0.000	0.000	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000
						Random Noise, *D* = 4						
1	0.667	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.120	0.700	0.082	0.045	0.027	0.026	0.000	0.000	0.000	0.000	0.000	0.000
4	0.064	0.076	0.700	0.068	0.037	0.029	0.026	0.000	0.000	0.000	0.000	0.000
5	0.000	0.143	0.143	0.542	0.043	0.043	0.043	0.043	0.000	0.000	0.000	0.000
6	0.000	0.063	0.063	0.063	0.662	0.038	0.038	0.038	0.038	0.000	0.000	0.000
7	0.000	0.032	0.033	0.034	0.050	0.700	0.050	0.034	0.033	0.032	0.000	0.000
8	0.000	0.000	0.032	0.033	0.034	0.050	0.700	0.050	0.034	0.033	0.032	0.000
						Random Noise, *D* = 5						
1	0.667	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.119	0.700	0.082	0.050	0.028	0.014	0.007	0.000	0.000	0.000	0.000	0.000
4	0.062	0.076	0.704	0.075	0.037	0.020	0.014	0.012	0.000	0.000	0.000	0.000
5	0.025	0.068	0.068	0.700	0.059	0.027	0.019	0.018	0.017	0.000	0.000	0.000
6	0.000	0.057	0.057	0.057	0.700	0.041	0.023	0.021	0.021	0.021	0.000	0.000
7	0.000	0.025	0.035	0.035	0.062	0.700	0.060	0.028	0.020	0.018	0.018	0.000
8	0.000	0.015	0.016	0.019	0.032	0.068	0.700	0.068	0.032	0.019	0.016	0.015
						SAFE						
1	0.680	0.288	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.408	0.472	0.073	0.006	0.040	0.001	0.000	0.000	0.000	0.000	0.000	0.000
3	0.208	0.514	0.101	0.015	0.153	0.008	0.000	0.001	0.000	0.000	0.000	0.000
4	0.077	0.440	0.122	0.026	0.262	0.058	0.002	0.013	0.000	0.000	0.000	0.000
5	0.022	0.294	0.112	0.046	0.337	0.111	0.023	0.053	0.002	0.000	0.000	0.000
6	0.004	0.157	0.085	0.051	0.347	0.154	0.052	0.136	0.010	0.000	0.002	0.000
7	0.000	0.037	0.070	0.044	0.294	0.182	0.087	0.198	0.071	0.004	0.013	0.000
8	0.000	0.009	0.015	0.035	0.203	0.164	0.119	0.244	0.123	0.044	0.042	0.002

SORT Special issue: Privacy in statistical databases, 2011, 77-98

The Microdata Analysis System at the U.S. Census Bureau*

Jason Lucero, Michael Freiman, Lisa Singh, Jiashen You,
Michael DePersio and Laura Zayatz

Abstract

The U.S. Census Bureau has the responsibility to release high quality data products while maintaining the confidentiality promised to all respondents under Title 13 of the U.S. Code. This paper describes a Microdata Analysis System (MAS) that is currently under development, which will allow users to receive certain statistical analyses of Census Bureau data, such as cross-tabulations and regressions, without ever having access to the data themselves. Such analyses must satisfy several statistical confidentiality rules; those that fail these rules will not be output to the user. In addition, the *Drop q Rule*, which requires removing a relatively small number of units before performing an analysis, is applied to all datasets. We describe the confidentiality rules and briefly outline an evaluation of the effectiveness of the *Drop q Rule*. We conclude with a description of other approaches to creating a system of this sort, and some directions for future research.

MSC: 62P20, 62Q05, 68P20

Keywords: Statistical disclosure limitation, Microdata analysis, Information retrieval, Cross-tabulations.

1. Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code, which prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. However, the agency

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.
Disclaimer: The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Received: November 2010
Accepted: March 2011

also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality, as described in Duncan et al. (2001) and Kaufman et al. (2005).

This paper discusses a Microdata Analysis System (MAS) that is under development at the U.S. Census Bureau. Much of the framework for the system was described in Steel and Reznek (2005) and Steel (2006). The system is designed to allow data users to perform various statistical analyses (regressions, cross-tabulations, correlation coefficients, etc.) on confidential survey and census microdata without seeing or downloading the underlying microdata.

In Section 2, we give some background on the MAS and the motivation for its development. In Section 3, we discuss the current state of the prototype system, including its capabilities and the rules that protect confidentiality. In Section 4, we briefly summarize a study of the effectiveness of the *Drop q Rule*, one of the disclosure avoidance measures taken within the system. In Section 5, we examine some other approaches to the problem of creating a remote access system such as the MAS. In Section 6, we conclude with remarks on future research and the further development of the system.

2. Background on the MAS

The Census Bureau conducts reidentification studies on our public use microdata files. In these studies, we attempt to link our public use files to external files that contain identifiers. It is reasonable to expect that with more publicly available data and expanded use of data mining tools, there will be an increase in the number and complexity of confidentiality threats. There is some concern that in order to meet the confidentiality requirements under which the Census Bureau operates, we may have to reduce the detail available in our data products and use more perturbation techniques to protect them, thus degrading the quality of the data.

This problem of data confidentiality—at the Census Bureau and other statistical agencies around the world—has motivated the creation of *remote access systems* which allow the user to request a statistical analysis and receive the result without having direct access to the underlying microdata. Common to almost all remote access systems is that the ability to receive desired results is not absolute: in some instances, the result might be based on perturbed data, and most proposals for remote access systems include the rejection of some queries to preserve confidentiality. The idea of a remote access system goes back at least to Keller-McNulty and Unger (1998), although the concept of allowing customized queries was proposed much earlier; see the description of the Geographically Referenced Data Storage and Retrieval System in Fellegi et al. (1969). Fellegi (1972) anticipates the need to screen the query results to ensure that confidentiality is adequately protected. Adam and Worthmann (1989) describe several restrictions on systems that release counts of numbers of people with particular characteristics. These include suppressing counts if the numbers are too close to 0 or to the full size of the database; requiring that multiple queries from the same user have only limited overlap; and keeping a log of each user's queries and checking each new query against the log to verify nondisclosure. However, they acknowledge that the last of these is sufficiently time consuming and storage intensive as to be unfeasible. They also consider the possibility of partitioning the data into indivisible units of two or more observations each and allowing only queries that operate on unions of the units, rather than on arbitrary sets of observations.

The Microdata Analysis System will allow the U.S. Census Bureau to provide a controlled, cost-effective setting in which data users have access to more detailed and accurate information than is currently available in our public use microdata files. The data accessible through the MAS can identify smaller geographic areas and show more detail in certain variables where our public use files would be coarsened. Our goal for the MAS is to allow access to as much high quality data as possible. An advantage of the MAS is that it lessens the need for data to be released in less secure or more expensive manners, such as those described in Weinberg et al. (2007). A predecessor of the MAS is discussed in Rowland and Zayatz (2001).

Unlike the proposal in Schouten and Cigrang (2003), our plan is to make the MAS available to anyone who wishes to use it. In a sense, the MAS will serve as a Research Data Center for the entire public, although there will be restrictions in place that a qualified researcher would not encounter at an established Research Data Center. The MAS will allow access to data from demographic surveys and decennial censuses, with the goal of eventually including economic survey and census data, as well as linked datasets. We will initially make available regression analyses and cross-tabulations, with other analyses to be added in the future. Currently, we intend to keep a record of all of the queries entered into the system, but not the identities of the users making the queries. Although the record will not directly affect the output that the system provides, it will allow us to see how the system is being used. Our goal in doing this is to improve the user experience and enhance the disclosure avoidance techniques if necessary.

Our current plan—as described in Chaudhry (2007)—is to offer the MAS through the Census Bureau's free DataFERRETT service with the intention that the system will be used by people needing fairly simple statistical analyses: news media, some policy makers, teachers, students, etc. The MAS has a graphical interface that allows users to select variables of interest from a list. In the case of regression, variables can be dragged into equations and, with a few clicks, users may create variable interactions and transformations of selected variables. Some users may feel the need to use the underlying confidential microdata for more exploratory data analysis, but it is not apparent how to allow this within the MAS without violating confidentiality. These users may find our public use files, when available, meet their needs if they account for the decreased accuracy inherent in our disclosure avoidance procedures. Having a limited range of allowable analyses is a weakness of the MAS, but, other than expanding the number of off-the-shelf analyses the system offers, it is difficult to see how to remedy it.

3. Overview of the MAS Confidentiality Rules

In 2005, the Census Bureau contracted with Synectics to develop an alpha prototype of the MAS using the SAS language. We also contracted with Dr. Jerome Reiter of Duke University to help in developing the confidentiality rules of the system and with Dr. Stephen Roehrig of Carnegie Mellon University to help in testing these rules. Some rules were developed and modified as a result of the testing. The beta prototype of the MAS implements a Java interface within DataFERRETT, which submits requested analyses to an R environment. We are using the publicly available data from the Current Population Survey March 2008 Demographic Supplement to test the system.

The MAS software is programmed with several confidentiality rules and procedures that uphold disclosure avoidance standards. The purpose of these rules and procedures is to prevent data intruders from reconstructing the microdata records of individuals within the underlying confidential data through submitting multiple queries. The confidentiality rules discussed in this section are quite complex, and this discussion does not delve into the complexities. More detail can be found in Lucero (2009, 2010a). All analyses are subjected to two logical checks, referred to as the *No Marginal 1 or 2 Rule* and the *Universe Gamma Rule*, which ensure that no query is answered if the universe is too small or if the universe can be used to carry out differencing attacks by comparing results of similar universes. Regression analyses are further subjected to restrictions on the use of predictor and response variables. We plan to explore whether additional rules are necessary for correlation coefficients.

3.1. Confidentiality Rules for Universe Formation

MAS users are allowed to run their statistical analyses on a universe, or sub-population, of interest. Users are presented with a set of variables and category levels from which they can define a universe using condition statements on the variables. For example, if the user selects $gender = 2(female)$ from the metadata, the universe is defined to be the sub-population of all females. A slightly more complicated universe is $gender = 1(male) \vee \textit{employment status} = 0(unemployed)$. One of the confidentiality rules requires that all variables used to define universes must be categorical.

Since a user may want to define a universe based on variables that are not inherently categorical (i.e., those that are continuous), raw numerical variables are presented to the user as categorical recodes based on output of a separate binning routine. This cutpoint program, outlined in Lucero et al. (2009b), creates bins of numerical values and ensures a pre-specified minimum number of observations between any two cutpoint values. Section 3.1.3 describes possible ways to generate cutpoints.

To define a universe using a numerical variable, a user is forced to choose from a predetermined list of ranges the range that best meets her goal. For example, if a user wished to run analysis on people with *income* = $46,000, the user would select the metadata *income* = 4, which is the range $(\$45,000, \$53,000]$ on the variable *income*

Table 1: *Table representation of the universe defined from (1) and (2).*

	income				
gender	\$0 to \$28,000	\$28,000 to \$39,000	\$39,000 to \$45,000	\$45,000 to \$53,000	Total
male	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$n_{1,.}$
female	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,.}$
Total	$n_{.,1}$	$n_{.,2}$	$n_{.,3}$	$n_{.,4}$	$n_{.,.}$

and defines the universe as the sub-population of all individuals whose income is between \$45,000 and \$53,000. Note that a user cannot define the universe to be the range $income = (\$39,000, \$46,000]$ unless \$39,000 and \$46,000 are among the pre-determined cutpoints. The user must choose a range of values consistent with the cutpoints that are given. This is a crucial restriction on what a user can do, since allowing arbitrary universe formation on continuous data could lead to a differencing attack disclosure, as described in Section 3.1.2.

3.1.1. Confidentiality by Minimum Universe Size Requirements

To define a universe in the MAS, the user would first select *m* recoded variables from the metadata, then select up to *j* bins for each of the *m* recoded variables. Universe formation on the MAS is performed using an implicit table server. For example, suppose a data user defines the universe as the union:

$$gender = \text{female AND } \$45,000 < income \leq \$53,000 \tag{1}$$

OR

$$gender = \text{male AND } \$28,000 < income \leq \$45,000 \tag{2}$$

This universe is represented as selected cells from a two-way table of counts for *gender* and *income*, as shown in *Table 1*. Note that there are $n_{2,4} + n_{1,2} + n_{1,3}$ total observations in this universe. For convenience, we will use the notation U(n) to denote a universe with *n* observations. In most cases, it should be clear from the context which *n* observations lie in the universe. In this example, the universe defined as the union of (1) and (2) will be referred to as U($n_{2,4} + n_{1,2} + n_{1,3}$).

In describing universes, we make a distinction between a simple universe and a complex universe. A simple universe is one that can be described using variable categories and the intersection set operator. A complex universe is constructed as the union of multiple simple universes.

All universes formed on the MAS must pass two confidentiality rules: the *No Marginal 1 or 2 Rule* and the *Universe Gamma Rule*. If a universe violates either of these rules, the MAS will reject the universe query and prompt the user to modify his selections. These rules are tested prior to performing the user's selected statistical analysis on the defined universe.

The *No Marginal 1 or 2 Rule* requires that for a universe defined using *m* variables, there may not be an $m-1$ dimensional marginal total equal to 1 or 2 in the *m*-way contingency table induced by the chosen variables. The universe U($n_{2,4}+n_{1,2}+n_{1,3}$) passes the *No Marginal 1 or 2 Rule* if:

$$(n_{i,\cdot} \geq 3 \text{ OR } n_{i,\cdot} = 0, \text{ for } i = 1,2) \text{ AND } (n_{\cdot,j} \geq 3 \text{ OR } n_{\cdot,j} = 0, \text{ for } j = 1,\ldots,4)$$

The *Universe Gamma Rule* requires that a universe must contain at least Γ observations; otherwise no statistical analysis will be performed. The value of Γ is not given here since it is Census confidential.

The way this rule is checked is dependent on whether the universe is disjoint or joint. A universe is classified as *disjoint* if its individual pieces do not share cell counts in common. For example, pieces (1) and (2) for the universe U($n_{2,4}+n_{1,2}+n_{1,3}$) are disjoint. Since U($n_{2,4}+n_{1,2}+n_{1,3}$) is a disjoint universe, the MAS would check that piece (1) and piece (2) each contain at least Γ observations. Note that the cutpoint bins of *income* are combined within piece (2) prior to performing the test; however, bins representing different classes of an inherently categorical variable would not be combined. In this case, since the $n_{1,2}$ and $n_{1,3}$ bins differ from each other only by a cutpoint variable, they are combined, and the MAS checks:

$$n_{2,4} \geq \Gamma \text{ AND } (n_{1,2}+n_{1,3}) \geq \Gamma$$

A universe is classified as *joint* if at least one of its individual pieces shares cell counts in common with at least one other piece. For example, suppose the user defines the universe U($n_{2,\cdot}+n_{1,3}+n_{1,4}$) = (3) OR (4), where (3) and (4) are given by

$$[gender = \text{ female}] \tag{3}$$

$$[\$39,000 < income \leq \$53,000] \tag{4}$$

In this case, the observations in $n_{2,3}$ and $n_{2,4}$ — females with income in the interval ($39,000 , $53,000] — are included in both pieces (3) and (4). See *Table 2*. Since U($n_{2,\cdot}+n_{1,3}+n_{1,4}$) is a joint universe, the *Universe Gamma Rule* would first check that pieces (3) and (4) contain at least Γ observations, following the disjoint universe scenario. Next, the intersection $I = (3) \cap (4) \neq \{\}$ would be checked to determine that *I* contains at least Γ* observations, where $\Gamma^* \leq \Gamma$ is another Census confidential parameter. In this example, the MAS checks that the following inequalities are satisfied before any results will be returned:

$$n_{2,\cdot} \geq \Gamma \text{ AND } (n_{\cdot,3}+n_{\cdot,4}) \geq \Gamma \text{ AND } (n_{2,3}+n_{2,4}) \geq \Gamma^*$$

Once again, the cutpoint bins of income are first combined within piece (4) and within *I* prior to the testing of the *Universe Gamma Rule*. In general, when a joint universe

Table 2: *Table representation of the universe defined from (1) and (2).*

	income				
gender	\$0 to \$28,000	\$28,000 to \$39,000	\$39,000 to \$45,000	\$45,000 to \$53,000	Total
male	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$n_{1,\cdot}$
female	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,\cdot}$
Total	$n_{\cdot,1}$	$n_{\cdot,2}$	$n_{\cdot,3}$	$n_{\cdot,4}$	$n_{\cdot,\cdot}$

$$\begin{array}{c|cc} \mathrm{T}_n & ES_1 & ES_2 \\ \hline G_1 & n_{1,1} & n_{1,2} \\ G_2 & n_{2,1} & n_{2,2} \end{array} - \begin{array}{c|cc} \mathrm{T}_{n-1} & ES_1 & ES_2 \\ \hline G_1 & n_{1,1} & n_{1,2}-1 \\ G_2 & n_{2,1} & n_{2,2} \end{array}$$

$$= \begin{array}{c|cc} \mathrm{T}_1 & ES_1 & ES_2 \\ \hline G_1 & 0 & 1 \\ G_2 & 0 & 0 \end{array}$$

Figure 1: *An Example of a Differencing Attack Disclosure.*

is considered, all of the non-empty intersections of the pieces of the universe must be checked to make sure they are sufficiently large.

3.1.2. Confidentiality by Random Record Removal

While the preceding rules provide some protection of the confidential data in the MAS, they do not completely prevent differencing attack disclosures. A *differencing attack disclosure* occurs when a data intruder attempts to reconstruct a confidential microdata record by subtracting the statistical analysis results obtained through two queries on similar universes. Suppose a data intruder first creates two universes on the MAS, U(n) and U($n-1$) (a proper subset of U(n)), where both contain the same n observations less one unique observation, i.e., $|\mathrm{U}(n)\backslash \mathrm{U}(n-1)| = 1$. The difference $\mathrm{U}(n)\backslash \mathrm{U}(n-1) = \mathrm{U}(1)$ is a manipulated universe that contains the single target observation. For illustration, suppose a data intruder has prior knowledge of demographics in a small geographic area, and in particular is aware of individuals, households or establishments with unique characteristics within that area. It may be the case that there is only one non-citizen among the n residents of the area. Then the intruder may create U(n) and U($n-1$), where U(n) is the full universe of people in the area and U($n-1$) is the universe consisting of citizens who live in the area. Suppose the data intruder then requests two separate cross-tabulations for gender by employment status on these universes, T_n and T_{n-1}, as shown in *Figure 1*. Since U(n) and U($n-1$) differ by a unique observation, T_{n-1} will be exactly the same as T_n, less one unique cell count.

We may perform the matrix subtraction $\mathrm{T}_n - \mathrm{T}_{n-1} = \mathrm{T}_1$, where T_1 is a two-way table of gender by employment status built upon the one unique observation contained in

$\mathrm{U}(n)\backslash\mathrm{U}(n-1) = \mathrm{U}(1)$. As shown in *Figure 1*, T_1 contains a cell count of 1 in the male non-employed cell with zeros in the remaining cells, which tells the data intruder that the one unique observation contained in U(1) is an unemployed male. By performing differencing attacks similar to the one just described, a data intruder can successfully rebuild the confidential microdata record for the one unique observation contained in U(1).

A differencing attack may also be a concern if there are two observations within an area that have a certain characteristic, particularly if the intruder is himself one of these two. Suppose, for example, that the universe contains only two non-citizens, one of whom is the intruder. The intruder could then construct the full universe $\mathrm{U}(n)$ and the portion of the universe consisting solely of citizens $\mathrm{U}(n-2)$. Since the intruder knows his own personal characteristics, he may manually remove himself from $\mathrm{U}(n)$ to get $\mathrm{U}(n-1)$ and then perform a differencing attack as above by comparing $\mathrm{U}(n-1)$ and $\mathrm{U}(n-2)$ to obtain information on the other non-citizen in the area.

To help protect against differencing attacks, the MAS implements a universe subsampling routine called the *Drop q Rule*. Traditionally, subsampling has usually been used to estimate parameters when a population is too large to analyze in an efficient manner and a (usually small) subset can give approximately the same results as the full population. Our aims are very different here: the *Drop q Rule* is intended to remove just enough observations from the dataset to thwart a differencing attack. In most cases, a differencing attack performed while the *Drop q Rule* is in place will not lead to a meaningful outcome, and even when it does, the intruder cannot be sure that the outcome found is the correct one.

The *Drop q Rule* works as follows. A user-defined universe that passes all of the previous rules has q records removed at random. To do this, the MAS will first draw a random value of $Q_v = q_1 \in \{2,\ldots,k\}$ from a discrete uniform distribution with probability mass function $\mathrm{P}(Q_v = q_1) = \frac{1}{k-1}$. Then, given $Q_v = q_1$, the MAS will subsample the universe $\mathrm{U}(n)$ by removing q_1 records at random from $\mathrm{U}(n)$ to yield a new subsampled universe $\mathrm{U}(n-q_1)$.

Within the MAS, all statistical analyses are performed on the subsampled universe $\mathrm{U}(n-q_1)$ and not on the original universe $\mathrm{U}(n)$. Each unique universe $\mathrm{U}(n)$ that is defined on the MAS will be subsampled independently according to the *Drop q Rule*. To prevent an “averaging of results” attack, the MAS will produce only one subsampled universe $\mathrm{U}(n-q_1)$ for each unique universe $\mathrm{U}(n)$, with this unique subsample persisting for the lifetime of the system. That is, all users who select a specific universe $\mathrm{U}(n)$ will have all analyses performed on exactly the same subsampled universe $\mathrm{U}(n-q_1)$. To avoid obvious storage issues, the MAS accomplishes consistent subsampling of universes by using the same random seed to perform the subsampling every time a given universe comes up. To receive the full disclosure protection offered by the *Drop q Rule*, it is necessary that the seed, while constant for a given universe, differs across universes, and this can be implemented by having the seed be a function of the set of units in the universe.

The discrete uniform distribution is ideal for this purpose because of all distributions on $\{2,\ldots,k\}$, it minimizes the probability that for two similar universes, the number of observations dropped will be the same for both universes, which is a necessary condition for an apparent disclosure to be made on a single observation.

Because each value used in the *Drop q Rule* is drawn from a discrete uniform distribution, a data intruder attempting the difference attack $\mathrm{T}_n - \mathrm{T}_{n-1} = \mathrm{T}_1$ may find results inconsistent with forming two universes where $\mathrm{U}(n-1) \subset \mathrm{U}(n)$, as shown in *Figure 2*. The values of x_{ij} are the random numbers giving the number of observations dropped from each cell of U(n) in forming U($n - q_1$). Similarly, the values of y_{ij} are the number of observations dropped from each cell of U(n – 1) in forming U(n – 1 – q_2) respectively. Hence:

$$\sum_i \sum_j x_{ij} = q_1, 0 \le x_{ij} \le q_1$$

$$\sum_i \sum_j y_{ij} = q_2, 0 \le y_{ij} \le q_2$$

Here, *i* and *j* index the rows and columns, respectively, of the contingency table, with the obvious generalizations involving higher order multiple sums for higher-dimensional data. The resulting table $\mathrm{T}_?$ *may* yield a successful disclosure of *gender* $= G_1$ (male) AND *employment status*$= ES_2$ (unemployed) for the one unique observation contained in U(1), but it is much more likely to supply nonsense to the intruder. Coupled with the difficulty of finding candidate differencing attack universes, data intruders will find their time better spent elsewhere. Section 4 contains a brief overview of the effectiveness of the *Drop q Rule* against differencing attack disclosures. The rule is a crucial part of our disclosure prevention strategy. The contracted work described by Roehrig et al. (2008) found several instances in which a prototype version of the MAS lacking this rule was susceptible to differencing attacks, not just in theory but also in practice. However, their approach was to run a large number of tabulation queries and search for universes that were almost the same. This method could be partly deterred by slowing down the system, requiring a wait time between each user query.

$$\begin{array}{r|cc} \mathrm{T}_{n-q_1} & ES_1 & ES_2 \\ \hline G_1 & n_{1,1}-x_{1,1} & n_{1,2}-x_{1,2} \\ G_2 & n_{2,1}-x_{2,1} & n_{2,2}-x_{2,2} \end{array} \; - \; \begin{array}{r|cc} \mathrm{T}_{n-1-q_2} & ES_1 & ES_2 \\ \hline G_1 & n_{1,1}-y_{1,1} & n_{1,2}-1-y_{1,2} \\ G_2 & n_{2,1}-y_{2,1} & n_{2,2}-y_{2,2} \end{array}$$

$$= \begin{array}{r|cc} \mathrm{T}_? & ES_1 & ES_2 \\ \hline G_1 & y_{1,1}-x_{1,1} & 1+y_{1,2}-x_{1,2} \\ G_2 & y_{2,1}-x_{2,1} & y_{2,2}-x_{2,2} \end{array}$$

Figure 2: *Differencing Attack Thwarted by the* Drop q Rule.

The *Drop q Rule* is a generalization of the previously used *Drop 1 Rule* and *Drop 2 Rule*, where a small and fixed number of observations were removed before analysis. These rules led to tables that were susceptible to differencing attacks. One notable vulnerability could be exploited by starting, as usual, with two universes U(n) and U($n-1$), identical with the exception of one unit, with the intention of performing a differencing attack. For example, an intruder might know that a certain geographical region contains exactly one Korean War veteran. The intruder could then consider the universe of all people in that region, as compared to the universe of all non-Korean War veterans in the region. However, instead of requesting a tabulation of these two universes, the intruder may augment each universe by adding to it the full population of a non-overlapping geographical region of size $N >> n$, such as a large state that does not contain the original region. Then a three-way tabulation could be done of veteran status versus state versus the variable that the intruder wishes to disclose for the augmented universes U($n+N$) and U($n-1+N$). In the case of the *Drop 2 Rule*, it is overwhelmingly likely that all four of the dropped observations will be in the large region of size N, thus leaving the portions of the provided tables representing the original region of interest unmodified. We are currently examining other disclosure rules to prevent this sort of "padding" attack.

A differencing attack leads to a correct inference when the difference between the two matrices represented by the modified tables contains a 1 in the correct cell and 0s in all other cells. In most cases, when the *Drop q Rule* is used, there are cells with both positive and negative numbers, and no inference can be reached by the intruder. It is also possible to obtain an apparent—but incorrect—inference, which occurs when the difference is a table with a 1 in one cell and 0s in all of the others, but the 1 is not in the correct cell.

3.1.3. Cutpoint Methods

The cutpoints used in universe formation in the MAS are generated by a separate program. Various methods exist in the program, and each provides a different set of cutpoints, as influenced by the empirical distribution of a variable. The methods implemented are fixed width, minimum width, increasing width, and partitioned binning. Cutpoints for each variable in the dataset can use a different strategy, but the final cutpoints for a given variable are generated only once, after choosing an appropriate strategy. What follows is a basic description of each strategy.

Fixed width binning ensures that all bins have the same width. This is implemented as finding a constant ω_{FW}, such as 10, so that the distance from the minimum value to the maximum value of each bin will be ω_{FW}. Because bin widths are constant, the number of observations in each bin will vary, causing some bins to be sparsely populated while others are dense. The fixed width is chosen to be the minimum value ω_{FW} such that all bins contain at least β_{FW} observations, for some pre-determined value β_{FW}. This can make ω_{FW} large, so that the resolution across dense areas of the data is too crude.

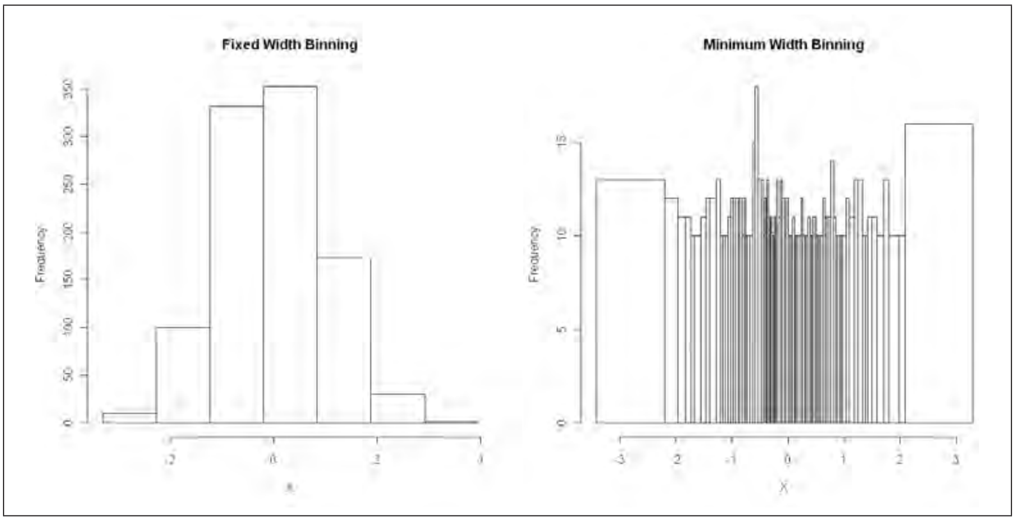

Figure 3: *Fixed and Minimum Width Binning on 1,000 N(0,1) random samples.*

In data following a Gaussian distribution, the bin width will be determined by the tails and the center bins will be quite dense.

Minimum width binning uses a value β_{MW} and creates bins such that each has as close to β_{MW} observations as possible. Identical realizations of the variable will not be split across multiple bins. For example, considering a numerical variable X with support $\mathbb{N}$, all observations with $X = 5$ will belong to the same bin regardless of the number of observations with $X = 5$. This approach tends to generate bins of smaller width than other approaches, since it allows for finer resolution in dense areas of the data but allows the bins to be much wider when covering sparse data in order to include at least β_{MW} observations.

Increasing width binning may be viewed as a compromise between fixed and minimum width binning. Increasing width binning starts with a fixed bin width, ω_{IW}, which gradually increases as the value of the variable increases. This corrects the problem in fixed width binning of bins tending to be large, while also allowing for a consistent bin width, which one does not get in minimum width binning. Considering income data, ω_{IW} might equal $25,000$ at $X = 0$, but when the cutpoint reaches $X = 100,000$, ω_{IW} may jump to 150,000 as a way to deal with sparser data in the tails. For sufficiently large X, we obtain a value of $\omega_{IW} = \infty$ once the number of remaining observations approaches some value $\alpha < 2\beta_{IW}$.

The previous binning methods are all referred to as bottom-up methods since they begin with some width value and starting point in the data and build bins from there. Alternatively, *partitioned binning* is a top-down binning strategy in that it uses the data as a whole in creating bins. Partitioned binning begins by sorting the data and then splits the set into two subsets containing approximately the same number of observations. These two subsets are themselves each split into two smaller subsets in the same fashion.

Table 3: *Bins created on the dataset* $\{1,1,2,2,4,4,5,6\}$.

Method	**Bin 1**	**Bin 2**	**Bin 3**	**Bin 4**
Fixed W.	1-2	3-4	5-6	NA
Min. W.	1-1	2-2	4-4	5-6
Inc. W.	1-2	3-6	NA	NA
Partitioned	1-1	2-2	4-4	5-6

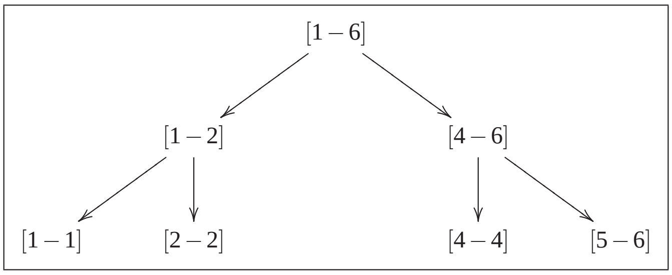

Figure 4: *Partitioned Binning on dataset* $\{1,1,2,2,4,4,5,6\}$.

This process continues as long as there are at least β_{PW} observations in each bin. The final result is a binary tree of bins of unequal width.

As a quick example of how each method performs on the same data, consider a dataset 1,1,2,2,4,4,5,6. *Table 3* shows the cutpoints, or boundaries, for each bin that the different algorithms will create. Assume that the minimum number of elements in each bin is $\beta_{MIN} = 2$.

The binary tree for the partitioned binning is shown in *Figure 4*. A user may choose pieces for the universe using any node shown in the diagram.

Each approach has its own strengths and weaknesses, so which performs best on a given variable depends both on the variable's support and distribution and on the properties desired by the user. However, none of the methods considers the underlying distribution of a variable in building the bins, so there is a necessity to analyze the performance of a chosen method. Consider how each would perform on a Gaussian distribution. Fixed width binning may not provide the resolution desired around the mean, and increasing width binning is primarily useful when the probability density function of the variable in question is decreasing over most of the range of the variable. Partitioned and minimum width binning will produce similar results, but the cutpoints in the minimum width and partitioned approaches may provide binning so fine that the exact values for some records are at risk.

3.2. Confidentiality Rules for Regression Models

The MAS implements a series of confidentiality rules for regression models, in addition to the universe restrictions already mentioned. For example, users may only select up to 20 independent variables for any single regression equation. Users are allowed to transform numerical variables only, and they must select their transformations from a pre-approved list. This prevents the user from performing transformations that deliberately overemphasize individual observations such as outliers. Currently, the allowable transformations are square, square root and natural logarithm.

Any fully interacted regression model that contains only dummy variables as predictors poses a significant potential disclosure risk, as described in Reznek (2003) and Reznek and Riggs (2004). Therefore, users are allowed to include only two-way and three-way interaction terms within any specified regression model, and no fully interacted models are allowed. Furthermore, a two-way interaction is allowed only if both of the interacted variables appear by themselves in the model, and a three-way interaction is allowed only if all three variables appear uninteracted in the model *and* each of the three associated two-way interactions appears. However, interactions do not count against the 20-variable limit (so that, for example, if a model includes two predictor variables and their interaction, this is considered two variables, not three, for the purpose of the limit). Categorical predictor variables are included in the model through the use of dummy variables for all categories except one reference category. The MAS uses the most common category as the reference category. In addition, each predictor dummy variable must represent a category containing a certain minimum number of observations; if this minimum is not met, the dummy variable is omitted from the model. In effect, this means that very sparse categories are absorbed into the reference category level. The minimum allowable number of observations in a category is not given here since it is Census confidential.

Prior to passing any regression output back to the user, the MAS also checks that R^2 is not too close to 1. If R^2 is too close to 1, then the MAS will suppress the output of the regression analysis, as releasing the results of the regression would allow estimation of the response variable with a high degree of accuracy if the values of the predictor variables for any unit were known. It may also be the case that the regression does not have an unreasonably high R^2, but that there is a subset of units for whom the response variable can be predicted unusually well given the predictor variables. Regressions with this feature may also be suppressed. The system may also suppress instances where an interaction term leads to a sparse combination of categories, as this may be a disclosure risk. If all of these requirements are satisfied, then the MAS will pass the estimated regression coefficients and the Analysis of Variance (or Deviance) table to the user without restrictions (except for the absorption of categories mentioned above). If the requirements are not satisfied, the system may attempt to absorb additional categories of any categorical predictors into the reference category, as this may result in a regression whose output is allowed to be released.

Sparks et al. (2008) propose some other confidentiality rules for regression, such as using robust regression to lessen the influence of outliers, although at the moment we still plan to use ordinary least squares regression when the response variable is numerical.

3.2.1. Synthetic Residual Plots

To determine whether the regression adequately describes the data, diagnostics such as residual plots are necessary. Actual residual values pose a potential disclosure risk, since a data intruder can obtain the values of the dependent variable by simply adding the residuals to the fitted values obtained from the regression model. Therefore, the MAS does not pass the actual residual values back to the user. To help data users assess the fit of their ordinary least squares regression models, diagnostic plots are based on synthetic residuals and synthetic real values. These plots are designed to mimic the actual patterns seen in the scatter plots of the real residuals versus the real fitted values, or of the real residuals versus the values of the individual variables.

The first step in creating synthetic residual plots is to create the synthetic dataset in such a way that the synthetic data mimic the actual data. Using the notation of Reiter (2003), let $\boldsymbol{x}_p$ be a variable in the collected dataset, for $p = 1, \ldots, d$. In the synthetic dataset, $\boldsymbol{x}_p^s$ corresponds to the original $\boldsymbol{x}_p$ variable, with the superscript s indicating the use of a synthetic dataset. There are various methods to generate $\boldsymbol{x}_p^s$, but this discussion will follow the method described in Reiter (2003), both for creating synthetic data and for creating synthetic residuals, and our exposition and notation here mostly follow his.

For categorical variables $\boldsymbol{x}_p$, $\boldsymbol{x}_p^s$ are generated from bootstrap sampling the collected data. If some categories are sparsely populated, there is the potential for averaging the synthetic residual values at the sparse category to disclose real residuals, but otherwise this part of the algorithm poses negligible disclosure risk. One possible approach to this problem is to suppress residuals for categories that are sufficiently sparse. For continuous variables $\boldsymbol{x}_p$, the distribution of the variable is approximated non-parametrically using a kernel density estimator, and then inverse-cdf sampling is used to generate $\boldsymbol{x}_p^s$ from the approximate distribution. When Reiter's method is used, there is no one-to-one correspondence between real observations and synthetic observations, so there need not be any particular relationship between the size of the actual dataset and the size of the synthetic sample. This feature helps to protect outliers, as an outlier in the original data may not appear in the synthetic plot or may appear more than once. In the case of categorical predictor variables, we let the synthetic sample size equal the actual sample size, while in the case of numerical predictor variables, we let the synthetic sample size be the minimum of 5,000 and the actual sample size. This is because when making the synthetic and actual sample sizes equal in the numerical case, we found that the system was slow when dealing with large datasets, and that the vast majority of the time that the analysis took was spent on creating the synthetic residual plots for numerical variables.

A shortcoming of the method for creating synthetic continuous predictors is that the kernel density estimator is not able to identify a probability mass at a single point, but rather will assume that the probability density function should be high in the neighborhood of that point. This should not invalidate the method, but it will affect the distribution along the x-axis for a predictor variable such as income, for whom many people have a true value of 0.

It should be noted that both of these methods for creating the synthetic data work with one variable at a time, i.e., $\boldsymbol{x}_p^s$ are drawn marginally, not jointly, and thus no valid analysis can be performed based on the joint distribution of the synthetic variables. This is not currently a major concern, as it is not our intention to release synthetic data through the MAS. However, this does impose a limitation on the range of diagnostics that we can make available in the future based on synthetic variables generated using this method.

The next step is to generate the standardized synthetic residuals $\boldsymbol{t}_p^s$ so that the relationship between $\boldsymbol{t}_p^s$ and $\boldsymbol{x}_p^s$ at any point x_{kp}^s in $\boldsymbol{x}_p^s$ is consistent with the relationship between $\boldsymbol{t}$ and $\boldsymbol{x}_p$ around point x_{kp}^s. To accomplish this, we must make a different set of synthetic residuals for each predictor variable. Note that x_{kp}^s, if numerical, will not necessarily be a value observed in continuous real data, but may be drawn with the inverse-cdf method.

For each variable, the goal is to give the user something akin to a plot of the standardized residuals of the full (possibly multiple) regression model versus the value of $\boldsymbol{x}_p$. For a variable p and an index k, define

$$t_{kp}^s = b_{kp} + v_{kp} + n_{kp}$$

The first term gives the expected value of the standardized residual for any given value of p; the second accounts for the variation of the actual standardized residuals around their expected values (which may change depending on the value of x_{kp} if heteroscedasticity is present); and the third adds noise to further prevent disclosure.

To calculate the first term b_{kp}, a generalized additive model (GAM) is built for $\boldsymbol{t}$ and $\boldsymbol{x}_p$. The value b_{kp} equals the value of the GAM curve at the point x_{kp}^s and is used to fit the values t_{kp}^s to the general relationship of $\boldsymbol{t}$ and $\boldsymbol{x}_p$, ignoring for the moment the variation of $\boldsymbol{t}$ around its local mean. Note that $\boldsymbol{t}_p^s$ will differ for every regression a user requests, and that it is important that the GAM not be overfit. In extreme cases, an overfit GAM can create some of the same disclosure risks as releasing a regression with a high R^2. There may be some difficulty in avoiding such an overfit in an automated setting. For categorical variables, a GAM cannot be fit, and we set $b_{kp} = 0$ because whenever a regression including a categorical variable is performed, the mean residual among observations with any particular level of that categorical variable is 0.

Next, t_{kp}^s is shifted off the curve b_{kp} by v_{kp}, which represents the amount by which the points in the real data around x_{kp}^s deviate from the curve. For the case where $\boldsymbol{x}_p$ is numerical, we consider the real data standardized residual t_j, where

$$j = \text{arg min}_i |x_{kp}^s - x_{ip}|$$

is the index of the unit in $\boldsymbol{x}_p$ whose value is closest to x_{kp}^s. Ties can be broken by selecting randomly from all tied choices. Having found j, we compute $v_{kp} = t_j - b_{jp}$ where b_{jp} is the value obtained from the GAM at x_{jp}. If $\boldsymbol{x}_p$ is categorical, j is the index of a randomly selected observation in the real data such that $x_{jp} = x_{kp}^s$, so we set $v_{kp} = t_j$, since $b_{jp} = 0$.

Finally, a noise term $n_{kp} \sim N(0,\sigma)$ is added to t^s_{kp} where, for each regression, σ should remain constant so that there is not artificial heteroscedasticity in the synthetic residuals. The same random seed should be used for all regressions using the same dependent variable; if this were not done, there would be the possibility of running the same or similar models a number of times and averaging the different results, creating a disclosure risk. Careful selection of σ is important, as a value that is too small may not provide enough protection against disclosure, while a value that is too large may cause patterns that are of interest to a legitimate user to be dwarfed by random variation.

When all steps are complete, the system creates a scatterplot of the synthetic residuals versus each numerical synthetic predictor variable, as well as a scatterplot of the synthetic residuals against the fitted value, with a kernel smoother used to show the general shape of the latter curve. To protect outliers, the scatterplot requires all synthetic standardized residuals to be in the interval [-4,4], with values that would otherwise be outside this range truncated appropriately.

Since categorical predictors do not lend themselves to scatterplots, the residual plots for categorical variables are replaced by side-by-side boxplots. Sparks et al. (2008) propose that numerical predictor variables be binned in a cutpoint-like fashion, and that the bins be used to create categories for side-by-side boxplots, which can be returned to the user instead of scatterplots, with Winsorization being performed to protect outliers. Since this binning lowers the resolution with which we can see the variable along the x-axis, Sparks et al. (2008) use it as a substitute for synthetic data.

We are beginning to implement regression diagnostics for logistic regressions in the manner described in Reiter and Kohnen (2005).

4. Evaluation: Effectiveness of the *Drop q Rule*

What follows is a generalization of some results in Lucero et al. (2009a), although that paper considered an earlier, less secure version of the *Drop q Rule* in which q was a fixed value chosen in advance. We present only a brief overview of this evaluation here; full details are in Lucero (2010b). Given a pair of similar universes, U(n) and U($n-1$), differing by only one unique observation, with n large, we consider the effectiveness of the *Drop q Rule* in preventing contingency table differencing attack disclosures of the form $\mathrm{T}_1 = \mathrm{T}_{n-q_1} - \mathrm{T}_{n-1-q_2}$, as was shown in *Figure 2*.

For this section, we will consider a contingency table giving the values of two categorical variables, with the same setup as described in Section 3.1. To make the notation somewhat less unwieldy, we denote the size of each cell in the contingency table using a single subscript, as shown in *Figure 5*, instead of the double subscript used previously. In the simplest case, the contingency table is 2×2 (two categories for each of two variables), but it could conceivably be larger—including either more categories for a particular variable or more variables, which would lead to more dimensions and would require a more elaborate graphical representation.

T_n	ES_1	ES_2
G_1	n_1	n_2
G_2	n_3	n_4

Figure 5: *Illustration of notation used in Section 4.*

We also let $\boldsymbol{\Pi} = (\Pi_1,\ldots,\Pi_4)$ denote the proportions of observations within each of the cells of T_n and let $\boldsymbol{\Psi} = (\Psi_1,\ldots,\Psi_4)$ denote the proportions within T_{n-1}. If n is large, then $\boldsymbol{\Pi} \approx \boldsymbol{\Psi}$. Furthermore, let $\boldsymbol{X}$ denote the vector giving the number of observations removed from each of the four cells when q_1 observations are dropped from T_n to produce T_{n-q_1}, and let $\boldsymbol{Y}$ denote the vector giving the number of observations removed from each of the four cells when q_2 observations are dropped from T_{n-1} to produce T_{n-1-q_2}. A correct disclosure will occur if and only if $\boldsymbol{X} = \boldsymbol{Y}$, and this may occur only when $q_1 = q_2$.

Since sampling with replacement is very similar to sampling without replacement when n is large, we can say that for a given q_1 and q_2, $\boldsymbol{X}$ is approximately a multinomial random variable with size q_1 and probabilities given by $\boldsymbol{\Pi}$, and $\boldsymbol{Y}$ is approximately a multinomial random variable with size q_2 and probabilities given by $\boldsymbol{\Psi}$. Substituting $\boldsymbol{\Pi}$ for $\boldsymbol{\Psi}$ and performing some other manipulations gives a formula for the approximate probability of disclosure for a given number of cells J, maximum number of cells dropped k and vector $\boldsymbol{\Pi} = (\Pi_1,\ldots,\Pi_J)$:

$$\xi_{J,k}(\Pi_1,\ldots,\Pi_J) = \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\ldots+x_J=q_1} \left(\frac{1}{k-1}\right)^2 \left(\frac{q_1!}{x_1!\cdot\ldots\cdot x_J!}\right)^2 \Pi_1^{2x_1}\cdot\ldots\cdot\Pi_J^{2x_J} \tag{5}$$

This formula has a total of $\binom{J+k}{J} - (J+1)$ summands within a rather involved summation, which makes it cumbersome, but it may be useful in assessing the risk involved with releasing a given table with a given value of k. Further research may focus on finding simpler approximations for the value in this sum.

A large number of differencing attacks were simulated, as described in Lucero (2010b), for a pair of tables, differing by one observation, with $n = 978$ and $k \in \{3,4,5,6,7\}$. The data were from the Current Population Survey March 2000 Demographic Supplement. The simulation led to the conclusion that the summation in (5) generally agrees with the empirical probability of a disclosure to two decimal places for this sample size.

It may also be desirable to find bounds on the summation in (5) in the case in which $\boldsymbol{\Pi}$ is not known. This would be useful, for example, if we were looking at the same table, but for a number of different universes. The derivation of bounds makes use of the fact that the function in (5) is a Schur-convex function of $\boldsymbol{\Pi}$; for more on Schur-convex functions, see Marshall and Olkin (1979) or Lucero (2010b). The Schur-convexity allows us to identify the most extreme cases, and leads to the following bounds:

$$\left(\frac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\ldots+x_J=q_1} \left(\frac{q_1!}{x_1!\cdot\ldots\cdot x_J!}\right)^2 \left(\frac{1}{J}\right)^{2q_1} \leq \xi_{J,k}(\Pi_1,\ldots,\Pi_J) \leq \frac{1}{k-1} \qquad (6)$$

The righthand portion of inequality (6) says that the probability of an accurate disclosure is at most the probability that the same value of q will be chosen for each of the two tables. The lefthand portion gives a best case for the probability of disclosure, upon which we cannot improve without modifying which cells are in the table or changing k (with the proviso that all probabilities are approximate). In particular, the best case is that all cells of the table include exactly the same proportion of the population, i.e. that $\mathbf{\Pi} = \left(\frac{1}{J},\ldots,\frac{1}{J}\right)$.

5. Other Approaches

Since the idea of a remote access system has been in existence for several years, a number of approaches have been proposed that differ from ours to varying degrees, and we survey some of them here.

Schouten and Cigrang (2003) present a variant of the idea of a remote access system, which allows outstanding versatility, but is also difficult to create and expensive and laborious to maintain. Their proposed system allows users to submit queries by email, written in any of several statistical programming languages. If a query is approved, the user receives the results by email. Before the analysis is performed, an automated system determines the legitimacy of the request, with particularly difficult cases handled manually. As with the MAS, certain types of output are allowed and certain types are not, but since the code is user-generated, rather than generated by the system behind the scenes, it is challenging to identify all unallowable queries. This is especially true because, as the authors emphasize, the validity of a query may depend on information already released as a result of previous successful queries. The authors write, "Computers are simply not fast enough and the construction of a system that fully evaluates the risk of disclosure may be too costly and complex and therefore not feasible." Thus, in a system like this, it may be necessary to perform some disclosure avoidance analysis on a query after the result of the query has already been returned. This is not ideal, as a query that is a disclosure threat might not be identified until its output has already been provided. However, such a method could be effective if the users are from large institutions and have signed a contract describing their research and pledging to uphold confidentiality. In this case, the fear of a user or institution's jeopardizing its future access to the data may serve as a sufficient deterrent to its deliberately submitting an invalid query. In this type of system, a username and password would be necessary so that individual users' actions could be properly tracked.

Sparks et al. (2008) propose a system—Privacy-Preserving Analytics®—that performs a number of methods for disclosure avoidance, including keeping track of the regression models a user requests and ensuring that only a limited (although large) number are run for each possible response variable. They also ensure that a user does not make too many closely related requests.

Gomatam et al. (2005) make a distinction between *static servers* and *dynamic servers*. A static server has a pre-determined set of queries to which it will provide an answer. A dynamic server receives a query and makes a decision on whether to provide an answer. A dynamic server—such as the one described in Schouten and Cigrang (2003)—would keep a running record of all previously answered queries, and whenever a new query was submitted, it would be compared against the list to determine whether providing an answer would lead to a disclosure risk when the new answer was combined with previously provided answers. A dynamic server has the highly undesirable property that the order in which queries are submitted by the collective group of users plays a large role in determining which queries are answered, and that eventually the server reaches a point where no new queries can be answered. Since queries are answered or rejected as they are received, the set of queries that are ultimately answered is not the result of a careful assessment of which analyses would provide the most utility to legitimate researchers while keeping disclosure risk at an acceptable level. Gomatam et al. (2005) write that "[w]hether dynamic servers are possible remains an open question." The MAS is at its heart a static server, since it operates under a set of rules that do not depend on previous queries. However, it operates in a dynamic fashion, since the rules are checked for each new query that is submitted, rather than comparing it to a pre-computed list, as creating such a list would be prohibitive. In a way, the MAS does not fit into the framework of Gomatam et al. (2005), as it sometimes will provide regression output that is less detailed than the user might have liked instead of refusing output altogether.

Another approach to protecting privacy from a query-accepting statistical database is to suppress from any tables any cells that are deemed a disclosure risk, either directly or indirectly. Adam and Worthmann (1989) discuss this possibility and note that in certain systems, cell suppression is not a feasible solution to the disclosure problem.

6. Future Work

The MAS will continue to be developed within DataFERRETT. We will soon be testing the software itself and the confidentiality rules within the MAS beta prototype to ensure that they properly uphold disclosure avoidance standards. We will draft a set of confidentiality rules for cross-tabulations, and add different types of statistical analyses within the system. We will explore other types of differencing attack disclosures, and investigate ways to prevent such differencing attacks. Also of potential interest is doing more theoretical explorations to evaluate disclosure risk. For example, it would be of interest to determine the probability of a correct disclosure given that there is an apparent disclosure resulting from a differencing attack. If this number were small enough, it could lead to a higher level of protection for the system, as an intruder would not be able to be highly confident of the correctness of an apparent disclosure.

References

N. Adam and J. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989. ISSN 0360-0300.

Z. Ben-Haim and T. Dvorkind. Majorization and applications to optimization. Technion Institute, 2004.

M. Chaudhry. Overview of the Microdata Analysis System. Statistical Research Division internal report, U.S. Census Bureau, 2007.

G. Duncan, S. Keller-McNulty, and S. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. In *Chance*. Citeseer, 2001.

I. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972. ISSN 0162-1459.

I. Fellegi, S. Goldberg, and S. Abraham. *Some Aspects of the Impact of the Computer on Official Statistics*. Dominion Bureau of Statistics, 1969.

S. Gomatam, A. Karr, J. Reiter, and A. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statistical Science*, 20(2):163–177, 2005. ISSN 0883-4237.

S. Kaufman, M. Seastrom, and S. Roey. Do disclosure controls to protect confidentiality degrade the quality of the data? In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 1218–1225, 2005.

S. Keller-McNulty and E. Unger. A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14:347–360, 1998. ISSN 0282-423X.

J. Lucero. Confidentiality rules for universe formation and geographies for the Microdata Analysis System. Statistical Research Division Confidential Research Report CCRR–2009/01, U.S. Census Bureau, 2009.

J. Lucero. Confidentiality rule specifications for performing regression analysis on the Microdata Analysis System. Statistical Research Division Confidential Research Report, U.S. Census Bureau, 2010a.

J. Lucero. Evaluation of the effectiveness of the Drop *q* Rule against differencing attack disclosures. Statistical Research Division Confidential Research Report CCRR-2010/03, U.S. Census Bureau, 2010b.

J. Lucero, L. Singh, and L. Zayatz. Recent work on the Microdata Analysis System at the Census Bureau. Statistical Research Division Confidential Research Report CCRR–2009/09, U.S. Census Bureau, 2009a.

J. Lucero, L. Zayatz, and L. Singh. The current state of the Microdata Analysis System at the Census Bureau. In *Proceedings of the American Statistical Association, Government Statistics Section*, 2009b.

A. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*, volume 143. Academic Press New York, 1979.

C. O'Keefe and N. Good. Regression output from a remote analysis server. *Data & Knowledge Engineering*, 68(11):1175–1186, 2009. ISSN 0169-023X.

J. Reiter. Model diagnostics for remote access regression servers. *Statistics and Computing*, 13(4):371–380, 2003. ISSN 0960-3174.

J. Reiter and C. Kohnen. Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75(11):889–903, 2005. ISSN 0094-9655.

A. Reznek. Disclosure risks in cross-section regression models. In *Proceedings of the Section on Government Statistics, JSM*, 2003.

A. Reznek and T. Riggs. Disclosure risks in regression models: Some further results. In *Proceedings of the Section on Government Statistics, JSM*, 2004.

S. Roehrig, S. Bayyana, S. Ganapatiraju, and S. Santhanam. Final report to the Bureau of the Census for the project "Auditing the Census Bureau's confidentiality preserving model server". Contracted report for the U.S. Census Bureau, 2008.

S. Rowland and L. Zayatz. Automating access with confidentiality protection: The American FactFinder. In *Proceedings of the Section on Government Statistics*, 2001.

B. Schouten and M. Cigrang. Remote access systems for statistical analysis of microdata. *Statistics and Computing*, 13(4):381–389, 2003. ISSN 0960-3174.

R. Sparks, C. Carter, J. Donnelly, C. O'Keefe, J. Duncan, T. Keighley, and D. McAullay. Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics®. *Computer Methods and Programs in Biomedicine*, 91(3):208–222, 2008. ISSN 0169-2607.

P. Steel. Design and development of the Census Bureau's Microdata Analysis System: Work in progress on a constrained regression server. Presentation at Federal Committee on Statistical Methodology Statistical Policy Seminar, November 2006.

P. Steel and A. Reznek. Issues in designing a confidentiality preserving model server. *Monographs of Official Statistics*, 9:29, 2005.

D. Weinberg, J. Abowd, P. Steel, L. Zayatz, and S. Rowland. Access methods for United States microdata. U.S. Census Bureau Center for Economic Studies paper CES-WP-07-25, August 2007. Paper for Institute for Employment Research Workshop on Data Access to Micro-Data, Nuremberg, Germany.

X. Zhang. Schur-convex functions and isoperimetric inequalities. *Proceedings of the American Mathematical Society*, 126(2):461–470, 1998. ISSN 0002-9939.

SORT Special issue: Privacy in statistical databases, 2011, 99-112

Multiplicative noise for masking numerical microdata with constraints

Anna Oganian[*]

Georgia Southern University

Abstract

Before releasing databases which contain sensitive information about individuals, statistical agencies have to apply Statistical Disclosure Limitation (SDL) methods to such data. The goal of these methods is to minimize the risk of disclosure of the confidential information and at the same time provide legitimate data users with accurate information about the population of interest. SDL methods applicable to the microdata (i.e. collection of individual records) are often called masking methods. In this paper, several multiplicative noise masking schemes are presented. These schemes are designed to preserve positivity and inequality constraints in the data together with the vector of means and covariance matrix.

MSC: 62Pxx, 62Hxx

Keywords: Statistical disclosure limitation (SDL), SDL method, multiplicative noise, positivity and inequality constraints.

1. Introduction

When statistical offices release information about individuals, they face two conflicting goals: preserve confidentiality of the data—identities of the data subjects and values of sensitive attributes— and at the same time, release useful information for policy, research or other purposes.

Data may be released in two formats: microdata (i.e. collection of individual records) and tabular data. Release of microdata is often considered to be more dangerous from the point of view of the disclosure risk, but at the same time the range of statistical analyses may be wider for the microdata comparative to the tabular data.

[*] Georgia Southern University, Department of Mathematical Sciences, Department of Computer Engineering and Mathematics, P.O.Box 8093, GA, 30460-8093. aoganyan@georgiasouthern.edu

Received: November 2010
Accepted: March 2011

This paper focuses entirely on the microdata releases. Multiple means of access to microdata records exist, including restricted data centers (*e.g.*, ANES; MEPS; SSDS), licensing [NCES] and remote access servers [Gomatam *et al.*]. These are effective, but they do not meet all needs, and many agencies also release deliberately altered microdata publicly.

For public microdata releases, the role of statistical disclosure limitation (SDL) is to alter the data in a way that maintains the utility but limits disclosure risk.

Many Statistical Disclosure Limitation (SDL) methods can be used to prepare microdata releases. Of course, the initial step is to remove explicit identifiers for individuals – names, addresses and social security numbers.

Almost always, removal of identifiers alone is inadequate. Rare attribute combinations (for example, a 17-year old widow) can lead to re-identification. Moreover, in high-dimensional data, virtually every subject may have a unique set of attributes. Therefore, almost invariably, released data attributes must be modified. Some SDL techniques coarsen the resolution of the data; for example, date of birth can be replaced by age, and age may be reported in five-year intervals. Extreme attribute values can be top- or bottom-coded.

Another approach is to generate synthetic records, which are draws from a distribution (typically, a posterior predictive distribution) representing the original data.

Other methods actually change attribute values. Examples are addition of noise, data swapping and microaggregation [Karr *et al.* (2006); Oganian and Karr (2006)]. We term methods whose output is a perturbed version of the original data *the perturbation methods*. This paper focuses on one of these – a perturbation by means of externally generated “noise.” Each specific perturbation method has consequences on both disclosure risk and data utility. Some limit risk effectively but are poor at preserving utility, while others yield high utility, but at the price of high risk. No method is superior with respect to both. Oganian and Karr (2006) show how to combine two methods with the goal of capturing the good aspects of each.

From a data utility perspective, it is important to preserve qualitative characteristics of data, for example, positivity constraints of the form $X \geq 0$ for some variables and inter-attribute relationships such as linear inequalities. Age, many economic variables (gross income, taxes) and many demographic variables (number of employees, number of students in the sixth grade) obey positivity constraints; examples of inequality constraints are “Federal taxes $\leq$ gross income”, “number of salaried employees $\leq$ number of employees” and “year of birth $\leq$ year of death.”

There is also a risk aspect. Because such characteristics are derived from domain knowledge available to both legitimate data users and intruders, failure to preserve them poses a disclosure risk: the extent to which constraints are violated can be informative about the nature and intensity of the SDL applied to the data.

Some SDL methods preserve such characteristics more by coincidence than by design, and only partially. For instance, data swapping preserves positivity, but not multi-attribute constraints. Microaggregation preserves positivity, but whether it preserves linear inequalities depends on specifics of the implementation.

In this paper, we present several SDL methods applicable to numerical data that *preserves positivity constraints, inequality constraints and the first two moments* – the vector of means and covariance matrix.

For the purposes of this paper, the original and released (which we hereafter term masked) databases are flat files in which rows represent data subjects (individuals, households, business establishments, ...) and columns numerical attributes of those subjects. We denote the original data by X_o and the released (masked) data by X_m. We assume that some variables in X_o are nonnegative, others can take positive and negative values. The goal is to obtain $X_m(j) \geq 0$ for those variables j which are nonnegative in the original data, also X_m should have the same mean and covariance matrix as X_o.

As background, the analogous procedure for addition of noise to unconstrained numerical data is as follows. Let Σ_o be the covariance matrix of X_o – in practice, one can use either the usual empirical estimator or a shrinkage-based estimator. Let $k > 0$ be a parameter chosen by the agency; then

$$X_m = E[X_o] + \frac{(X_o - E[X_o]) + E}{\sqrt{1+k}}, \tag{1}$$

where the noise E has distribution $N(\mathbf{0}, k\Sigma_o)$, has the requisite properties [Oganian and Karr (2006)]. Note that the value of k need not be released, even if it were made known that the method of SDL is addition of noise. As $k \to \infty$, X_m becomes a very simplistic form of synthetic data [Reiter (2002)], and any non-normal distributional characteristics of X_o are lost.

The structure of this paper is the following: several multivariate noise protocols that preserve the first two moments are presented in Section 2, close forms for higher order moments are given in Section 3, the extension of these protocols to satisfy inequality constraints is described in Section 4 and the results of the numerical experiments are reported in Section 5.

2. Multiplicative noise protocols

Suppose that X_o contain n records, each with d numerical attributes. Some of the attributes are nonnegative, denote them ${X_o}^p$. We wish to construct and release a masked data set X_m with these characteristics:

$${X_m}^p \geq 0 \tag{2}$$

$$E[X_m] = E[X_o] \tag{3}$$

$$\Sigma(X_m) = \Sigma(X_o), \tag{4}$$

where ${X_m}^p$ are the masked values of ${X_o}^p$ and $\Sigma(\cdot)$ means “covariance matrix of $(\cdot)$.”

Oganian and Karr (2011) proposed a masking scheme which preserves the positivity, means and covariance matrix. The basis of this scheme is to use multiplicative noise, implemented by taking logarithms, applying additive, normally distributed noise and exponentiating. This scheme works only if all the variables in the data set are nonnegative. Below are the details.

Let E be noise that is conditionally independent of X_o given $E[X_o]$ and $\Sigma(X_o)$, and satisfies

$$E[X_o \circ \exp(E)] = E[X_o] \tag{5}$$

$$\Sigma(X_o \circ \exp(E)) = (1+k)\Sigma(X_o), \tag{6}$$

where $k > 0$ is an agency-chosen parameter and $\circ$ denotes elementwise matrix multiplication (Schur or Hadamard product). That is, the exponentiation in (5), (6) and elsewhere below also takes place componentwise. Then

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] + [X_o \circ \exp(E)]}{\sqrt{1+k}} \tag{7}$$

satisfies (2)–(4).

For normally distributed noise E, Oganian and Karr (2011) showed that the following vector of means μ_E and the covariance matrix Σ_E should be chosen for E to satisfy (5) and (6):

$$\Sigma_E(i,j) = \log\left(1 + \frac{k\Sigma_o(i,j)}{E[X_o(i)X_o(j)]}\right), \qquad i,j = 1,\ldots,d \tag{8}$$

$$\mu_E(i) = -\sigma_E(i)/2, \qquad i = 1,\ldots,d. \tag{9}$$

Here, d is the number of the dimensions in the data.

Note that the fact that the original data are multiplied by the lognormal noise does not mean that such a noise introduce a significant skewness to the data. In fact, because of the specific choice of the parameters of the lognormal distribution, the introduced skewness is minimal. In particular, from (8), the variance of the lognormal noise is less than k, where k is a parameter of the method and typically small, *e.g.* 0.15. The lognormal noise with such a small variance is practically symmetrical with very slight skew to the right, to the point that its distribution is almost indistinguishable from a normal distribution.

If the data set contains not only nonnegative variables but variables with negative values as well, the scheme described above cannot be applied directly. The variables with negative and positive values may lead to

$$1 + \frac{k\Sigma_o(i,j)}{E[X_o(i)X_o(j)]} < 0 \tag{10}$$

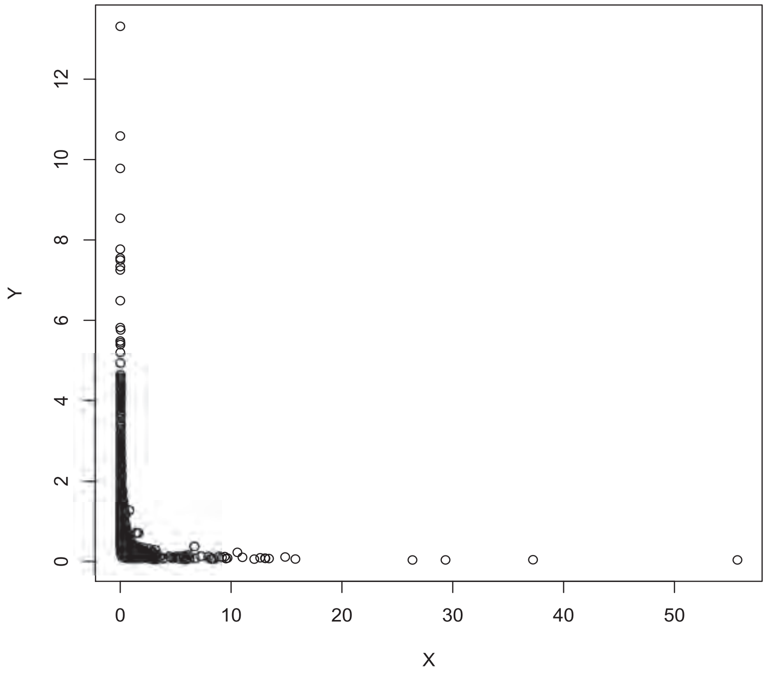

Figure 1: *Example of a data set when covariance matrix for noise cannot be computed.*

so, the covariance matrix (8) cannot be computed. After experimentation with different data sets, it was noticed that for some very rare distributions of values of ${X_o}^p$, (10) may still hold. This will happen when positive variables are negatively correlated and when

$$E[X_o(i)X_o(j)] < \frac{k}{1+k}E[X_o(i)]E[X_o(j)]$$

For positive variables this may happen only when the values of the variables are strongly aligned along the axes. Example of such distribution is shown in Figure 1.

One possible solution to this problem is described in [Oganian (2010)] which consist of converting all the variables to z-scores and making these z-scores nonnegative by adding some value (or vector – for multivariate data) lag, such that $lag \geq |min(Z)|$. Denote these nonnegative z-scores by Z_p. Then masking scheme described by (7), (8) and (9) can be applied to Z_p and after that the resulting data are returned to the original scale:

$$Z_m = \frac{(\sqrt{1+k}-1)lag + [Z_p \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{11}$$

$$X_m = (Z_m - lag) \circ \sigma_o + E(X_o) \tag{12}$$

where σ_o is the main diagonal of Σ_o and the noise E^{z_p} has the following mean and covariance matrix:

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1+\frac{k\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]}\right), \qquad i,j=1,\ldots,d \tag{13}$$

$$\mu_{E_{z_p}}(i) = -\sigma_{E_{z_p}}(i)/2, \qquad i=1,\ldots,d. \tag{14}$$

where $\Sigma_{z_p}(i,j)$ is the (i,j) element of the covariance matrix of Z_p.

Masked data X_m in this case can be represented as

$$\begin{aligned} X_m &= \Big(\frac{(Z_p\circ\exp(E^{z_p})+(\sqrt{1+k}-1)lag}{\sqrt{1+k}} - lag\Big)\circ\sigma_o + E(X_o) = \\ &= \frac{[X_o\circ\exp(E^{z_p})]-E(X_o)\circ exp(E^{z_p})+\sigma_o\circ lag\circ exp(E^{z_p})}{\sqrt{1+k}} + \\ &+ \frac{E(X_o)\sqrt{1+k}-lag\circ\sigma_o}{\sqrt{1+k}} \end{aligned} \tag{15}$$

It is easy to see that such scheme preserves the means and covariance matrix:

$$\begin{aligned} E(X_m) &= \frac{1}{\sqrt{k+1}}[E(X_o)-E(X_o)+\sigma_o\circ lag - \sigma_o\circ lag + \\ &+E(X_o)\sqrt{1+k}] = E(X_o) \end{aligned} \tag{16}$$

The equality in the formula above follows from the fact that the noise is independent from X_o and $E(exp(E^{z_p}))=1$.

$$\begin{aligned} \Sigma_m(i,j) &= \Sigma\Big(\frac{Z_p(i)\exp(E^{z_p}(i))\sigma_o(i)}{\sqrt{1+k}}, \frac{Z_p(j)\exp(E^{z_p}(j))\sigma_o(j)}{\sqrt{1+k}}\Big) = \\ &= \frac{\sigma_o(i)\sigma_o(j)}{1+k}(1+k)cov(Z_p(i),Z_p(j)) = \\ &= \sigma_o(i)\sigma_o(j)cor(X_o(i),X_o(j)) = \Sigma_o(i,j) \end{aligned} \tag{17}$$

where $cov(\cdot)$ and $cor(\cdot)$ denote covariance and correlation of $(\cdot)$ respectively. Note that the second equality in the formula above follows from the property (6).

Oganian (2010) shows that masking scheme (15) with the specific choice for lag will never lead to the case described by (10).

In particular, first, let us see what are the possible values for lag in this scheme. lag should be greater than $|min(Z)|$, however, a very big lag may lead to a negative masked data (this follows from equation(12)), which violates positivity constrants for the variables ${X_o}^p$.

From (11), Z_m is minimized when $E_n \to -\infty$:

$$min(Z_m) > \frac{(\sqrt{1+k}-1)lag}{\sqrt{1+k}}$$

From (12), $min(X_m)$ is larger than

$$\frac{-lag}{\sqrt{1+k}}\sigma_o + E(X_o) \tag{18}$$

To preserve positivity in the masked data, it would be enough to require positivity of (18). So, we have an upper bound for lag:

$$lag \leq \frac{E(X_o)}{\sigma_o}\sqrt{1+k}$$

where division is done componentwise.

The lower bound for lag is $|min(Z)|$. For nonnegative variables with zeros $|min(Z)| = E(X_o)/\sigma_o$. So, the lower and upper bound for lag are:

$$\frac{E(X_o)}{\sigma_o} \leq lag \leq \frac{E(X_o)}{\sigma_o}\sqrt{1+k} \tag{19}$$

Let us consider a few choices for lag in this range. If we choose $lag = E(X_o)/\sigma_o$, then the scheme with z-scores transformation (15) is equivalent to the scheme without transformation (7). In fact, it is straightforward to verify that the masked data in this case can be written as:

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] + [X_o \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{20}$$

Expression (20) is almost identical to (7) except the second term in the nominator: $[X_o \circ \exp(E^{z_p})]$.

Below we will show that even this term is identical in both schemes. In particular, after the application of our masking scheme to the positive z-scores, noise E^{z_p} has the mean and covariance matrix defined by (14) and (13) respectively.

Note, that

$$\frac{\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]} = \frac{cor(X_o(i),X_o(j))}{E[(\frac{(X_o(i)-E(X_o(i)))}{\sigma_o(i)} + lag(i))(\frac{(X_o(j)-E(X_o(j)))}{\sigma_o(j)} + lag(j))]} =$$

$$= \frac{cor(X_o(i),X_o(j))}{E[X_o(i)/\sigma_o(i) * X_o(j)/\sigma_o(j)]} = \frac{\Sigma_o(i,j)}{E[X_o(i)X_o(j)]}$$

So, when $lag = E(X_o)/\sigma_o$, the transformation to positive $z-$scores does not make any changes in the original scheme (7).

Now let us consider another extreme for lag: $lag = \sqrt{(}1+k)E(X_o)/\sigma_o$.

It is easy to verify that masked data in this case can be written as:

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] \circ \exp(E^{z_p}) + [X_o \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{21}$$

Covariance matrix for the noise for this scheme is:

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1 + \frac{k\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]}\right) \tag{22}$$

To prove that the expression under logarithm of (22) is always positive, let's express it in terms of original data.

$$Z_p(i) = \frac{X_o(i) + E(X_o(i))(\sqrt{1+k}-1))}{\sigma_o(i)}$$

It is easy to see that

$$E[Z_p(i)Z_p(j)] = \frac{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}{\sigma_o(i)\sigma_o(j)}$$

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1 + \frac{k\sigma_o(i)\sigma_o(j)cor(X_o(i),X_o(j))}{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}\right) =$$

$$= \log\left(\frac{(1+k)E[X_o(i)X_o(j)]}{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}\right) \tag{23}$$

The expression under the logarithm in (23) is always positive for the nonnegative X_o, so we can always compute $\Sigma_{E_{z_p}}$. In the same way, it is possible to show that no other value for lag (in the range of its possible values) can guarantee positivity of (10) for all possible data sets.

When the data set contains variables which can take positive and negative values together with nonnegative variables, the scheme with z-scores transformations will work too. First the data should be made nonnegative by adding $|min(X_o)|$ and then scheme (21) is applied to this data. Last, to return the data to the original location, we have to substract $|min(X_o)|$ from the result of the previous step.

3. Preservation of higher moments

Multivariate noise protocols described in Section 2 maintain positivity and the first two moments. Exact preservation of higher-order moments is not guaranteed. Here we consider the extent to which higher-order moments can be distorted by the scheme with z-scores transformation, which has a wider range of applicability than the scheme without z-scores transformation

Consider the P-th (mixed) moment, $E(X_{m_1}^{p_1}X_{m_2}^{p_2}\cdots X_{m_d}^{p_d})$, where $P=\sum_{i=1}^{d}p_i$:

$$\begin{aligned}
&E[\prod_{j=1}^{d}X_m(j)^{p_j}] = E\Big[\prod_{j=1}^{d}(\frac{(\sqrt{1+k}-1)E(X_o(j))\exp(E^{z_p}(j))+}{\sqrt{1+k}}\\
&+\frac{X_o(j)\exp(E^{z_p}(j))}{\sqrt{1+k}})^{p_j}\Big] = \frac{1}{(\sqrt{1+k})^{\sum_{j=1}^{d}p_j}}E\Big[\prod_{j=1}^{d}\Big(\sum_{i_j=0}^{p_j}\binom{p_j}{i_j}\times\\
&\times(\sqrt{1+k}-1)^{p_j-i_j}E^{p_j-i_j}(X_o(j))\exp((p_j-i_j)E^{z_p}(j))X_o(j)\times\\
&\times\exp(i_jE^{z_p}(j))\Big)\Big] = \frac{1}{(\sqrt{1+k})^{\sum_{j=1}^{d}p_j}}\sum_{i_1=0}^{p_1}\cdots\sum_{i_d=0}^{p_d}E\Big[\prod_{j=1}^{d}X_o(j)^{i_j}\Big]\times\\
&\times E\Big[\exp(\sum_{j=1}^{d}p_jE^{z_p}(j))\Big]\times(\sqrt{1+k}-1)^{\sum_{j=1}^{d}(p_j-i_j)}\prod_{j=1}^{d}\binom{p_j}{i_j}E^{p_j-i_j}[X_o(j)].
\end{aligned}$$

Note that $W=\sum_{j=1}^{d}p_jE^{z_p}(j)$ is a weighted sum of d normal variables that are not independents but are jointly normal. So, W is a normal variable too. Thus, $\exp(W)$ is log-normal with mean equal to $\exp(\mu_W+0.5Var_W)$. Then, since

$$\begin{aligned}
&E[\exp(W)]=\\
&=\exp\left[\sum_{j=1}^{d}p_j\mu_E(j)+0.5\left(\sum_{j=1}^{d}p_j^2\Sigma_{E_{z_p}}(jj)+\sum_{j<l}2p_jp_l\Sigma_{E_{z_p}}(jl)\right)\right]\\
&=\exp\left[-0.5\sum_{j=1}^{d}i_j\Sigma_{E_{z_p}}(jj)+0.5\left(\sum_{j=1}^{d}i_j^2\Sigma_{E_{z_p}}(jj)+\sum_{j<l}2i_ji_l\Sigma_{E_{z_p}}(jl)\right)\right]\\
&=\prod_{j=1}^{d}\left(\frac{(1+k)E[X_o(j)^2]}{E[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}}\prod_{j<l}\left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_jp_l}
\end{aligned} \tag{24}$$

we obtain

$$E\left[\prod_{j=1}^{d} X_m(j)^{p_j}\right] = \sum_{i_1=0}^{p_1}\cdots\sum_{i_d=0}^{p_d} E\left[\prod_{j=1}^{d} X_o(j)^{i_j}\right] \frac{(\sqrt{1+k}-1)^{\Sigma_{j=1}^{d}(p_j-i_j)}}{(\sqrt{1+k})^{\Sigma_{j=0}^{d} p_j}} \times$$

$$\prod_{j=1}^{d}\binom{p_j}{i_j} \times \prod_{j=1}^{d}\left(\frac{(1+k)E[X_o(j)^2]}{E[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}} \times \tag{25}$$

$$\times \prod_{j<l}\left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_j p_l} \prod_{j=1}^{d} E^{p_j-i_j}[X_o(j)].$$

Now, (25) can be written as

$$E\left[\prod_{j=1}^{d} X_m(j)^{p_j}\right] = E\left[\prod_{j=1}^{d} X_o(j)^{p_j}\right] \frac{A}{(\sqrt{1+k})^{\Sigma_{j=1}^{d} p_j}} + \sum_{i_1=0}^{u_1}\cdots\sum_{i_d=0}^{u_d} E\left[\prod_{j=1}^{d} X_o(j)^{i_j}\right] \times$$

$$\times \prod_{j=1}^{d}\binom{p_j}{i_j} E^{p_j-i_j}[X_o(j)] \frac{A(\sqrt{1+k}-1)^{\Sigma_{j=1}^{d}(p_j-i_j)}}{(\sqrt{1+c})^{\Sigma_{j=1}^{d} p_j}}, \tag{26}$$

where $u_1 \in \{(p_1-1), p_1\}$, $u_2 \in \{(p_2-1), p_2\} \cdots$ $u_d \in \{(p_d-1), p_d\}$, such that $u_1, u_2 \cdots u_d \neq \{p_1, p_2 \cdots p_d\}$ and

$$A = \prod_{j=1}^{d}\left(\frac{(1+k)E[X_o(j)^2}{E\,[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}} \times$$

$$\times \prod_{j<l}\left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_j p_l}$$

From (26) we see how the moments of the original and masked data are related. If the agency decides to release information about masking algorithm – in particular the value of k, then this formula can be reported to data users, allowing them to adjust their analyses and to calculate the original moments. To compute the original moments users would employ expression (26) recursively: first and second order moments of the original data in (26) can be substituted by the corresponding moments computed on the masked data. All higher order original moments can be computed recursively using formula (26). However, the safety of the releasing k is problematic in some scenarios, because doing so might lead to attribute disclosure risk for some records.

A question of practical interest is how large the expression (26) can be, compared to the corresponding original moments. Because the masked data are scaled to have the same covariance matrix as the original data, higher-order moments seem unlikely to be grossly inflated, but it is possible. In most our experiments with different data sets, third-order moments were only 2.5% larger than the original moments on average for skewed original data with outliers, such as the lognormal data sets described in Section 5). For the symmetrical data sets with the same covariance matrix as the lognormal ones, they were only .15% larger than the corresponding original ones. Fourth-order moments were about 15% larger on average for the lognormal original data and only .8% larger for the symmetrical data. In general, the discrepancy increases with the order of the moment, but only slowly.

4. Inequality constraints preservation

Suppose our original data in addition to positivity constraints also have inequality constraints of the form $X > Y$. For example, masking an income data with the variables “Gross income” and “Federal taxes“ should produce a masked data such that “Gross income $>$ Federal taxes”. The protocols described above can be used as building blocks of a new scheme which would guarantee the preservation of inequality constraints. This scheme is the following:

- Apply the multiplicative noise scheme to $(Y_o, [X_o - Y_o])$. Denote the result by $(Y^*, [X_o - Y_o]^*)$
- The masked data corresponding to (X_o, Y_o) are $(X_m, Y_m) = (Y^* + [X_o - Y_o]^*, Y^*)$

It is easy to see that this scheme preserves the means and covariance matrix.

$$\begin{aligned}
E(X_m, Y_m) &= E(Y^* + [X_o - Y_o]^*, Y^*) = \\
&= (E(Y_o) + E[X_o - Y_o]), (E(Y_o)) = (E(X_o), E(Y_o)) \\
cov(X_m, Y_m) &= cov(Y^* + [X_o - Y_o]^*, Y^*) = var(Y_o) + \\
&+ cov([X_o - Y_o]^*, Y^*) = var(Y_o) + cov([X_o - Y_o], Y_o) = \\
&= var(Y_o) + cov(X_o, Y_o) - var(Y_o) = cov(X_o, Y_o) \\
var(X_m) &= var((Y^* + [X_o - Y_o]^*) = var(Y_o) + var([X_o - Y_o]) + \\
&+ 2cov(Y_o, [X_o - Y_o]) = var(X_o)
\end{aligned}$$

The scheme can be readily extended for the cases when multiple variables are related by inequality constraints. For example, suppose
$X_{o_1} > X_{o_2} > X_{o_3}$, then $X_{m1} = X_3^* + [X_{o_2} - X_{o_3}]^* + [X_{o_1} - X_{o_2}]^*$, $X_{m2} = X_3^* + [X_{o2} - X_{o3}]^*$ and $X_{m3} = X_3^*$.

Or in general case if $X_{o_1} > X_{o_2} > \cdots > X_{o_{l-1}} > X_{ol}$

$$X_{mi} = X_l^* + \sum_{j=i+1}^{l} [X_{o_{j-1}} - X_{o_j}]^* \tag{27}$$

5. Numerical experiments

Both multiplicative noise schemes (with and without z-scores transformation) were implemented and evaluated on different data sets. These data sets have different distributional characteristics: a skewed distribution with many outliers and a symmetrical one without outliers. The symmetrical data sets had a multivariate normal distribution and the skewed sets were log-normally distributed. 500 replicates of three-dimensional normal and lognormal sets were generated. Each set had $10,000$ records. They were moderately correlated ($cor = 0.5$). The log-normal sets had means around 2 and variances ranging from 4 to 16. These sets had outliers – values close to 50 or larger.

The normal sets had means around 3.5 and variances ranging from 5 to 10. The variance inflation factor k was chosen to be 0.15 as recommended in Oganian (2003).

The experiments showed that means were very well preserved for both schemes and both types of data: the ratio of masked and original means showed only a very small variation around 1. The results on variance/covariance matrix were different for skewed and symmetrical data sets. The experiments showed that covariance matrix was preserved for the symmetrical data sets without outliers. There was slight variability in variance/covariance matrix inflation, defined as Σ_m / Σ_o, where $/$ denotes elementwise division. Values of this ratio ranged from 0.98 to 1.02.

There was more variability in variance/covariance matrix inflation for the skewed data sets with outliers. Values of this ratio ranged approximately from 0.7 to 1.3. The scheme with z-transformation resulted to be slightly more stable: variance/covariance inflation ranged approximately from 0.8 to 1.2. However, the average and most frequent value were 1 in both schemes and both types of data sets, as expected.

Such variability over replications is not very surprising in light of the nature of the noise and the variation in log-normal original data, which as noted above had a number of large outlying values. Records in the original data with big values – especially outliers – can undergo significant changes when multiplied by noise, distorting the covariance matrix.

One possible solution to reduce variability in the resulting masked data when the original is skewed and/or has many outliers is to apply different levels of noise to different zones of the data, as discussed in Oganian and Karr (2011). It is illustrated in the Figure 2, where zone 1 is masked with the parameter k_1 and zone 2 with the parameter $k_2 < k_1$. Because all the protocols presented in Section 2 are designed to preserve the mean and covariance matrix of the original data, we can apply different independent noises to different zones of the data and the covariance matrix of the masked data should be the same as that of the original data.

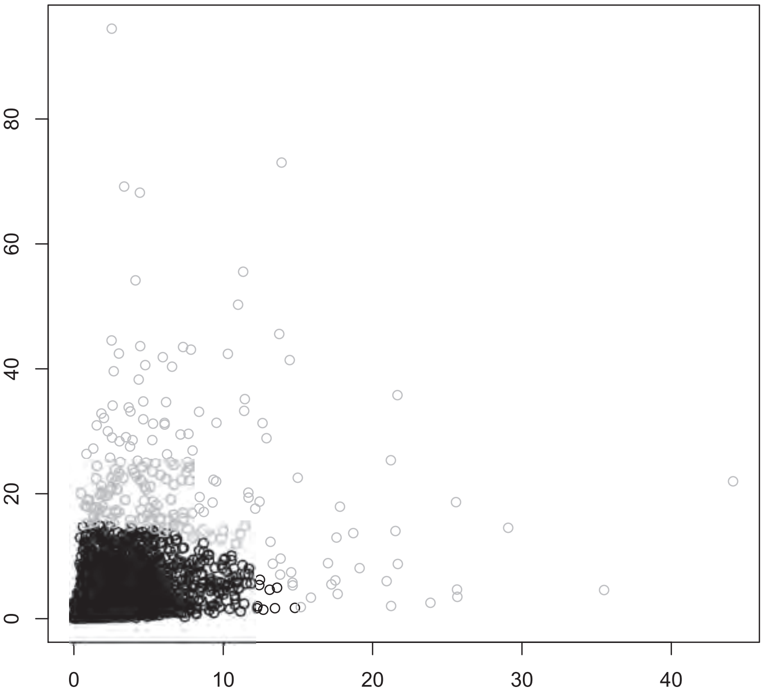

Figure 2: *Two zones of masking: black points correspond to the first zone of masking and grey points to the second zone.*

Two-zone masking was implemented with different values of k for the same three-dimensional lognormal data as in the experiment with only one zone. The first zone consisted of all the points from 0 to 15; all the other records were included in the second zone. For the second zone we chose $k_2 = 0.01$. For the first zone we chose $k_1 = 0.15$.

This approach reduced variability in the covariance matrix of the skewed data significantly: in 95% of replicates of the masked data X_m/X_o was in the interval of $[0.98, 1.02]$.

Optimal ways of variability reduction in the masked data when the original have outliers and severe skewness are the subject of our current and future research.

Note, that the multiple-zone masking may be used for other goals. For example, suppose a numerical variable in the data set has a lot of zeros, which happens often in the household data. Suppose the same numerical variable is paired with an indicator variable I, such that when $I = 0$, it is strictly positive and when $I = 1$, it is zero. Examples of I are “In the labor force” or “Income is greater than taxable min”. If the agency wants to preserve such a relationship in the masked data, they can separately mask records paired with different values of the indicator variable leaving zeros in the numerical variable unchanged. Again, because our protocols preserve means and the covariance matrix, the first two moments of the overall data should be preserved.

Last, we want to discuss the disclosure risk associated with the method. Our measures of disclosure risk focus on re-identification disclosure risk. Re-identification disclosure is defined as an average percentage of correctly identified records when record linkage techniques [Jaro (1989)] are used to match the original and masked data. Specifically, we assume that the intruder tries to link the masked file with an external database containing a subset of the attributes present in the original data [Oganian (2003)]. The overall re-identification risk of the multiplicative noise is very small. Our experiments showed that only about 0.3% of records could be correctly identified in both schemes. So, the multiplicative noise can be successfully compared with the most protective methods, like microaggregation and rank swapping, at the same time performing significantly better than those in terms of utility.

Acknowledgments

This research was partly funded by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS). The sincerest thanks go to Alan Karr.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

ANES. American National Election Studies Resticted Data Access, http://www.electionstudies.org/rda/anes_rda.htm

Gomatam, S., Karr, J. P. A. F., Reiter, J. P. and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers, *Statistical Science*, 20, 163–177.

Jaro, A. M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414–420.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil A. P. (2006). Framework for Evaluating the Utility of Data Altered to Protect Confidentiality, *The American Statistician*, 60, 224–232.

MEPS Medical Expenditure Panel Survey, Restricted Data Files Available at Data Centers, http://www.meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp.

NCES Confidentiality procedures, http://nces.ed.gov/StatProg/confproc.asp.

Oganian, A. (2003). *Security and Information Loss in Statistical Database Protection*, PhD thesis, Universitat Politecnica de Catalunya.

Oganian, A. (2010). Multiplicative Noise Protocols, *Privacy in Statistical Databases 2010, Lecture Notes in Computer Science*, 6344, 107–117.

Oganian, A. and Karr, A. F. (2006). Combinations of SDC Methods for Microdata Protection, *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science*, 4302, 102–113.

Oganian, A. and Karr, A. F. (2011). Masking Methods that Preserve Positivity Constraints in Microdata, *Journal of Statistical Planning and Inference*, 141, 31–41.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets, *Journal of Official Statistics*, 18, 531–544.

SSDS Social Science Data Services, http://libraries.mit.edu/guides/subjects/data/access/restricted.html.

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.cat) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a $\LaTeX\ 2_{\varepsilon}$.

In any case, upon request the journal secretary will provide authors with $\LaTeX\ 2_{\varepsilon}$ templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/˜feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name

Organisation

Street Address

Zip/Postal code City

State/Country Tel.

Fax NIF/VAT Registration Number

E-mail

Date

Signature

I wish to subscribe to **SORT** ***(Statistics and Operations Research Transactions)*** for the year 2011 (volume 35)

Annual subscription rates:
– Spain: €22 (4% VAT included)
– Other countries: €25 (4% VAT included)

Price for individual issues (current and back issues):
– Spain: €15/issue (4% VAT included)
– Other countries: €17/issue (4% VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number
□□□□ □□□□ □□ □□□□□□□□□□

☐ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT ***(Statistics and Operations Research Transactions)***

The undersigned ____

authorises Bank/Financial institution ______________________________

located at (Street Address) _______________________________

Zip/postal code __________________ City __________________

Country ____

to draft the subscription to **SORT** ***(Statistics and Operations Research Transactions)*** from my account

number □□□□ □□□□ □□ □□□□□□□□□□

Date ____________________

Signature

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45