

# **k-anonymization May Be NP-Hard, but Can it Be Practical?**

**David Wilson**

RTI International

dwilson@rti.org

## **1 Introduction**

This paper discusses the application of k-anonymity to a real-world set of microdata with particular emphasis on computational complexities and data quality. A set of microdata is said to provide k-anonymity “when the contained data do not allow the [microdata] recipient to associate the released information to a set of individuals smaller than k.”[1]

While it is well-known that producing a k-anonymous microdata set is computationally difficult [2], there is very little information on application of k-anonymity to real-world microdata. This paper attempts to examine the practicality of producing a k-anonymous microdata set and assess whether or not the resulting microdata can be used to produce estimates reasonably close to estimates produced using the original microdata. This paper does not take the position that k-anonymity, by itself, is sufficient to limit risk of disclosure from all possible intrusion scenarios; rather, this paper seeks to explore whether or not k-anonymity microdata sets may be produced in a reasonable amount of time and whether k-anonymity may be used as part of a larger statistical disclosure treatment process.

An overview of k-anonymity and a discussion of the specific “k” used in the application of k-anonymity are given in Section 2. The microdata, selected quasi-identifiers, and computational processing times are discussed in Section 3. Data quality impacts are described and discussed in Section 4. Conclusions are discussed in Section 5.

## **2 K-anonymity**

K-anonymity was introduced in [1] and may be seen as a generalization of statistical disclosure limitation methods discussed in [3]. Specifically, Willenborg and de Waal discuss the concept of “minimum unsafe combinations [(MINUCS) of the values of variables in a microdata set]” which refers to a specific set of unique combinations of values of all subsets of a group of variables. The unique combinations in this set have the property where setting any single value in a M-dimensional unique combination to missing produces an M-1 dimensional combination that is not unique. Willenborg and de Waal discuss how to minimize the number of local suppressions required to produce a microdata set that has no minimum unsafe combinations and their approach is one way to implement 2-anonymity.

K-anonymity, as discussed in [1], extends the concepts discussed in [3] by describing a process that produces a microdata set, or table, where each combination of values of a set or subset of a group of variables (quasi-identifiers in their terminology) occurs k ( $k \geq 2$ ) or more times, depending upon the desired value of k. Furthermore, methods to remove unique combinations of variables are extended from local suppression to a broader method of data generalization that includes variable recoding.

In order to apply k-anonymity, one must select a value of k and a method for removing combinations of a set or subset of variables that occur less than k times. A value of 2 was selected for the application of k-anonymity used for this paper and local value suppression was used to remove unique combinations of a set or subset of variables.

When local suppression is used, the process of creating a 2-anonymous microdata set consists of three steps:

- 1) Identify the set of variables, referred to as quasi-identifiers, which contain data assumed to be known or available in other data sources, to microdata recipients. These variables do not include direct identifiers such as names or addresses; such variables are assumed to be excluded from the microdata.

- 2) Identify the minimum unsafe combinations (MINUCS) of the quasi-identifiers. The set of MINUCS is a subset of all unique combinations of values of all subsets of the quasi-identifiers. If there are  $Q$  quasi-identifiers, this step requires the creation and review of all  $2^Q - 1$  frequency tables in the following fashion.
  - a. Create all  $Q$  one-way frequency tables and identify values that are unique. These unique values define the set of one-dimensional MINUCS.
  - b. Create all  $\binom{Q}{2}$  two-way tables and identify combinations of values that are unique and where neither of the values is unique in the corresponding one-way table. These unique values define the set of two-dimensional MINUCS.
  - c. Create all  $\binom{Q}{3}$  three-way tables and identify combinations of values that are unique and where none of the values is unique in the corresponding one-way table and none of the pairs of values is unique in the corresponding two-way tables. These unique values define the set of three-dimensional MINUCS.
  - d. Continue the process until all  $2^Q - 1$  tables have been produced and the set of MINUCS has been created.

The term “minimum” in MINUCS refers to the fact that when a single value of an  $M$ -dimensional MINUC is suppressed, then the resulting  $M - 1$  dimensional combination of values is not unique. Note that a MINUC is associated with one and only one record in a microdata set. Accordingly, the set of MINUCS may be partitioned into disjoint non-overlapping sets where each set is associated with one and only one record. This partitioning simplifies that process of determining the minimum number of local suppressions required to produce 2-anonymity.

- 3) For each record in the microdata set, identify those MINUCS associated with that record, and use Boolean integer programming to determine the minimum number of local suppressions per record and then apply those suppressions to produce the 2-anonymized microdata set.

Note that it is assumed that a missing value arising from a local suppression is not informative and therefore does not need to be used to determine if 2-anonymity has been achieved. The reason this assumption is required is because, theoretically, application of the algorithm described above could produce combinations of missing and non-missing values that are unique. In the simplest case, one record could be the only record with a missing value for a given variable. Such a record would not be identical to another record in the data set so the data set would technically violate 2-anonymity. The application of 2-anonymity described herein permits such deviations from  $k$ -anonymity as discussed in [1].

### 3 Source Data and Methods

#### 3.1 Data

The data for this study come from the public-use microdata of the second follow-up of the Education Longitudinal Study of 2002 (ELS:2002)<sup>1</sup>. A student-level data file was downloaded using the National Center of Education Statistics’ online Education Data Analysis Tool (EDAT<sup>2</sup>). The student-level file contained hundreds of variables and 16,197 records (students).

The first step in creating 2-anonymity was to identify variables that could potentially contain data known to, or able to be learned by, data recipients. The variables in the student-level file were reviewed in order to subjectively assess the degree to which the information contained in those variables is “visible.” The term “visible” is used to specify variables whose values may somewhat be inferred by relatively limited observation of a person or school. For example, one may generally tell a person’s sex and approximate age by visual observation. Similarly, one may know the geographic location of school by simply observing the school. We do not assume that data sources exist

<sup>1</sup> <http://nces.ed.gov/surveys/els2002/>

<sup>2</sup> <http://nces.ed.gov/edat/>

that include all of the information about all students and schools from which the ELS:2002 schools and students were sampled; rather, it is likely that no such overall data sources exist. Furthermore, it must be strongly emphasized that the data in all ELS:2002 public-use microdata files have been treated with statistical disclosure limitation methods so it is impossible to say if any particular value of any of the selected variables is true or not.

Table 1 lists and describes the 19 variables selected for the application of 2-anonymity. In order to simplify the process of identifying unique combinations of the 19 quasi-identifiers, student records were excluded when any one of the 19 quasi-identifiers was missing. 11,204 of the 16,197 student records had values for all nineteen identifiers<sup>3</sup>.

**Table 1. Variables Used As Quasi-Identifiers**

<b>Variable</b>	<b>Label (Number of Values)</b>
BYSEX	Sex-composite (2)
BYRACE	Student's race/ethnicity-composite (7)
BYHOMLNG	Student's native language-composite (6)
BYSTLNG2	Sample member's English fluency (5)
BYPARACE	Parent's race/ethnicity-composite (7)
BYPARLNG	Parent's native language-composite (6)
BYPLANG	Parent's English fluency (5)
BYFCOMP	Family composition (9)
BYPARED	Parents' highest level of education (8)
BYMOTHED	Mother's highest level of education-composite (8)
BYFATHED	Father's highest level of education-composite (8)
BYGPARED	Highest reported level of education among parents' parents (8)
BYOCCUM	Mother/female guardian's occupation-composite (17)
BYOCCUF	Father/male guardian's occupation-composite (17)
BYINCOME	Total family income from all sources 2001-composite (13)
BYSCTRL	School control (3)-Public, Catholic, Other Private
BYURBAN	School urbanicity (3)
BYREGION	Geographic region of school (4)
BYDOB_P <sup>1</sup>	Student's Year and Month of Birth (6)

### 3.2 Identifying Unique Combinations of Quasi-identifiers

The process of determining which combinations of values of the 19 quasi-identifiers were unique required the construction, and review, of each frequency table for each of the  $2^{19}$  possible combinations of the 19 variables.

Following the process outlined in [3], all one-way frequency tables were produced and reviewed. Unique values, and the record in which they appeared, were noted and recorded. All two-way frequency tables were produced and reviewed. Unique combinations of values that did not include unique values from any one-way table were noted and recorded. All three-way frequency tables were produced and reviewed. Unique combinations of values that did not include unique values from any one-way table or unique combinations of values from any two-way table were noted and recorded. This process was implemented similarly for four-way through 19-way frequency tables.

<sup>3</sup> There was one exception. 26 of the 11,204 records had a missing value for Year of Birth but were included in the analysis. A numeric value of 0, indicating “missing”, was used for these 26 records.

Initial attempts at constructing frequency tables were implemented using SAS software but the running-time was prohibitive and not practical. A standalone C++ program was created using Microsoft Visual Studio 2010 Express in order to produce frequency tables for the 11,204 records and 19 variables describe above. The process of calculating and recording unique values took approximately 9 hours to run on a desktop computer running 64-bit Windows 7 on an Intel Core i7 940 CPU with 6 gigabytes of random access memory.

Note that the naïve implementation of this algorithm, namely calculating all possible frequency tables, would require approximately double the running time if one additional quasi-identifier variable was added; so twenty variables would take about 18 hours instead of the 9 hours that nineteen variables required. The running-time required to identify all MINUCS could be reduced by noting that any combination of variables that has only unique values does not need to be included in higher dimensional tables. Furthermore, any unique combination of a set of variables does not need to be included in subsequent higher dimension tables.

Because the identification of MINUCS is somewhat sequential in nature, running-time could also be reduced by using multiple computers or CPUs to generate all tables of a particular dimension. Of course, for tables of a given dimension, all CPUs would need to have access to all lower dimension frequency tables in order to determine which combinations should be added to the set of MINUCS.

Since MINUCS are associated with a single record, the algorithm for identifying MINUCS requires space for identifying the combination of variables that produces a MINUC and space to identify the record containing the values of that MINUC. Each MINUC recorded by the C++ program required 19 bits to store the particular combination of variables associated with the MINUC and an integer of 2 bytes was used to store the record number; for a total of 35 bits.

### 3.3 Identifying Values to Suppress

2-anonymity may be achieved by suppressing those combinations of values that are unique as in [4] or by generalizing (recoding) the unique values, alone or possibly in conjunction with non-unique values, as in [1]. In this study, suppression was used to achieve 2-anonymity with one slight deviation from the suppression described in [4]; instead of suppressing all of the values in a unique combination, only some values were suppressed as long as none of the post-suppression non-missing combinations of values were unique.

For illustration, suppose the combination (1, 2) for two variables was unique in a data set. Full suppression would set both of these values to missing. In this paper, we allow the possibility that only one of the values is suppressed. If, for example, the value of 1 was suppressed, we would allow this to occur if the value 2 occurred more than once.

In order to minimize the number of value suppressions, we note that any unique combination of values of the 19 variables occurs in one and only one record. Therefore, minimizing the total number of values suppressed may be achieved by minimizing the number of suppressed values for each record independently.

Determining the minimum number of suppressions per record was solved by forming a binary integer programming problem for each record and minimizing the number of suppressions subject to the requirement that at least one value in each unique combination of values was suppressed. SAS software version 9.2 and its OPTMODEL procedure was used to solve the 11,148<sup>4</sup> binary integer programming problems and was able to find solutions to all problems in approximately five hours.

The time required to find the minimum number of suppressions was primarily driven by the number of records, not the number of variables. Doubling the number of records would roughly double the running-time, but doubling the number of variables would most likely not double the running time.

Unlike the process required to identify MINUCS, the process of finding the minimum number of suppressions is highly parallelizable. Conceptually, one could assign one record to each of 11,148 CPUs and run all 11,148 minimizations in parallel.

Unfortunately, this optimization problem becomes much more complex if more than 2-anonymity is desired. Theoretically, a single optimization routine could be created that uses all records at once but the number of constraints and number of terms in the optimization function would be prohibitively large. Alternatively, the set of

---

<sup>4</sup> 56 records were not unique with respect to the full combination of 19 quasi-identifiers so no data had to be suppressed for these records.

records could be partitioned into disjoint sets where records in a set share one or more MINUCS. This does not occur for 2-anonymity because a MINUC is associated with one and only one record; but, under 3-anonymity or higher, MINUCS can be associated with more than one record. A separate optimization process could be applied to each of the disjoint sets.

#### **4 Impact on Data Quality**

In surveys like ELS:2002, missing data are often imputed. Once the data values were identified for suppression and set to missing, an imputation procedure could have been applied. Imputation may have produced a post-treated microdata set that was closer to the original microdata set than just the 2-anonymized microdata. The primary reason imputation was not performed was because imputation would have made it difficult to separate out the impact of optimal suppression with the impact of imputation. Furthermore, any imputation would have had to make sure to not undo the suppression by replacing a suppressed value with the original value. So, for example, if the value of “Male” was suppressed then the imputation procedure would only have been allowed to impute a value of “Female”; this would be disconcerting to say the least.

The impact on data quality was examined by calculating the percent and count of suppressed values by variable and by examining which values were suppressed most often. Since the ELS:2002 sample is designed to support inferences about target populations of students, one way to measure the impact on data quality is to calculate and compare weighted estimates using the original data with weighted estimates using the suppressed microdata.

##### **4.1 What was Suppressed?**

The percentage of records with a suppressed value of a given variable varied from a minimum of 2.4% for Parent’s native language requirement to a maximum of 49.8% for Father/male guardian’s occupation-composite. The percentage of suppressed records by variable is shown in Table 2.

**Table 2. Percentage<sup>1</sup> and Count of Records with a Suppressed Value by Variable**

<b>Variable</b>	<b>Label</b>	<b>Percent and Count</b>
BYSEX	Sex-composite	8%/859
BYRACE	Student's race/ethnicity-composite	10%/1,165
BYHOMLNG	Student's native language-composite	3%/315
BYSTLNG2	Sample member's English fluency	4%/436
BYPARACE	Parent's race/ethnicity-composite	7%/806
BYPARLNG	Parent's native language-composite	2%/272
BYPLANG	Parent's English fluency	3%/379
BYFCOMP	Family composition	18%/1,967
BYPARED	Parents' highest level of education	6%/643
BYMOTHED	Mother's highest level of education-composite	15%/1,678
BYFATHED	Father's highest level of education-composite	13%/1,412
BYGPARED	Highest reported level of education among parents' parents	28%/3,170
BYOCCUM	Mother/female guardian's occupation-composite	47%/5,210
BYOCCUF	Father/male guardian's occupation-composite	50%/5,581
BYINCOME	Total family income from all sources 2001-composite	43%/4,803
BYSCTRL	School control	9%/1,004
BYURBAN	School urbanicity	19%/2,163
BYREGION	Geographic region of school	20%/2,239
BIRTHYR (bydob_p)	Birth Year	19%/2,114

The values of each of the 19 variables were suppressed at different rates. Table 3 shows the value of each variable that was suppressed the most along with the number of records that had that value suppressed. For a given variable, the displayed percentage is out of all records that had a value of that variable suppressed.

**Table 3. Variable Value Suppressed the Most among Records with Suppressed Values**

<b>Variable</b>	<b>Label</b>	<b>Value</b>	<b>Percent and Number of Suppressed Values</b>
BYSEX	Sex-composite	Male	56%/478
BYRACE	Student's race/ethnicity-composite	More than one race, non-Hispanic	28%/327
BYHOMLNG	Student's native language-composite	Other language	34%/107
BYSTLNG2	Sample member's English fluency	Fluent	43%/188
BYPARACE	Parent's race/ethnicity-composite	Hispanic, race specified	26%/207
BYPARLNG	Parent's native language-composite	Other language	36%/98
BYPLANG	Parent's English fluency	Fluent	54%/203
BYFCOMP	Family composition	Mother only	25%/488
BYPARED	Parents' highest level of education	Attended college, no 4-year degree	21%/136
BYMOTHED	Mother's highest level of education-composite	Attended 2-year school, no degree	17%/292
BYFATHED	Father's highest level of education-composite	Attended 2-year school, no degree	18%/249
BYGPARED	Highest reported level of education among parents' parents	Graduated from high school or GED	15%/472
BYOCCUM	Mother/female guardian's occupation-composite	Service	12%/627
BYOCCUF	Father/male guardian's occupation-composite	Manager, administrator	12%/644
BYINCOME	Total family income from all sources 2001-composite	\$35,001-\$50,000	16%/761
BYSCTRL	School control	Other private	42%/420
BYURBAN	School urbanicity	Urban	38%/820
BYREGION	Geographic region of school	West	27%/602
BIRTHYR (bydob_p)	Birth Year	1985	46%/978

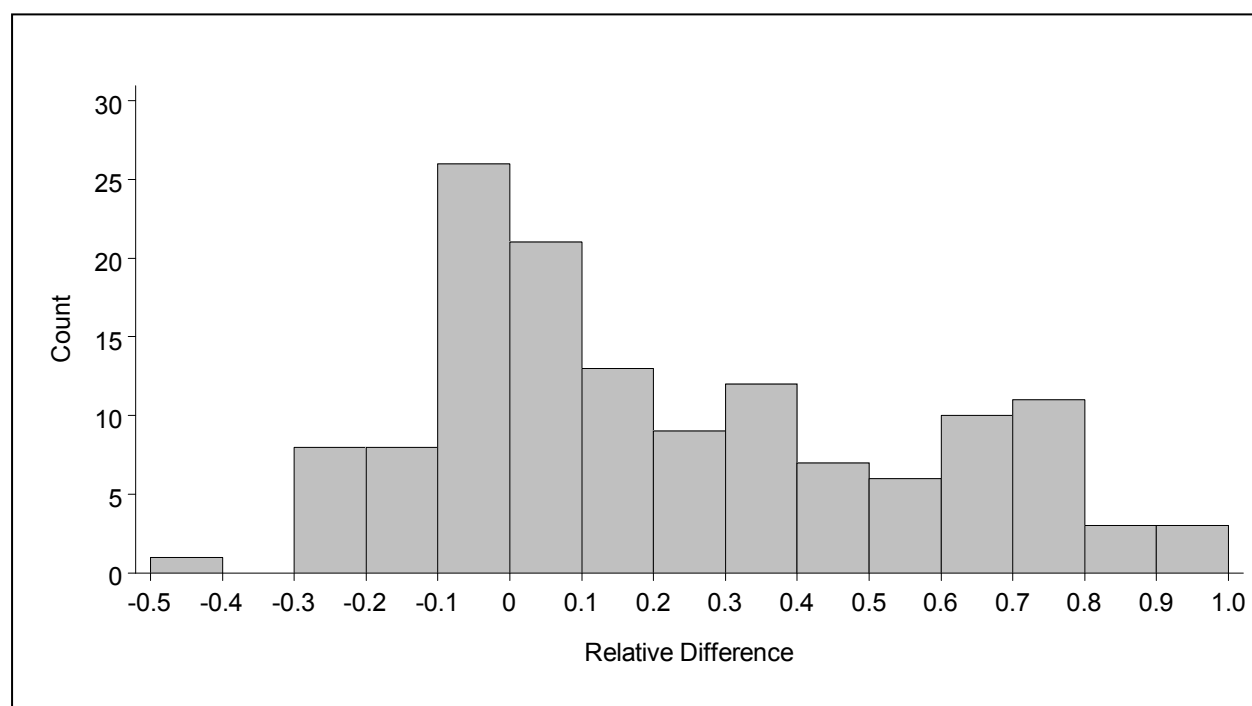
#### 4.2 Impacts on Univariate Distributions

The ELS:2002 surveys provide a variety of analysis weights; including cross-sectional weights for each round (base-year, first follow-up, and second follow-up) and a variety of longitudinal weights used to analyze data collected in multiple rounds. In order to assess how data suppression affected population estimates, a single analysis weight was selected. While any of the available analysis weights could have been used, the cross-sectional analysis weight, F2BYWT was used. An assessment could have been performed using other available analysis weights but, as the results below show, population estimates are affected significantly by creating 2-anonymity so using another weight

that showed more or less significant changes does not change the ultimate conclusion; that 2-anonymity, by itself, changes estimates significantly.

The suppression of values caused the distributions of the 19 variables to change. There were three variables that had one or more values completely suppressed. The response value “Non-native English speaker, fluency unknown” for the variable Parent’s English Fluency (BPLANG) was completely suppressed. The response value “Military” for the variable “Mother/female guardian's occupation-composite” (BYOCCUM) was completely suppressed. The response values “1983” and “0” (where “0” indicates missing) for the variable “Student’s Birth Year” were completely suppressed.

The changes in the distributions of the 19 variables, for the values that were not completely suppressed, were examined by calculating the weighted percentage of responses for each variable’s set of response values. Weighted percentages were calculated using the original ELS:2002 data and using the suppressed ELS:2002 data. The difference between weighted percentages, calculated as full data percentage minus suppressed data percentage, varied from -10.5 to 2.9 with a mean of 0 and median of 0.5. Figure 1 shows the distribution of the relative difference between weighted percentages (scaled by the full data weighted percentage).



**Figure 1. Relative Difference Between Pre- and Post-suppression Weighted Percentages**

In addition to comparing the weighted percentage point estimates, 95% confidence intervals of the weighted percentages were also examined. Confidence intervals overlapped for 70 of the 138 values (51%).

### 4.3 Impacts on Multivariate Relationships

In an attempt to assess the degree to which suppressions, alone, impact multivariate relationships, the relationship between three other ELS:2002 variables and the 19 quasi-identifiers was examined through the use of linear regression. Three variables, BYTXCSTD (Standardized test composite score-math/reading), BYTXMSTD (Math test standardized score), and BYTXRSTD (Reading test standardized score) were selected as dependent variables for a regression analysis.

Three separate models, using the same covariates, were created for each dependent variable. One model used the full ELS:2002 data, a second model used the suppressed data arising out of the application of 2-anonymity, and a third model used the suppressed data but replaced the missing values with a numeric value of 99; a poor imitation of an imputation.



Because of collinearity among the 19 quasi-identifiers and because of the desire to use the same set of covariates in each model, all models used only 8 of the 19 quasi-identifiers as covariates. These eight covariates are BYSEX, BYRACE, BYSTLNG2, BYPARACE, BYPLANG, BYPARED, BYFATHED, and BYSCTRL. Table 4 lists, for each model, the covariates statistically significant at the 0.05 level.

**Table 4. Model Significant Variables**

<b>Model Outcome</b>	<b>Data</b>	<b>Significant Covariates</b>
BYTXRSTD	Full Data	All but BYPLANG
BYTXRSTD	Excluding Suppressed Data	All but BYPLANG
BYTXRSTD	Suppressed Data with 99	All but BYPLANG
BYTXMSTD	Full Data	All
BYTXMSTD	Excluding Suppressed Data	All but BYSTLNG2
BYTXMSTD	Suppressed Data with 99	All
BYTXCSTD	Full Data	All but BYSEX
BYTXCSTD	Excluding Suppressed Data	All but BYSEX and BYPLANG
BYTXCSTD	Suppressed Data with 99	All but BYSEX

The models using the full ELS:2002 data and the models using the suppressed data with a value of 99 in place of a missing indicator yielded the same significant covariates. The models of BYTXMSTD and BYTXCSTD that exclude records with suppressed data yield one less statistically significant covariate than the other two types of models. The model of BYTXRSTD that excludes records with suppressed data yields the same statistically significant covariates as the other two types of models.

## 5 Conclusions

The primary objective of the research behind this paper was to determine if it was practical to apply the notion of 2-anonymity to a real world microdata set. Much of the current effort was spent developing a C++ software program that could process a data set and identify, in a reasonable amount of time, unique patterns of quasi-identifiers. The current version of the software is able to process the ELS:2002 data (11,204 records and 19 variables) in a practical amount of time: 9 hours. The subsequent SAS program used to determine the minimum number of suppressions takes about 5 hours. It seems reasonable that 2-anonymity, at least when the number of quasi-identifiers is 20 or less, could be integrated into a broader statistical disclosure treatment process. Improvements to the algorithm used to identify MINUCS and utilizing parallel computational capabilities, it seems likely that up to 30 quasi-identifiers could be reasonably handled on a single multi-core CPU system.

A secondary objective of the research was to determine how suppression of data via application of 2-anonymity impacted statistical relationships. The initial application of 2-anonymity to the full ELS:2002 data shows that some variables have a higher percentage of values suppressed than others. The weighted distributions of the quasi-identifiers are also impacted by the initial application of 2-anonymity with roughly half of the confidence intervals of the weighted percentages of each value not overlapping between the full data and suppressed data percentages. The application of 2-anonymity also affects the multivariate relationships of the quasi-identifiers with other variables. For 2 out of 3 outcome variables, the models that exclude records with suppressed data result in slightly different sets of statistically significant covariates.

The degree of perturbation in point estimates and the non-overlapping confidence intervals indicate that 2-anonymity should not be applied directly as a means to limit disclosure risk. Rather, it seems reasonable to use the notion of 2-anonymity to review the variables and suppressed values identified by the MINCUS. This review could lead to decisions to drop variables entirely or to suggest recoding of variables that could be applied. This leads to the concept of identifying MINUCS, dropping and recoding variables based on review of the MINUCS, re-identifying MINUCS using the reduced (and recoded) microdata, and repeating the process of dropping, recoding, and re-identifying MINUCS until the degree of data coarsening and variable suppression is no longer acceptable.

After variable suppression and data coarsening have occurred, optimal local suppression could be implemented but changes in estimates should be reviewed to determine if too much suppression has occurred. If too much

suppression has occurred, imputation could be used to replace missing values with imputed values but the possibility of recreating unique combinations in the original microdata remains. A related alternative to imputation is the creation of revised analysis weights that attempt to provide the ability to produce weighted estimates close to the weighted estimates produced using the original microdata. Weight adjustment procedures are complicated to develop and there is no guarantee that multivariate relationships would be preserved.

## References

- 1 Samarati P, Sweeney L (1998). Generalizing data to provide anonymity when disclosing information (Abstract). In Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, p. 188, Seattle, WA, USA.
- 2 Meyerson, A., Williams, R. (2004). On the Complexity of Optimal KAnonymity. Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (2004), pp. 223-228, Paris, France
- 3 L. Willenborg and T. De Waal. Statistical Disclosure Control in Practice. Springer-Verlag, 1996.
- 4 Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering, Volume 13, Issue 6.
- 5 L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.