Controlled shuffling, statistical confidentiality and microdata utility: a successful experiment with a 10% household sample of the 2011 population census of Ireland for the IPUMS-International database

Abstract. IPUMS-International disseminates more than two hundred-fifty integrated, confidentialized census microdata samples to thousands of researchers world-wide at no cost. The number of samples is increasing at the rate of several dozen per year, as quickly as the task of integrating metadata and microdata is completed. Protecting the statistical confidentiality and privacy of individuals represented in the microdata is a sine qua non of the IPUMS project. For the 2010 round of censuses, even greater protections are required, while researchers are demanding ever higher precision and utility. This paper describes a tripartite collaborative experiment using a ten percent household sample of the 2011 census of Ireland to estimate risk, mask the microdata using controlled shuffling, and assess analytical utility by comparing the masked data against the unprotected source microdata. Controlled shuffling exploits hierarchically ordered coding schemes to protect privacy and enhance utility. With controlled shuffling, the lesson seems to be the more detail means less risk and greater utility. Overall, despite substantial perturbation of the masked dataset (30% of adults on one or more characteristic), we find that data utility is very high and information loss is slight, even for fairly complex analytical problems.

Keywords: controlled shuffling, population census, microdata sample, data privacy, data utility, statistical disclosure controls, IPUMS-International, Ireland

1 Introduction

IPUMS-International disseminates integrated, confidentialized census microdata samples to researchers world-wide at no cost[1]. Currently, 259 samples (561 million person records) encompassing 82% of the world's population (79 countries) are available to more than 10,000 registered users, representing over one hundred nationalities. Each year the database expands with the addition of samples for the 2010 round of censuses and for more countries, as the tasks of integrating microdata and metadata are completed.

Protecting the confidentiality and privacy of individuals represented in the microdata is a sine qua non for the IPUMS project. Access is restricted by means of a rigorous vetting process. To be granted access, researchers must demonstrate their bona fides, agree to abide by the stringent conditions of the user license, and demonstrate a specific research need. The microdata are further protected by the fact that researchers do not obtain complete copies of samples, but instead must submit an individual ("extract") request, specifying not only the sample or samples but also the precise variables and even sub-populations required. In other words, each extract is unique, and none is complete. This process of dissemination provides additional safe-guards against researchers sharing data with unauthorized persons.

Technical measures, such as sampling of households, suppression of variables and codes, and swapping of records, are also used to protect the confidentiality of the microdata. For the 2010 round of censuses, even greater protections are required due to the explosion in available microdata, the development of ingenuous techniques of data mining and matching, and the threat of unethical behavior facilitated by the internet. Honesty, trust and professional responsibility continue to be held in highest esteem by all but the tiniest minority of researchers. Nonetheless, census microdata must be protected such that the slightest allegation of violation of confidentiality may be immediately and credibly debunked.

The threat of de-anonymization in the age of "Big Data" is real. Despite the fact that to gain access to the IPUMS-International database the conditions of use license endorsed by each user expressly prohibits any attempt to identify individuals in the census microdata, before release strong technical measures must be applied to protect the microdata against even the remote likelihood of re-identification. At the same time we must assure researchers that the microdata are of the highest precision and utility.

This paper describes a tripartite collaborative experiment to estimate risk (Comerford), protect the data using controlled shuffling (Muralidhar and Sarathy), and assess the analytical utility (McCaa and Esteve). Thanks to the cooperation of the Central Statistical Office of the Republic of Ireland, a 10% household sample of the 2011 census was used as a test case. The sample is richly detailed with 474,535 person records, 117,945 families, and 79,785 couples described by 43 variables and more than 1,400 unique attributes. Person records include variables for single year of age (0-85+), occupation (number of categories=90), industry (110), country of birth (92), nationality (75), relationship to reference person (12), educational level (7), etc. Before beginning the experiment, we recoded "County of usual residence" (35) into region (8), thereby sacrificing geographical detail to facilitate analysis of social, demographic, cultural and economic attributes.

2 k-anonymity

A standard approach to the assessment of disclosure risk addresses three key aspects in the literature: the data environment, the sensitivity of the data and the data characteristics. Examples of this type of approach can be seen in [2], [3]. In our analysis we interpreted these three aspects in the following ways. The data environment is an attempt to capture information about the world outside of the data under consideration for release. This information is used to demonstrate the a priori knowledge of a would-be intruder and can be configured in a number of ways to simulate different intruder scenarios. In our experiments we wanted to provide a robust analysis and therefore chose a deliberately conservative re-identification key. This was based on the growing concerns about the amount of information publicly available online through social networking sites, e.g., Facebook and LinkedIn. Searching public profiles on LinkedIn using one of our author's names revealed a number of individuals that share a very detailed personal curriculum vitae, without the need for a 'friend request' style level of security.

Extrapolating the information from social media we constructed our conservative key with the following variables from the census sample: sex, age, marital status, nationality, ethnicity, level of education, occupational group, industry classification, region of usual residence, region of birth, country of usual residence and country of birth. This assumes a high level of knowledge for an intruder and should be seen as a worst case scenario.

In this context, 'data sensitivity' means the extent to which the data's subjects might consider the information held in the dataset to represent a threat to their privacy. This is often considered aside from the legal obligations of the data holders. For example, projects like the Scottish Health Informatics Programme (SHIP) use this aspect of risk assessment to build trust with the data subjects, holding focus groups with patient representatives. For our experiments the data sensitivity contributed to the selection of our test parameters as set out below, taking into account also that we are working with a sample of the population.

The data characteristics take the information gathered from the environment and the data sensitivity and seek to describe the data in an empirical analysis. For this purpose we used k-anonymity, a well-established tool for highlighting re-identification risk. K-anonymity is satisfied if a record is indistinguishable from k-1 other records for a given key. Despite certain criticisms and enhancements k-anonymity still offers a reliable test providing the results are interpreted within the test's definition. For a discussion of k-anonymity see [4]. Given our assessment of the data sensitivity in this case, we set the k-anonymity threshold at 3, and the key as referenced above.[1]

[1] K-anonymity tests were carried out using the NIAH algorithm available from: https://sourceforge.net/projects/niahsdc/

The first pass of the data, using a k-anonymity threshold of three, flagged 78% of records as not meeting the k-anonymity criteria. This high level was to be expected given such a strong key. This allowed us to look at those records that did meet the criteria and unpick their apparent homogeneity. The results showed that at this level young people made up the bulk of our records meeting the k-anonymity criteria because they share a number of values in our key i.e. they tend not to have been married, they do not work, and they are still in school.

For the second pass of the data we experimented by removing variables from the key to see what effect this would have on the k-anonymity rate. After each k-anonymity test we analyzed the remaining risky records to inform the order in which variables could be removed from the key. Once an order was chosen those records that flipped from 'not satisfying' to 'satisfying' k-anonymity were flagged with a dummy variable indicating which variable had affected the change.

We concluded that the variables age, education, occupational group and industry classification followed by the geographical variables should be considered for our later data shuffling experiments.

3 Controlled Data Shuffling to Prevent Disclosure and Preserve Utility

The purpose of disclosure risk assessment is to identify the extent to which the unmodified release of the data could result in potential re-identification of the records and, possibly, the subsequent disclosure of sensitive information regarding individuals. If the risk of such disclosure is deemed low, then it may be appropriate to allow users to analyze the original data resulting in the highest level of analytical utility. When the risk of disclosure is high, then it may be necessary to modify the data prior to dissemination so as to prevent re-identification and disclosure of confidential information. The process of modifying the data prior to allowing access is often referred to as data masking.

There are a wide variety of data masking solutions that are available. At the broadest level, they can be classified as input or output masking. In input masking, the original data is masked and all analyses are performed on the masked data. In output masking, the analyses are performed on the original data and the results of the analyses are masked prior to release. For static data, which includes all the samples integrated into the IPUMS-International database, input masking is generally preferred since it provides the assurance that the results of the same analysis on the same data performed at any point in time will *always* yield the same results. Maintaining consistency at this basic level is crucial to maintain users trust in the validity of the data. For output masking, unfortunately, it is extremely difficult (if not practically impossible) to ensure consistent results. Hence, in the remainder of this paper, we limit our discussion to input masking.

There are many input masking techniques that are available. Hundepol et al. provide an excellent discussion of these techniques [5]. Given that we have used k-anonymity to identify risky records, it seems reasonable that input masking through aggregation, simple aggregation for categorical data [6] and micro-aggregation for numerical data [7] would be relevant. Unfortunately, given that close to 80% of the records were identified as being at risk, the level of aggregation that is required in order to prevent disclosure is so high the types of analyses that can be performed on the aggregated data would be severely limited. In order to provide users with greater flexibility in analyzing the data, we chose to investigate alternative procedures.

Input masking through data perturbation is one approach that can be used in these situations. There are many data perturbation techniques that are available (see [5]). Most of these techniques rely on modifying the original data through random noise, and the values in the masked data are different from those in the original data. This would be perfectly acceptable for traditional numerical data. The treatment of nominal data is a more difficult problem for data perturbation approaches, and only a few select techniques are capable of perturbing nominal data (see Hundepool et al [5] for a comprehensive discussion).

Recently Domingo-Ferrer et al [8] identified the specific problem of taxonomic data, that is, data whose values are nominal but also have a hierarchical structure such as medical diagnosis coded using the International Classification of Diseases [9]. In the Irish data, there are two variables that fall under the category of taxonomic data (Industry classification with 110 hierarchical categories and Occupation group with 90 hierarchical categories). For example, the 90 3-digit occupation groups are divided into 9 1-digit groups. Group 1, "Managers, Directors and Senior Officials", contains 12 3-digit occupations, while Group 9, "Elementary Occupations", has only 9. By controlling the shuffling to take into account the hierarchical codes, the perturbed data are more likely to preserve associations with other variables, such as education, industry, and even age.

One approach to handling taxonomic data is to convert them to purely nominal data (by representing every unique code within the taxonomy as a nominal variable). The problem with this approach is that it results in a very large number of nominal variables making it extremely difficult to carry out the perturbation. More importantly, this transformation ignores the inherent taxonomy that is an integral part of the variable. Hence, in the presence of taxonomic data, perturbation approaches that "generate new values" for the original values are not appropriate.

Among data perturbation techniques, there are two that differ from all others in the fact they do not replace the original values with newly generated values, but reassign the original values between records. These two techniques are data swapping [10] and data shuffling [11]. In data swapping, the values of a variable are exchanged between two records within a specified proximity. The process will then have to be repeated for every variable that is to be masked. The problem with this approach is that the swapping is performed on a univariate basis and it is difficult to maintain consistent levels of swapping across many variables. Swapping also results in attenuation of the relationship both between the swapped variables and between the swapped and un-swapped variables.

Data shuffling, by contrast, is a multivariate procedure where the values of the individual records are reassigned to other records in the data set based on the rank order correlation of the entire data set. One of the key features of data shuffling is that the rank order correlation of the masked data is asymptotically the same as that of the original data. This ensures that all monotonic relationships between the variables are preserved by the shuffling process. When compared to data swapping, data shuffling provides a higher level of utility and lower level of disclosure risk [12]. Data shuffling is capable of handling all types of data. Numerical and ordinal data inherently lend themselves to data shuffling. Nominal data are converted to binary data prior to shuffling. And for taxonomic data, numerical mapping proposed by Domingo-Ferrer et al [8] is used.

Data shuffling can be briefly described as follows. Let X represent the set of confidential variables and let S represent the set of non-confidential variables. Let Y represent the masked confidential variables. Data shuffling models the joint distribution of {X, S, Y} as a multivariate normal (Gaussian) copula. Let {X*, S*} represent the normalized values of the {X, S}. The perturbed normalized values Y* are created using the conditional distribution {X*, S*}. Once the values of Y* have been generated in this manner, the original values of X are reverse mapped to Y* to result in the perturbed values Y. For a complete description of data shuffling please refer to Muralidhar and Sarathy (2006).

Data shuffling offers the following advantages:

1. The shuffled values Y have the same marginal distribution as the original values X. Hence, the results of all univariate analyses using Y provide exactly the same results as that using X.

2. The rank order correlation matrix of {Y, S} is asymptotically the same as the rank order correlation matrix of {X, S}. Hence, the results of most multivariate analysis using {Y, S} should asymptotically provide the same results as using {X, S}.

One of the key features of data shuffling is that the process is based on joint rank order correlation matrix of *all variables* {X, S, Y}. This provides the data administrator with the ability to control for disclosure risk by specifying the appropriate relationship between the original (X) and masked (Y) variables. This specification can range anywhere from no protection (no shuffling), to maximum protection (where X and Y are conditionally independent given S), and any level in between. Prior illustrations of data shuffling have used the maximum level of protection. We use the term *controlled data shuffling* to indicate that the desired level of disclosure protection has been specified by the data administrator. This new approach provides a much higher level of flexibility in implementing data shuffling. We now provide the results of implementing data shuffling for the Irish data.

4. Assessing analytical utility

The primary purpose of IPUMS-International is to provide researchers access to harmonized census microdata for countries around the globe. Hence, a successful data protection mechanism must ensure not only that the masked microdata are sufficiently confidentialized but also provide results that are similar to those using the original, unharmonized microdata held by the National Statistical Office-owners.

In this section, we assess analytical utility of the 10% household sample for the 2011 census of Ireland entrusted to the IPUMS-International project. One important aspect of this evaluation is that the microdata were masked without knowledge of the subsequent analyses that would be done on the dataset. Hence, this evaluation provides a more genuine assessment of the effectiveness of the masking procedure. As is the universal rule for official census microdata, we agreed to not report details regarding which variables were perturbed or the degree of perturbation. To do so would increase confidentiality risks for the microdata.

We consider confidentiality protection, taken as a whole, to be strong. One or more characteristics were perturbed for 29.9% of adults aged 20 years or more. For couples (excluding same-sex unions, which are too few in number to successfully shuffle as a conditional characteristic), joint attributes were taken into account to maintain husband-wife associations. For individuals, joint characteristics were controlled so as not to attenuate associations between occupation, industry, social class, educational attainment and socio-economic group. Six cycles of experiments were required to produce a "Goldilocks" dataset—one that was neither over- nor under-confidentialized and with utility at the highest possible levels—and therefore acceptable to all parties.

Overall, the results show excellent analytical utility. Consider age, for example. We compared mean age for 34,517 effective subgroups from over ten million permutations of six key variables: sex (2), level of educational attainment (7), industry (110), occupation (90), social class (8), and socio-economic group (11). As expected, differences are inversely proportional to the size of the cell counts. For combinations with counts of equal to or less than ten, the mean difference in age ranges between +/- 0.6 years. With cell counts of 30 or higher, the range shrinks to +/- 0.4

Log-linear models of cells counts.

We use log-linear models to test whether complex analytical models—original and shuffled—produce the same best fitting models. To illustrate the method consider a four-way cross-classification of Age (20-85+, 66 categories), Sex (2), Marital status (4),and Region of usual residence (8). First we model the original source microdata using seven models. Second, we compute the same models using the shuffled data. Finally we compare the differences in goodness of fit between the two datasets. If the goodness of fit statistics for the original and shuffled data differ substantially, the masking procedure has distorted the results by introducing bias. Our model specification allows for unrestricted associations between all variables.

Our baseline or independent model can be written as follows:

$$\ln(F_{ijklm}) = \mu + \mu_i^A + \mu_j^S + \mu_k^M + \mu_l^R , \qquad [1]$$

where $\ln(F_{ijkl})$ is the log of the expected cell frequency of the cases for cell *ijkl* in the contingency table; *i, j, k, and l* refer to the categories within the variables Age (A), Sex (S), Marital status (M), and Region (R). μ is the overall mean of the natural log of the expected frequencies; μ_i^A is the effect of age *i* has on the cell frequencies (the same for μ_j^S , μ_k^M , and μ_l^R).

Table 1 describes each of the models and goodness-of-fit statistics. Model 1 corresponds to the baseline or independent model described above. The modeling strategy consists of adding two level interactions between variables and testing for improvement in the fit of the model. To assess fit, we use the Likelihood Ratio Chi-squared statistic (L^2) and the Bayesian Information Criterion (BIC), which is based on the L^2 statistic [13]. BIC introduces a penalty term for the number of parameters in a model. Thus, it is possible to improve the fit of a model by adding more parameters, but if this adds unnecessary complexity in terms of a reduction in degrees of freedom, BIC will indicate a poorer fit.

Models 2 to 7 include two and three level interactions between age, sex, marital status and region. Comparing goodness of fit statistics of the shuffled data with those from the original source data for each model reveals no significant differences in either L2 or BIC. Model 2 includes all two way interactions between age and sex, marital status and region. Model 2 indicates a substantial improvement over the baseline model in goodness of fit both in terms of L2 and BIC. Model 5 offers the most parsimonious fit for both datasets according to BIC (BIC_5 = -37894.8 and -37860.1, respectively). Model five includes all possible two way interactions between age, sex, marital status and region. Three way interactions yield a tighter fit, but the loss in degrees of freedom is proportionally greater than the gains in goodness of fit, so BIC tells us that the additional model complexity is unwarranted.

Table 1. Log-linear models of original and shuffled data show small differences in goodness of fit.

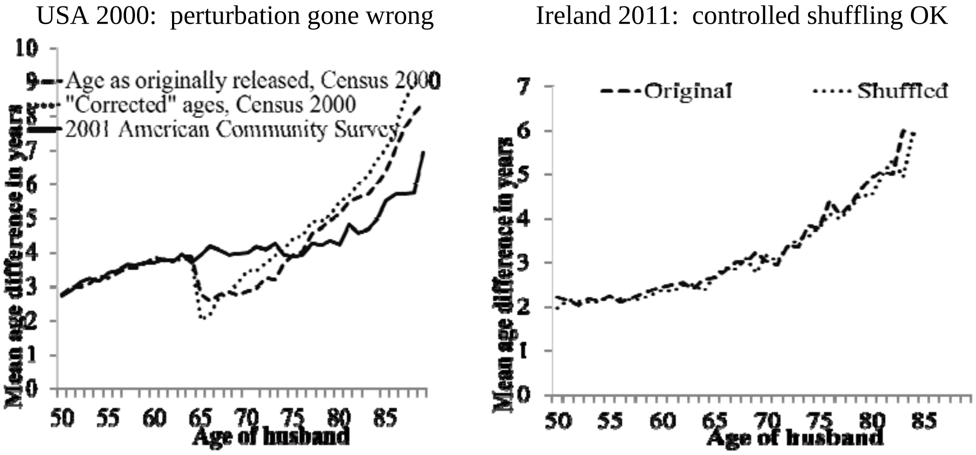
			Goodness of fit				Percentage Difference*	
			Original Data		Shuffled Data			
Model		df	L^2	BIC	L^2	BIC	L^2	BIC
1	A, S, M, R	4147	203400.7	150522.8	203226.1	150348.2	0.1	0.1
2	AS, AR, AM	3432	13013.3	-30747.7	13052.5	-30708.5	-0.3	0.1
3	AS, AR, AM, SM	3429	6536.4	-37186.4	6573.5	-37149.3	-0.6	0.1
4	AS, AR, AM, SM, SR	3422	6471.8	-37161.7	6508.8	-37124.7	-0.6	0.1
5	**AS, AR, AM, SM, SR, RM**	**3401**	**5471.0**	**-37894.8**	**5505.7**	**-37860.1**	**-0.6**	**0.1**
6	ASM, ASR	2772	4748.0	-30597.4	4812.6	-30532.9	-1.4	0.2
7	ASM, ASR, SRM	2730	3426.3	-31383.6	3487.8	-31322.1	-1.8	0.2

Note: A (66) Age 20-85+, S (2) Sex, M (4) Marital Status, R (8) Region.
*Percentage difference = ((Original-Shuffled)/Original)*100
Source: Author's calculations from 10% household sample of the 2011 population census of Ireland

What is striking from Table 1 is that both the original and shuffled datasets lead to the same best fitting model and the differences in goodness of fit between original and shuffled are trivial, less than 0.3% for BIC. All in all, the results clearly suggest that, with regard to the variables analyzed, there are no statistically significant differences between the shuffled and the original dataset. We conclude that distortions introduced by shuffling have not significantly diminished analytical utility.

Age gap between spouses.

For a second test, consider the gap in ages between spouses, a challenging correlation to maintain with masked microdata. A notorious example of perturbation gone wrong is the sample of the 2000 census of the USA, which contains an embarrassing error due to masking of ages for persons 65 years and older. Later, the Census Bureau "corrected" the error, but seemingly worsened the discrepancy (see left panel, Figure 1).



McCaa et al [1], p. 185. Calculations by authors.

Fig 1. Masking effects on age gap between spouses: two examples

In contrast, for the 10% household sample of Ireland (right panel, Figure 1), comparing the unperturbed and shuffled microdata reveals surprisingly minor discrepancies throughout the age range, despite the fact that in 4% of the cases age was masked for both members of the pair and in 20% at least one. The age gap between spouses is a strong test of data utility, a test that the shuffled Irish sample readily passes.

Own-Child fertility.

As a final test, we focus on fertility. Fertility is fundamental for demographic research, and population censuses offer valuable insights on fertility levels, trends, and differentials. Where the census does not ask questions on fertility, estimates can still be derived indirectly from household samples, using the "own-child method". Children aged 0-14 are matched to their mothers by means of the relationship to reference person variable. Then, a 15 year fertility series is constructed from the ages of mothers and their co-resident children. (The data are adjusted both for children who cannot be matched to mothers and for mortality). A challenging test for masked data is to replicate the age differences between mothers and their children.

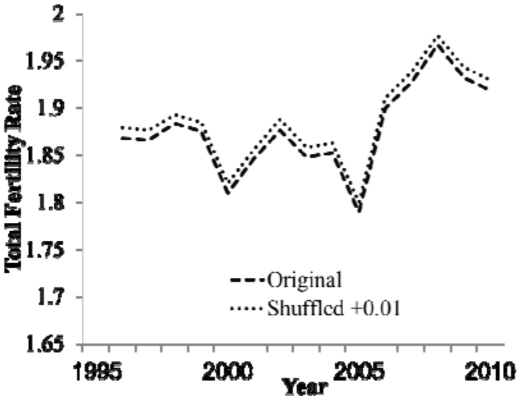


Fig 2. A 15 year series of Total Fertility Rates from a household sample of the 2011 census of Ireland: Shuffled microdata track the unperturbed source data so closely that to highlight separation its line is shifted +0.01. Otherwise differences are imperceptible

Figure 2 shows that the shuffling strategy yields astonishingly robust results in spite of the fact that the data were masked without any knowledge that it would be tested in this way. Differences in total fertility rates between the original data and the shuffled are at 3 decimal places. They are so imperceptible that for illustration purposes 0.01 was added to the shuffled series in Figure 2 to make the point that there are indeed two sets of data portrayed. Drilling down, we find that both datasets reveal declining fertility for ages 15-29 and rising fertility for ages 30-49. While this is not news to experts on Irish fertility, what is surprising is that the pattern is unmistakable even in the shuffled data.

5. Conclusion.

Data shuffling is widely recognized as a robust masking procedure for confidentializing microdata. Controlled shuffling allows the data administrator greater flexibility to protect privacy and enhance utility. The success of this experiment was possible thanks to the close cooperation by the microdata owner/steward, statisticians, data administrators, and researchers. Initially, for the 2011 census sample of Ireland, a reduced set of variables, including age in five year bands, was offered to IPUMS, severely diminishing the utility of the sample. Following further discussion, CSO-Ireland agreed to furnish single year ages and more than a dozen additional variables provided the microdata could be confidentialized satisfactorily.

We were surprised at the difficulty in striking a balance between protection and utility for household samples with a rich array of variables. Six trials were required. The first was deemed acceptable by the CSO and the statistical experts, but not by the data administrators or researchers. With each successive round of experimentation we learned more about how to control the perturbations to retain utility, specifically associations within person records and between person records within households, yet protect the data. One important lesson is that subject matter specialists must be included in the confidentializing process for quality assurance to test the analytical utility of perturbed data. Failure to do so may lead to unmeasured bias, particularly in the associations between couples, parents and children, and even between variables within a single person record. Statistical properties of variables are not synonymous with analytical properties of individual records nor associations within households.

Counter-intuitively, the higher the sample density and the more detailed the variables, the finer grained the shuffle, the better the protection and the greater the utility. For example, shuffling across 1, 2, or 3 single years of age perturbs the data less than shuffling across 5, 10 or 15 years of quinquennial age bands. Likewise, for occupation, shuffling at the third digit level within the second digit distorts the data much less than shuffling across first digit boundaries.

Controlled shuffling offers the ability to model hierarchically ordered coding schemes common to census microdata. The promising results of this experiment may be of interest not only for masking census microdata but for all types of microdata with explicit hierarchical codes, whether based on international standards such as ISCO, ISIC, NACE, NUTS, etc. or ex post facto integrated codes such as those developed by IPUMS. The robustness of the shuffle is enhanced by taking into account associated characteristics—within households and within clusters of variables, such as correlations between occupation, industry, educational attainment, social class, etc.

With the success of this experiment, the CSO entrusted single years of age for the 2002 and 2006 samples which, before release to researchers, were confidentialized using controlled shuffling. Prior to publication of this paper, all three samples were successfully

integrated into the IPUMS-International database and released to the research community for dissemination on a restricted access basis. The confidentialized sample of these and eight other censuses of Ireland (1971-2011) along with over 250 other samples for more than 75 countries may be downloaded at:
https://international.ipums.org/international/sample_designs/sample_designs_ie.shtml.

Researchers must heed the warning that microdata subjected to any masking procedure, including controlled shuffling, introduces bias. Moreover, the smaller the frequency of a combination of characteristics, the greater the proportion of cases perturbed. As indicated by the percentage differences in Table 1, the more complex the statistical analysis, the greater the distortion caused. Nonetheless of the hundreds of models tested (of which only 7 are reported here), the analytical differences are trivial. Compared with the enormous loss occasioned by aggregating age to five year groups and by the suppression of more than a dozen variables due to confidentiality concerns by the data producer, controlled shuffling offers an elegant solution to the conundrum of protecting statistical confidentiality, yet retaining the highest utility in the source microdata.

To assuage concerns regarding analytical validity and disclosure risk, the corresponding author extends an invitation to researchers to conduct analysis on the microdata that can be compared against the original (unmasked) data.

Acknowledgements. The authors greatly appreciate the cooperation of the Central Statistics Office of the Republic of Ireland in providing for this experiment a ten per cent household sample of the 2011 census. The authors alone are solely responsible for the contents of this paper. Funded in part by the National Science Foundation of the United States, Grant Nos. SES-0433654 and 0851414; National Institutes of Health, Grant Nos. R01HD047283 and R01HD044154.

References

1. McCaa, R., Cleveland, L., Ruggles, S. and Sobek, M. When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata. J. Domingo-Ferrer and I. Tinnirello, eds., Privacy in Statistical Data PSD2012 Proceedings, LNCS 7556 (2012), 179-187.
2. Elliot, M. et al. Data Environment Analysis and the Key Variable Mapping System. Privacy in Statistical Databases PSD2004 Proceedings. Berlin: Springer-Verlag (2004),138–147. Available at: http://www.springerlink.com/index/6KL805434G016U15.pdf [July 13, 2012].
3. Elliot, M. & Dale, A. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Statistics (1999), 14:6–10.
4. Domingo-Ferrer, J. & Torra, V. A critique of k-anonymity and some of its enhancements. In Availability, Reliability and Security. ARES 08. Third International Conference. (2008), 990–993. http://ieeexplore.ieee.org/xpls/abs_ all.jsp?arnumber=4529451 [Accessed July 14, 2012].

5. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. and de Wolf, P.-P. Statistical Disclosure Control, Wiley Series in Survey Methodology. London: John Wiley & Sons (2012).
6. Sweeney, L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems (2001), 10:557–57.
7. Domingo-Ferrer, J. and Mateo-Sanz, J.M. “Practical data-oriented microaggregation for statistical disclosure control,” IEEE Transactions on Knowledge and Data Engineering (2002), 14(1):189-201, 2002.
8. Domingo-Ferrer, J., K. Muralidhar, K. and Ruffian-Torrell, G. Anonymization Methods for Taxonomic Microdata. J. Domingo-Ferrer and I. Tinnirello, eds., Privacy in Statistical Data PSD2012 Proceedings, LNCS 7556, Berlin: Springer-Verlag (2012), 90-102.
9. World Health Organization. International Classification of Diseases. Geneva (2008), 9th Revision, Clinical Modification, Sixth Edition. http://icd9cm.chrisendres.com/
10. Dalenius, T. and Reiss, S. P. Data-swapping: A Technique for Disclosure Control. *Journal of* Statistical Planning and Inference (1982), 6 73-85.
11. Muralidhar, K. and Sarathy, R. Data Shuffling- A New Masking Approach for Numerical Data. Management Science (2006), 52(5), 658-670.
12. Muralidhar, K., Sarathy, R., and Dandekar, R. (2006). Why Swap when you can Shuffle? A Comparison of the Proximity Swap and the Data Shuffle for Numeric Data. Domingo-Ferrer and Franconi, eds., Privacy in Statistical Databases PSD2006 Proceedings, LNCS 4302, Berlin: Springer Verlag (2006), 164-176.
13. Raftery, A.E. (1986). Choosing models for cross-classifications. American Sociological Review, 51(1), 145-146.