Westat

Improving lives through research™

The Unlinkable Data Challenge: Advancing Methods in Differential Privacy

Proposers: WesTeam

Sigurd Hermansen
Natalie Shlomo
Tom Krenzke
Jane Li
Marcelo Simas

July 20, 2018

Prepared for:
Unlinkable Data Challenge

Prepared by:
WesTeam
1600 Research Boulevard WB 310
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Chapter **Page**

Figures

1. Introduction

During the late 1990s, when many still believed that data stripped of IDs were unlinkable and therefore confidential, Federal agencies responded to privacy concerns among individuals and business enterprises about public release data. The Internal Revenue Service (IRS), for example, conducted a study of a possible linkage attack on a pending release of the Individual Tax Model Public Use File. A team of data linkage specialists and statisticians determined that "… there is no easy solution to minimize information loss and maximize disclosure protection …" (Winglee *et. al.* (2002)). Based on what we know today, one would have to excise the word "easy". The question has become "What compromises in accuracy of information or privacy protection, or both, do we accept?"

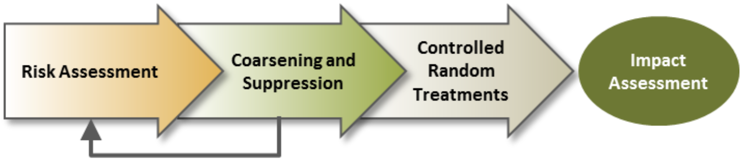
Growing demands for data and the escalating amount of available public information from various data sources, including from social media, administrative and large data warehouses (e.g., data.gov), combine with increasing capabilities and skills in record linkage and data science to prompt serious concerns about data privacy. The focus on re-identification of individuals or businesses within a file, once the basis for developing approaches for protecting privacy interests, has broadened, as stated in PCC (2017), to include linkage attacks based on any and all available public or commercial data sources. A data release can have an impact on a person's privacy rights whether or not that person's data appears in that release. Given the various degrees of "de-identified" datasets, record linkages compile data from various sources, add even more information to individual records, and greatly increase exposure risks. This so-called mosaic effect (OMB 2013) can compromise data privacy by accident or by providing more data for a malicious attack. Improved public access to masses of data and more powerful data transfer and computing technologies are having a global impact (e.g., AmStat News (2018)) particularly at national statistical institutes where there is a long tradition of releasing statistical data in the form of microdata (social surveys) and tables (survey responses, census, and business data). In addition, health researchers also have traditionally released data from clinical trials. Cases and controls have to consent to contribute data to, say, a trial of a new treatment. In medical research in particular, data privacy boards require informed consent of patients prior to a release of their data for research and shifts decisions to release data from owners of administrative data to individuals. This move toward privatization of research data shifts decisions to release research data from data owners, who follow public choice data utility-privacy risk guidelines for making decisions in the public interest, to individuals whose concerns about their own privacy tend to override the value they place on scientific progress of benefit to many. For example, while McLaughlin (2010) presents a legal case for collection and de-identified release of state cancer registry data in the USA, state legislatures and registry administrators have reacted to threats of data exposures by curtailing releases of data to researchers who request them. For similar reasons the Social Security Administration (SSA) has largely blocked access of researchers to the full SSA Death Master File (DMF) and made searches for decedents among members of research cohorts a longer and more expensive process. Strong guarantees of individual privacy would help researchers make a stronger case for releasing high quality data.

1.1 Traditional Statistical Disclosure Control Approaches

Statistical disclosure control (SDC) treatments of data suppression, coarsening and perturbation were generally developed to address traditional types of disclosure risks, such as identity and attribute disclosure. In times when data custodians keep strict controls on the release of data, there was little concern for inferential disclosure where an intruder can infer knowledge about a data subject with high probability through manipulating data releases.

Figure 1 shows a typical SDC process, starting with a risk assessment. The risk analysis will, at the very least, identify combinations of variables that may help to identify a respondent's data when considered together. The disclosure risk assessment focused on identity and attribute disclosure and included probabilistic record linkage matching procedures (e.g., Jaro 1989) that could determine whether a data record is a likely match with an external file record, and re-identification risk estimation based on probabilistic modelling (Reiter 2005b, Skinner and Shlomo 2008-). The purpose of conducting disclosure risk analysis is to inform the SDC process for each data file, for instance, the coarsening and suppression, and controlled random treatments (e.g., perturbation). In the end, an impact assessment should incorporate a risk-utility mapping, where results from various iterations of SDC methods and their parameters lead to determining the optimal SDC approach having the best trade-off for the data product in terms of holding disclosure risk below a tolerable threshold at the onset and maximizing data utility given that constraint. The SDC process typically involves a reasonable approach that can ensure protection of confidentiality while minimizing the impact of SDC on the integrity of the data.

Figure 1. General SDC process



It is now much more difficult for data producers to disseminate detailed data for individuals in ways that are convenient to researchers. This objective of balancing risk reduction with retention of data utility remains, while there is pressure to develop better approaches for releasing and protecting data.

1.2 New Forms of Data Dissemination and Differential Privacy

There is growing demand for more open and easily accessible data, especially as disseminated through national statistical institutes. Some web-based platforms have evolved or are under development to supply users with opportunities for research through flexible table builders, remote analysis servers, and use of synthetic microdata without the need for human intervention to check for disclosive outputs. This means that the paradigm of disclosure risks is shifting away from the traditional identity and attribute disclosures and is moving towards concerns about inferential disclosure where perturbative approaches are needed to protect the confidentiality of data subjects. Researchers are recognizing the need for perturbation with larger degrees of information loss in return for more flexible and accessible data.

Given concerns about inferential disclosure, this has led to the statistical community actively reviewing and researching the formal privacy framework developed in computer science and known as differential privacy (DP), in which data are protected by additive noise and/or randomization (see: Dwork, et al., 2006b, Dwork and Roth, 2014 and references therein). Why DP? DP subsumes all disclosure risks through the worst case scenario while employing privacy-by-design so disclosure risk is quantified *a priori*. Furthermore, while DP bounds the amount of risk protection, it also controls the amount of infused noise. The theory ensures that when dropping any one record of the file, very little can be learned about that record in estimates derived from data that is protected by a DP mechanism.

The concise and strong theoretical foundation of DP leaves many questions to address during the course of implementing a DP mechanism. The basic form of a privacy guarantee presented by Dwork (2006a) is the ε- Differential Privacy defined for a perturbation mechanism *M* as follows:

$$P(M(a) \in S) \leq \ e^{\varepsilon} P(M(a') \in S)$$

for all subsets S of the range of *M* and neighboring databases a and a' differing by one individual. A relaxed DP mechanism is the (ε, δ)-Differential Privacy which adds a new parameter δ as follows:

$$P(M(a) \in S) \leq \ e^{\varepsilon} P(M(a') \in S) + \delta$$

Here we use the definition of (ε, δ)-Differential Privacy instead of ε –Differential Privacy since the parameter δ can be thought of as a utility measure. It is the degree that we allow a 'slippage' in the bounded ratio of ε –Differential Privacy and, assuming that δ is very small, it is possible to substantially improve the utility of the released data. For example, δ represents the probability of not perturbing beyond a certain cap in a table of counts thus narrowing the range of possible perturbations.

Assuming $\delta = 0$, we define a loss function l with an argument of ε representing the privacy loss budget and an argument α controlling loss of accuracy in the protected output relative to the original output. l(ε,α) expresses the combined loss corresponding to a choice of two parameter values. Requiring e = 0.1, for instance, enables a DP mechanism M to add noise to a given output and generate a protected output with a maximum possible accuracy level of l = 1 - a on a convex PAF curve. Requiring a = 0.95 (l = 0.05), in contrast, limits e to a much higher value on the same PAF curve.

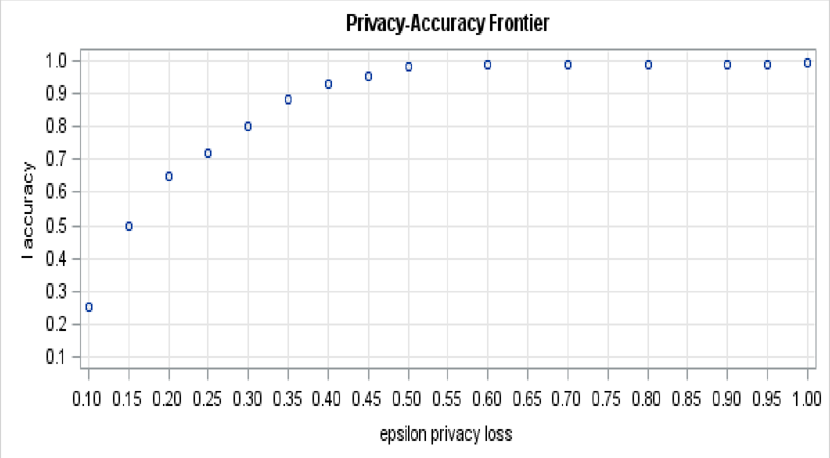
The possible pairs of values of ε and l lie on or below a privacy-accuracy frontier (PAF) similar to the one displayed in Abowd (2017). The frontier shows the optimal pairs of parameter values ε and l for an original output and mechanism *M*.

The PAF with an overlaid social welfare function (SWF), in the case of a data provider who in effect owns data, or a data supply curve in the case of entities who may choose to provide or refuse to provide data, serves as an aid to assigning values to the ε and α parameters. At the least, a PAF with a SWF or supply curve would show a sketch of the options for mechanism M given the original output.

The current state of the mechanism M in the context of a public release data set would likely generate a PAF that has privacy loss values ε acceptable to data providers paired with α values unacceptable to data users, and the other way around. Even so, the exercise of creating a somewhat formal framework for assessing trade-offs between tighter privacy controls and better data accuracy will improve dialogs between information technologists and statisticians, system developers and researches, data collectors and providers, and statistical institutes and the data users they serve. For no other reason, it would be worth the effort to avoid producing differentially private output that compromise data privacy and has no value to researchers or to the public.

The range of ε on a horizontal axis of a graph represents the acceptable range of privacy loss. The range of l on the vertical axis has a scale of $\{0 \ldots 1\}$. A simplified, hypothetical example is presented in Figure 2.

Figure 2. Privacy-Accuracy Frontier (PAF) for hypothetical DP Mechanism *M* and an original output



With some exceptions, data providers granting informed consent to privacy loss focus on the risk to them as individuals and place much less weight on accuracy of data and statistics. All the more reason, it follows, to find ways to shift the PAF upward and make data release decisions easier for data providers.

One path toward improving the terms of the privacy-accuracy trade-off would be to define accuracy of output not merely by comparison to private data and their distributions, but also by the accuracy of estimates derived from them. Carefully filtering of noise improves the accuracy of small area statistics, time series cycles, statistical/machine learning of image and sound classes, and many other specialties, yet filtering may also remove incidental variations that distinguish individual observations from others in a sample. Any one sample from a population may include extreme observations that would identify an entity contributing data to a public release data set. Resampling from a large sample generates sub-samples in some instances that contain fewer of the identifying extreme observations. Releasing a DP sub-sample along with the distribution of the original sample would improve data utility and still guarantee a prescribed risk of exposure. An array of methods, including Bayes factors and posterior distributions, raking and regularization, multiple imputation, and time series smoothing, as examples, may add further resistance to linkage attacks while improving the power of a sample to detect truly significant differences among classes of entities or discover clusters of similar entities. We intend to exploit noise filtering to improve data accuracy and reduce risk of data breach.

Ligett (2017) advocates turning the impact on accuracy of a privacy guarantee on its head by searching in an automated but principled manner for mechanisms that achieve a maximum privacy guarantee given an acceptable level of accuracy. Empirical risk management (ERM) methods offer a complementary strategy for controlling risks of privacy loss.

Given a choice, researchers and the public usually prefer to have access to microdata and tables rather than summary statistics or model parameter estimates. Data users tend to suspect information loss in synthetic tables. Further, many statistical institutes and government agencies have supplied organizations and interest groups with microdata and tables or have contracted for public use microdata and tables. Generating superset microdata or tables by multiple imputation may perhaps prove more representative of a population than a single sample drawn from that population, and this feature may help allay some of the concerns of researchers and the public about synthetic and highly perturbed data.

1.3 Examples of Relevant WesTeam's Experience

The range of experiences of the WesTeam include applications of all the skills necessary to develop a DP-based data product:

- Creating numerous public use microdata files;
- Developing risk assessment software (*InitialRisk* for the National Center for Education Statistics);
- Developing perturbation software, including *DataSwap* for the National Center for Education Statistics, *SDCPert* for the Census Bureau, and the proprietary *WesSDC Toolbox*;
- Developing data analysis tools (National Household Travel Survey R Toolkit[1] and Westat's WesDaX®);
- Developing record linkage software (*WesLink*) and performing research with record linkage approaches;
- Developing software for disclosure risk assessment through advanced probabilistic modelling;
- Developing noise infusion algorithms for table generating systems; and
- Research into DP under contract to the National Center for Science and Engineering Sciences.

2. Challenge Proposal

In this section we provide a conceptual solution that describes a combination of methods in SDC and differential privacy (DP) when publicly releasing microdata files that contain numerical, geo-spatial, and categorical data. We propose algorithms that include randomized mechanisms and additive noise with the aim of optimizing privacy and utility for a range of statistical analyses methods: exploratory analysis, generation of count data, regression, classification, and clustering analysis and other models not yet specified. An important advantage of using a DP mechanism is that it is not secret. The parameters of the noise addition/ random mechanism can be made public

[1] More information at https://github.com/Westat-Transportation/summarizeNHTS.

and researchers are then able to account for the perturbation in their statistical inferences. This greatly increases the utility of the data.

There are two potential approaches that would allow the breadth of statistical analyses models mentioned in the challenge:

1. Produce a synthetic dataset (or multiple synthetic datasets) using a mechanism that is defined as differentially private (in combination with other SDC methods) and would depend on Bayesian posterior distributions as mentioned in Section 1.2. Then, all further exploratory and statistical analyses carried out on the protected data, including those specifically mentioned in the challenge, would also be differentially private. The synthetic data generation needs to consider numerical and categorical variables as well as geo-spatial data.

2. Develop an online web-based remote analyses server that will go beyond generating tables, exploratory analyses and basic statistics, and also support forms of regression modeling (linear, GLM, binary) and other statistical modelling procedures. It has been established in Rinott, et al. (2018) that online flexible table generation, which we will call ‘ Tablebuilder’, can be protected under a mechanism that is (ε, δ)- differential private in combination with standard SDC techniques that maintains a privacy guarantee with reasonable utility. Here we propose to extend the Tablebuilder to a remote analyses server. This proposal is more similar to the spirit of DP which is defined as an output perturbation mechanism.

We address each approach separately:

2.1 DP Synthetic Data

Differentially private synthetic data (or multiple releases of synthetic data) is an area of active research currently under investigation across both statistical and computer science communities. Under this setting, the synthetic data can be used in place of the original data and all statistical analyses methods can be applied, for example, regression modelling, classification and clustering. Here, we propose to follow the approach proposed by Raghunathan, et al. (2001) and van Buuren (2007) based on multiple imputation sequential regression models which depend on conditional Bayesian posterior distributions and can handle both continuous, binary and categorical variables. Bowen and Liu (2016) provide an overview of approaches that have been explored for generating synthetic data for counts, histograms and numerical variables and list their pros and cons. They also propose differentially private data synthesis techniques but our proposed approach has less reliance on distributional assumptions and all variables can by synthesized together thus improving the properties of joint distributions. We also discuss protection of geo-spatial variables in Section 2.1.4.

2.1.1 Initial SDC Approaches

Prior to carrying out the synthetic data generation, common SDC approaches should first be applied. In consultation with the users of the data, an initial SDC step is carried out based on defining which variables need to be in the data and how they should be defined. Direct identifiers are removed from the data. Age and other quasi-identifiers are typically coarsened into groupings. The level of geographical information is to be determined according to the requirements of the users. In this stage, we can also apply *k*-anonymity and related approaches to ensure that coarsened quasi-identifiers have some *a priori* privacy protection. In DP all variables are considered identifiable, but it may not be plausible in practice. For example, there may be some variables that do not need to be masked, e.g., due to legislation, such as some demographic variables in the US Census. Therefore, the dataset may need to be split into two sets of variables: $y = (y_1, \dots, y_L)$ variables that will need to be masked and $x = (x_1, \dots, x_R)$ variables that do not need to be masked.

2.1.2 Description of Multiple Imputation Sequential Regression Modelling

We first describe the multiple imputation procedure as set out by Raghunathan, et al. (2001) and shown to be useful for the purpose of generating synthetic data in Raghunathan, et al. (2003) and Reiter (2005a) in the SDC literature. The joint distribution for generating synthetic values is developed through a sequence of conditional regression models where each successive regression includes variables from the preceding regressions. We generate the data by drawing values from the corresponding predictive distributions. The types of regression models used can be linear, logistic, Poisson, generalized logit or a mixture of these depending on the type of variable to be synthesized. Multiple copies of the synthetic data are generated and inference carried out on each of the data sets and results combined for point and variance estimates under well-established combination rules.

Using the notation of Raghunathan, et al. (2001) and under the simple case of a continuous variable Y in the data (possibly transformed for normality), we fit a linear regression model $\mathrm{Y} = \mathrm{Ub} + \mathrm{e}, \mathrm{e} \sim N(0, \sigma^2 I)$ where U is the most recent predictor matrix including all predictors and previously generated variables based on x and y. We assume that $\theta = (\mathrm{b}, log\sigma)$ has a uniform prior distribution.

The coefficient b is estimated by solving the score function $Sc(\mathrm{b};\sigma) = \sum_i U_i'(\, Y_i - U_i\mathrm{b}) = 0$ and obtaining $\mathrm{b} = (\mathrm{U}'\mathrm{U})^{-1}\mathrm{U}'\mathrm{Y}$. The residual sum of squares is $SSE = (\mathrm{Y} - \mathrm{Ub})'(\mathrm{Y} - \mathrm{Ub})$ having $df = rows(\mathrm{Y}) - cols(\mathrm{U})$. Let T be the Cholesky decomposition such that $\mathrm{T}'\mathrm{T} = (\mathrm{U}'\mathrm{U})^{-1}$. To draw from the posterior predictive distributions we generate a chi-square random variable u with degrees of freedom df and define $\sigma_*^2 = \frac{SSE}{u}$. We then generate a vector $z = (z_1, \dots, z_p)$ of standard normal random variables where $p = rows(\mathrm{b})$ and define $\beta_* = \mathrm{b} + \sigma_* \mathrm{T}z$. The synthetic values for Y are $Y_* = \mathrm{U}\beta_* + \sigma_* v$ where v is an independent vector of standard normal random variables with dimension $rows(\mathrm{U})$. More details are in Raghunathan, et al. (2001) as well as descriptions for other types of models for binary and categorical variables that all depend on solving score functions.

We note that there are other ways to produce synthetic data in the SDC literature but using the multiple imputation sequential regression modelling approach is conducive to our proposal for adding a layer of protection based on DP.

2.1.3 Proposal for DP Noise Addition

We propose to add random noise to the estimating equations when estimating the coefficient b as described in Section 2.1.2 similar to the approach taken in Chipperfield and O'Keefe (2014) as follows:

- Define a random perturbation vector $v = (v_1, \dots, v_p)$ and solve the score function $Sc(\mathrm{b};\sigma) = \sum_i U_i'(\, Y_i - U_i\mathrm{b}) = v$.
- We define $v_i = s_i l_i$ where s_i is the maximum contribution a record on the microdata makes to the *i*'th coefficient of the estimating equation. Then we multiply by $l = (l_1, \dots, l_p)$ independently generated from the Laplace distribution having the range (-1,1). Since E(v)=0 and $\tilde{\mathrm{b}} = \mathrm{b} + (\mathrm{U}'\mathrm{U})^{-1}v$ we obtain a value of $\tilde{\mathrm{b}}$ that is an unbiased estimate of b. It remains to be seen in simulation studies the stability of score functions for obtaining an estimate $\tilde{\mathrm{b}}$ under different types of models and for the case of skewed data. Under these scenarios, a rejection-acceptance algorithm can be applied.
- We then follow the approach for generating the synthetic values replacing the coefficient b with the coefficient $\tilde{\mathrm{b}}$ in Section 2.1.2

We use the same intuition for adding Laplace noise in estimating equations for other types of models in the Raghunathan, et al. (2001) approach for binary and categorical variables.

The perturbation is of the same order as adding or removing the record with the largest contribution of each estimating equation. To obtain the (-1,1) interval under the Laplace Distribution, we first note that we draw random perturbations with probability proportional to $\exp(-\frac{\varepsilon|u|}{\Delta u})$ where *u* is a utility function defining the maximal difference between original and perturbed values and Δu is the sensitivity. In this case for an interval (-1,1), u takes on the value of 2 and assuming a sensitivity of 1 we obtain that for the unit interval we generate Lap(0, 1/ $\varepsilon)$ noise (Dwork and Roth, 2014, p. 31). We propose that the shape parameter in the Laplace distribution is 0.5 to ensure the interval (-1,1) but further investigation is needed through simulation studies to assess the level of protection and the utility under the proposed approach. We note that since the multiple releases of synthetic datasets are defined as DP for a given privacy budget, any subsequent statistical analysis performed on the datasets and combining rules will also be differentially private.

Finally, we note that the proposed approach can be modified for the case of generating multilevel data or longitudinal data where individuals can appear more than once in each wave of the survey and the residuals are correlated. Sakshaug and Raghunathan (2014) create synthetic data by hierarchical Bayesian modeling. In the first step, sequential regression modeling is carried out to approximate the joint distribution and in the second step the regression coefficients undergo an additional hierarchical model to preserve the between-cluster relationships. In this case, we can add Laplace noise to the estimating equations in the first stage as proposed above.

2.1.4 Geo-spatial Variables

There has been some work related to generating synthetic values for geo-coded data under the SDC framework. Wang and Reiter (2012) use a conditional bivariate regression model on latitude-longitude coordinates and then generate synthetic values of new latitude and longitude values from these models. To ease computational burden, they also propose using CART to develop the classes within which latitude-longitude should be synthesized. Starting with longitude, within each node of the tree, they carry out a Bayesian bootstrap and then use a Gaussian kernel density estimate to smooth the values. The synthetic longitude is drawn from the estimated mixture density. Synthesizing latitude is carried out similarly but conditional on the synthetic longitude as well.

Zandbergen (2014) also discusses random perturbation to spatial masking within an SDC framework and illustrates different approaches that can be implemented. These include randomly perturbing the location based on a circle of fixed radius around the true coordinates. For example, the random perturbation can be based on a bimodal Gaussian displacement. In summary, both Zandbergen (2014) and Wang and Reiter (2012) depend on distributions to generate synthetic coordinates and it is possible to add random Laplace noise either through the estimating equations as shown in Section 2.1.3 or directly on to the synthetic coordinates themselves to ensure differential privacy. The utility of these approaches will be the subject of further investigation.

2.2 Remote Analysis Server

It is well known that it is better to induce perturbation at the output level rather than the input level with respect to improving data utility, although this comes at the risk of depleting a privacy budget. The approach of producing synthetic microdata described in Section 2.1 from which all statistical analysis methods can be carried out remains to be tested on whether the generated data has analytical value. Recent publications, however, in both the computer science and statistical literature have shown that it is possible to produce differentially private tabular data with good utility. Rinott, et al., 2018 have addressed the implementation of differential privacy through an exponential mechanism on a web-based flexible table builder (which we call 'Tablebuilder') for census counts from both a theoretical and applied perspective. Other examples in the computer science literature are Barak, et al. (2007), Yaroslavtsev, et al. (2013) and Qardaji, et al. (2014). As mentioned, DP is an area of active research and has been only applied to a few operating systems (Garfinkel, 2015), though increasingly in recent years[2].

Rinott et al. (2018) show that perturbing tabular data through a Tablebuilder, based on a discretized Laplace distribution has a guarantee of privacy while showing good utility for ε, in the range of 1 or 2, and a very small δ for an (ε, δ) −differentially privacy approach. The δ basically defines the cap on the amount of perturbation that is allowed in the Tablebuilder. In addition, they show that with public knowledge of the DP mechanism, statistical inference for contingency tables can be adjusted for the perturbation. In fact, it has been recommended at the US Census Bureau that to produce a synthetic dataset of the US Census, they propose to perturb tabular data according to a DP

[2] See for example https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/

mechanism and then reconstruct the microdata based on the set of perturbed tables. This option may perform well if there are few variables in the microdata, with some variables that will not be subject to perturbation, but it will likely not provide a dataset that can handle the range of statistical analyses methods mentioned in this challenge.

The approach in developing a Tablebuilder with DP protection is to make it into a non-interactive mechanism. This means that perturbations are consistent across counts having the same domain and categories of variables spanning the table and hence there is no privacy budget spent beyond the initial budget set up in the development stage. This is carried out as follows: each record in the underlying microdata of the Tablebuilder is assigned a 'key'. Any time individuals in the microdata are aggregated to produce a count within a table defined by a domain, the keys are also aggregated and then adjusted by the domain total to form the seed of the perturbation. Therefore, counts in the same domain and having values of categorical variables spanning the table will always have the same seed and hence the same perturbation, no matter how many times the table is generated. As an illustration, assume we request a table of two cells with a query on whether individuals in a geographical region have or have not a disease. Recall that the principle of DP is that the output will be indistinguishable for a database a and a neighboring database a' with one individual removed. Under the consistency principle of a Tablebuilder, if we do not account for the domain total in the microdata key, it will be evident in which cell the target individual will belong since only one cell will change and the other cell will not change. However, by adjusting the key under the consistency property to include the domain total, both cells will change in this case and it will not be possible to learn in which cell the target individual is located.

For this challenge, we propose to extend the Tablebuilder to a remote analyses server allowing other types of statistical analyses methods besides the generation of tables and the analyses of categorical variables including the principle of consistency to preserve the privacy budget. The users access the data via an online interface similar to the Tablebuilder where the statistical model can be called via the interface. SDC literature on remote analyses servers include: O'Keefe et al. (2009) on regression outputs from a remote analyses server, O'Keefe, et al. (2012) on survival analyses, and Atikur and O'Keefe (2017) on generalized linear models.

[As an example of a remote analysis server (differential privacy mechanism not implemented to date), see https://wesdaxdemo.wesdemo.com/Default.aspx . Enter the word "demouser" (but don't include the quotation marks) for the User Name and the word "Westat!1" for the Password (but don't include the quotation marks).]

As discussed in Section 2.1.1, initial SDC approaches first need to be applied. There are many 'rules-of-thumb' when developing a remote analyses server, such as the types of statistics, models and outputs that would be allowed, minimum population threshold levels on the covariates of the models and more. Generally, there are restrictions in the outputs which do not allow single data points to be disseminated so the server will not allow minimum and maximum values, original scatter and residual plots, etc. Similar to the Tablebuilder, any statistical method that is based on sufficient statistics can be made to have consistent perturbations through the use of microdata keys similar to the concept in the Tablebuilder.

Here we propose to make the remote analyses server differentially private by adding Laplace noise to estimating equations (see Section 2.1.3) for regression modelling or directly into sufficient statistics of counts and exploratory statistics as described in Rinott, et al (2018). The use of microdata keys to ensure the consistency property described above fixes the privacy budget at the design stage of the DP mechanism. More specific details will be developed and fine-tuned for each statistical analysis method that we develop within the remote analysis server. For example, for the regression modelling in the remote analysis server, we can apply the technique in Section 2.1.3 of adding Laplace noise to the estimating equations and for exploratory analysis we can add Laplace noise directly into the statistics, and in both cases we can control the seed of the perturbation through the microdata keys to ensure consistency across domains. All scatter plots and residual plots would be presented as sequential box-plots with Laplace noise added perturbation. For geo-coded variables, the remote analyses server can produce maps containing statistical information that would be protected through coarsening and additive Laplace noise-infusion. Another way to protect statistical information in maps is to provide heat maps. Other combinations of SDC and DP will be exploited to provide high utility in the remote analysis server.

3. Data for Case Studies and Simulations

The case studies and simulations needed to develop the two DP approaches can be carried out on available public-use files that are freely available for download over the internet. These datasets have typically undergone the type of initial SDC approaches mentioned in Section 2.1.1 with direct identifiers removed and the coarsening of quasi-identifiers. Some examples of such datasets are: Survey of Doctorate Recipients 2013 or 2015 and other data available at https://ncsesdata.nsf.gov/datadownload/ and the Life in Transition Surveys available at https://www.ebrd.com/cs/Satellite?c=Content&cid=1395236498263&d=Mobile&pagename=EBRD%2FContent%2FContentLayout. In addition, there is a range of census and survey data from around the world that are available from the IPUMS website at https://www.ipums.org/ . Another important type of data to test our proposed approaches is longitudinal data, such as datasets from Understanding Society https://discover.ukdataservice.ac.uk/?q=understanding+society and English Longitudinal Study of Ageing datasets https://discover.ukdataservice.ac.uk/?q=elsa available from the UK Data Service. We note the importance of conducting the evaluation on both census/administrative data as well as survey-weighted data.

Since geo-coded data are typically not available on public files, there will be a need to generate these data based on known geographical information and clustering algorithms. The fact that geo-coded data may not be 'true' data will not negate the development and testing of the two proposed DP approaches. Once algorithms have been fully developed, they can be tested within a national statistical institute on real geo-coded data.

4. Risk-Utility Assessment

As mentioned in Section 1.2, it is necessary to carry out a risk-utility mapping to assess the DP mechanism for a given output according to a selection of parameters ε and α. Besides the privacy-by-design parameters of DP, we also propose to measure disclosure risk for identity and attribute disclosure using standard methods of probabilistic record linkage and probabilistic modelling to estimate a probability of re-identification. In addition, we also assess utility through a range of distance functions between the original and perturbed data (for example, Hellinger's Distance and the Kullback–Leibler divergence, propensity score metrics, etc.) as well as a range of metrics based on information loss in statistical analyses (for example, confidence interval overlap, the impact on Chi-Square tests for independence or R^2 in regression modelling, etc.). We will conduct the risk-utility mappings as demonstrated in Section 1.2 to identify optimal parameters under the DP approaches.

5. Summary

Traditional methods of privacy protection focus on removal of direct identifiers, coarsening, perturbation, and data suppression have failed under assault by new linkage and integration attacks within an ever-widening data environment. Any data element related to an individual entity identifies the entity to some degree, especially when linked to other data sources containing similar attributes. Some elements contribute to research and public information in general. A data release has to fit its purpose. We have to capture information about a lesion on a persons' palm, for instance, while blurring fingerprints of no value to a medical researcher. A differential privacy standard guarantees that each record released contributes to useful information about a group sufficiently large to support meaningful statistical estimates and predictions. Bayesian posterior distributions, regularization, and multiple imputation contribute to SDC in that they blur incidental features of entities that would distinguish records of one entity but contribute little to knowledge of groups of entities. These techniques are being explored in the context of DP.

We have recommended next steps forward to focus on pragmatic ways of implementing differential privacy standards in public releases of data and statistics. We have shown two proposals to produce differentially private analyses in combination with standard SDC approaches, the first based on generating synthetic microdata and the second based on a remote analyses server. In the first case of synthetic microdata, all statistical analyses methods such as regression, classification, clustering and other algorithms which are not known in advance can be carried out. It remains to be seen what level of utility can be obtained; that is, whether the level of accuracy in synthetic microdata will support useful analytic research. In the second case of a remote analyses server, we can develop the capability of most desired statistical analyses methods within the server, for example, forms of regression modelling (linear, GLM, binary (classification) etc.), clustering, exploratory analyses and mapping based on geo-coded data. In addition, the remote analyses server includes flexible table generation from which chi-square tests for independence and log-linear modelling can be carried out. Finally, since parameters of differential privacy are not secret, they can be used to adjust for the perturbation in statistical analyses. For example, Rinott, et al. 2018 demonstrate how the knowledge of the DP mechanism and parameters in a Tablebuilder can be used to adjust statistical inference on tabular data to produce unbiased test statistics.

We believe that both DP options are viable and simulation studies should be undertaken. It is likely that statistical analyses conducted through the remote analyses server will lead to higher utility compared to the case of generating synthetic microdata and then carrying out the statistical analyses, as it is well known that output perturbation outperforms input perturbation (O'Keefe and Shlomo, 2012) with respect to utility requirements. Future developments and improvements to the remote analyses server will increase the range of statistical modelling that can be carried out leading to future improvements. We note that the computing requirements and feasibility of producing synthetic data and developing a remote analyses server are well within the capabilities of national statistical institutes and other organizations and both proposed approaches can handle new forms of data as well as allow for the balance between privacy and utility.

References

Abowd J.M. and Schmutte I. (2017). Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods. https://www2.census.gov/ces/wp/2017/CES-WP-17-37.pdf.

AmStat News (2018). The EU General Data Protection Regulation is Affecting – Maybe – Your Work. Written by ASA Privacy and Confidentiality Committee, July 2018.

Atikur, R. K. and O'Keefe, C.M. (2017). Disclosure risk reduction for generalized linear model output in a remote analysis system. Data and Knowledge Engineering, Vol. 111, 90-102.

Barak, B. Chaudhuri, K., Dwork, C., Kale, S., McSherry, F. and Talwar, K. (2007). Privacy, accuracy and consistency too: a holistic solution to contingency table release. Symposium on Principles of database systems. ACM 2007, 273-282.

Bowen, C. and Liu, F. (2016). Comparative study of differentially private data synthesis methods. arXiv:1602.01063.

Chipperfield, J.O. and O'Keefe, C.M. (2014). Disclosure- protected inference using generalised linear models, International Statistical Review, Vol. 82 (3), 371-391.

Dwork C., Kenthapadi K., McSherry F., Mironov I., Naor M. (2006a) Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: Vaudenay S. (eds) Advances in Cryptology - EUROCRYPT 2006. EUROCRYPT 2006. Lecture Notes in Computer Science, Vol 4004. Springer, Berlin, Heidelberg

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In 3rd IACR Theory of Cryptography Conference 265-284.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, Vol. 9, 211-407.

Garfinkel, S. (2015). De-Identification of Personal Information. NISTIR 8053. https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

Jaro, M.A. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, 84, pp.414-20.

Ligett, K., Neel, S., Roth, A., Waggoner, B. and Wu, Z.S.. (2017) Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. In NIPS, 2017.

McLaughlin, R. H., Clarke, C. A., Crawley, L. M., & Glaser, S. L. (2010). Are Cancer Registries Unconstitutional? *Social Science & Medicine (1982)*, *70*(9), 1295–1300. http://doi.org/10.1016/j.socscimed.2010.01.032

O'Keefe, C.M. and Good, N.M. (2009). Regression output from a remote analysis system. Data and Knowledge Engineering, Vol. 68, 1175–1186.

O'Keefe, C.M., Sparks, R., McAullay, D. and Loong, B. (2012). Confidentializing survival analysis output in a remote data access system. Journal of Privacy and Confidentiality, Vol. 4, 127-154.

O'Keefe, C.M. and Shlomo, N. (2012). Comparison of remote analysis with statistical disclosure control for protecting the confidentiality of business data. Transactions on Data Privacy, Vol. 5 (2), 403-432.

OMB (2013). "Memorandum Number M-13-13: Open Data Policy—Managing Information as an Asset." Office of Management and Budget, Executive Office of the President. Washington, DC: OMB, May 9, 2013. Available at http://www.whitehouse.gov/omb/memoranda_default/

PCC (2017). Developing a Privacy Policy. Dated December 20, 2017. Written by the American Statistical Association's Privacy and Confidentiality Committee. Available at: http://higherlogicdownload.s3.amazonaws.com/AMSTAT/284c0271-770e-46ef-975c-d99a87486bd3/UploadedImages/PC_Privacy_Policy_Guidelines-2017-Dec20.pdf

Qardaji, W., Yang, W. and Li, N. (2014). Preview: practical differentially private release of marginal contingency tables. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, ACM 2014, 1435-1446.

Raghunathan T.E., Lepkowksi J.M., van Hoewyk J., Solenbeger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, Vol. 27, 85-95.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics, Vol. 19, 1-16.

Reiter, J. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A, Vol. 168 (1), 185-205.

Reiter, J.P. (2005b). Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association 100, 1103-1112.

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. To be published in Statistical Sciences.

Sakshaug, J. and Raghunathan, T.E. (2014). Generating Synthetic Microdata to Estimate Small Area Statistics in the American Community Survey. Statistics in Transitions, New Series, Issue 3, 341-368.

Skinner, C. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. J. Amer. Statist. Assoc. 103 989–1001.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, Vol. 16(3), 219-242.

Wang, H. and Reiter, J. (2012). Multiple imputation for sharing precise geographies in public use data. Annals of Applied Statistics, Vol. 6, 229–252.

Winglee, H., Valliant, R., Clark, J. and Lim, Y., Weber, M., and Strudler, M. (2002). Assessing Disclosure Protection for a SOI Public-use File, ASA Conference presentation.

Winkler, W. (1993). Matching and Record Linkage. U.S. Census Bureau.

Yaroslavtsev, G., Cormode, G., Procopiuc, C.M. and Srivastava, D. (2013). Accurate and efficient private release of datacubes and contingency tables. In ICDE, 2013.

Zandbergen, P. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. Advances in Medicine. Available at: https://www.hindawi.com/journals/amed/2014/567049/