November 2013

A New Method for Protecting Interrelated Time Series with Bayesian Prior Distributions and Synthetic Data

Matthew J. Schneider
mjs533@cornell.edu

John M. Abowd
Cornell University, jma7@cornell.edu

A New Method for Protecting Interrelated Time Series with Bayesian Prior Distributions and Synthetic Data

Abstract
Organizations disseminate statistical summaries of administrative data via the Web for unrestricted public use. They balance the trade-off between confidentiality protection and inference quality. Recent developments in disclosure avoidance techniques include the incorporation of synthetic data, which capture the essential features of underlying data by releasing altered data generated from a posterior predictive distribution. The United States Census Bureau collects millions of interrelated time series micro-data that are hierarchical and contain many zeros and suppressions. Rule-based disclosure avoidance techniques often require the suppression of count data for small magnitudes and the modification of data based on a small number of entities. Motivated by this problem, we use zero-inflated extensions of Bayesian Generalized Linear Mixed Models (BGLMM) with privacy-preserving prior distributions to develop methods for protecting and releasing synthetic data from time series about thousands of small groups of entities without suppression based on the of magnitudes or number of entities. We find that as the prior distributions of the variance components in the BGLMM become more precise toward zero, confidentiality protection increases and inference quality deteriorates. We evaluate our methodology using a strict privacy measure, empirical differential privacy, and a newly defined risk measure, Probability of Range Identification (PoRI), which directly measures attribute disclosure risk. We illustrate our results with the U.S. Census Bureau's Quarterly Workforce Indicators.

Keywords
synthetic data, zero-inflated mixed models, informative prior distributions, probability of range identification, PORI, statistical disclosure limitation, SDL, empirical differential privacy, big data privacy, Bayesian

Comments
Presentation given by LDI Graduate Assistant, Matthew Schneider, on big data privacy at Temple University on Nov 18, 2013.

A New Method for Protecting Interrelated Time Series with Bayesian Prior Distributions and Synthetic Data

Matthew J. Schneider[1] and John M. Abowd[2]

Abstract Organizations disseminate statistical summaries of administrative data via the Web for unrestricted public use. They balance the trade-off between confidentiality protection and inference quality. Recent developments in disclosure avoidance techniques include the incorporation of synthetic data, which capture the essential features of underlying data by releasing altered data generated from a posterior predictive distribution. The United States Census Bureau collects millions of interrelated time series micro-data that are hierarchical and contain many zeros and suppressions. Rule-based disclosure avoidance techniques often require the suppression of count data for small magnitudes and the modification of data based on a small number of entities. Motivated by this problem, we use zero-inflated extensions of Bayesian Generalized Linear Mixed Models (BGLMM) with privacy-preserving prior distributions to develop methods for protecting and releasing synthetic data from time series about thousands of small groups of entities without suppression based on the of magnitudes or number of entities. We find that as the prior distributions of the variance components in the BGLMM become more precise toward zero, confidentiality protection increases and inference quality deteriorates. We evaluate our methodology using a strict privacy measure, empirical differential privacy, and a newly defined risk measure, Probability of Range Identification (PoRI), which directly measures attribute disclosure risk. We illustrate our results with the U.S. Census Bureau's Quarterly Workforce Indicators.

[1]Address for Correspondence: Samuel C. Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853: Email: mjs533@cornell.edu

[2]Department of Economics, Department of Statistical Sciences, and Labor Dynamics Institute, Cornell University, Ithaca, NY 14853 USA. Email: john.abowd@cornell.edu

We acknowledge funding from NSF grants BCS 0941226, SES 9978093, ITR 0427889, SES 0922005, and SES 1131848. We thank Stephen Fienberg, Sachin Gupta, Sharan Jagpal, Aleksandra B. Slavkovic, and Lars Vilhuber for helpful comments. The scientific inputs and programming used in this article have been deposited in a curated data repository run by the Cornell University Libraries http://digitalcommons.ilr.cornell.edu/ldi/20/.

Keywords: synthetic data; zero-inflated mixed models; informative prior distributions; administrative data; statistical disclosure limitation (SDL); empirical differential privacy

I. Introduction

Various organizations across the globe disseminate statistical summaries of administrative data via the Web for unrestricted public use. These products describe individuals or businesses either in micro-data or tabular format. Disclosure avoidance methods (also called statistical disclosure limitation methods) are required to enable dissemination beyond the trusted users.

Agencies choose disclosure avoidance methods to balance "the level of protection provided and the effects on the ability of users to draw valid inferences" (Duncan et al. 1993, pages 10-11). For example, in 2003, the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program released the Quarterly Workforce Indicators (QWIs), a collection of highly detailed local labor market time series with 32 economic measures (*e.g.*, the number of jobs created, accession, jobs destroyed, and separations) categorized by gender, age, race, ethnicity, education, U.S. county, U.S. state, North American Industrial Classification (NAICS), and business ownership. At the most detailed level, the micro-data contain many zeros. Current disclosure avoidance rules suppress the publication of some of these data, a problematic method because it degrades the ability of users and researchers to draw valid inferences about the underlying micro-data. Another challenge is that formal privacy measures (*e.g.*, differential privacy, Dwork 2006) do not apply well to very detailed data sets where many of the candidate publication cells are empty, such as the QWI data, because the level of protection they impose is too great for meaningful inference (Abowd and Schneider 2011). Consequently, more realistic protection methods and measures are needed for models that accurately capture the information present in such detailed data sets.

Hotz et al. (1998) conducted a comprehensive review of agencies around the world and found that disclosure avoidance techniques varied in practice. In the United States, the Federal Committee on Statistical Methodology, which is composed of members from the principal statistical agencies of the federal government, maintains a statistical policy paper (FCSM 2005) that catalogues the disclosure limitation practices of these official statistics publishers. International statistical agencies have also discussed these issues regularly. See OECD (2012).

Agencies frequently used suppression, swapping, sub-sampling, coarsening, top coding for continuous variables, limitation of geographic details, aggregation, and discretization of continuous variables, as well

as newer methods like controlled tabulation, noise infusion, and limited synthesis (sometimes called "blank and impute"). The disclosure avoidance technique used depends on the type of data being protected. For tabular data, such as count tables, applicable methods included cell suppression, interval publication, cell rounding, cell perturbation, and controlled tabulation. For micro-data, methods included noise addition, noise multiplication, suppression, coarsening, aggregation, and swapping. However, many of these methods overprotect and therefore, researchers who require more detail in order to make valid inferences on micro-data often request, and are granted, restricted access via agency supervised enclaves. For a very complete summary of methods see Duncan, et al. (2011).

Recent developments in advanced disclosure avoidance techniques for micro-data include noise infusion (OECD 2012) and synthetic data (Machanavajjhala et al. 2008; Reiter 2005b; Kohnen and Reiter 2009). The first large-scale use of noise infusion in any official statistical product occurred in 2003 on the U.S. Census Bureau's Quarterly Workforce Indicators (Abowd et al. 2009). Drechsler and Reiter (2010) showed improvements by using synthetic data with classification and regression trees compared to common disclosure avoidance techniques like sampling.

Rubin (1993) and Little (1993) proposed the use of fully and partially synthetic data in the same special issue of the *Journal of Official Statistics*. Rubin's method was based on multiple imputation, using a Bayesian model to fit the confidential data, then releasing samples in which all released variables were draws from the posterior predictive distribution, and thus not actual respondent data. Little's method, known as partially synthetic data, used similar methods but only replaced some of the variables or some of the records with synthetic data before release.

Machanavajjhala et al. (2008) performed the first formal privacy analysis with synthetic data that was adopted by a U.S. statistical agency (the Census Bureau). Their method was developed for sparse multinomial data. The analysis was formal because the released data were provably protected with a level of differential privacy, a theoretical concept in which the relative change in knowledge produced by the release (the posterior odds ratio) is bounded. In related work, Abowd et al. (2012) proposed eliminating the limited suppressions that often used in conjunction with noise infusion by combining a synthetic data model with multiplicative noise distortion applied directly to the micro-data. Abowd and Schneider (2011) added formal noise directly to the parameter estimates of a Linear Mixed Model without the use of synthetic data. They found that valid inferences on detailed data sets like the QWIs were not possible when differential privacy for a data model with bounded release variables and a limited number

of parameters. Consequently, Abowd, Schneider, and Vilhuber (2013) relaxed the definition of differential privacy to empirical differential privacy, which bounds the posterior odds ratios of a Bayesian model with and without the most influential observation. They used a non-informative prior to measure empirical differential privacy for a normally distributed dependent variable, but did not protect the Bayesian model; that is, they did not provide a method for controlling the privacy protection in the Bayesian model by direct manipulation of the prior distributions.

Building upon these recent developments, our research protects the best fitting Bayesian synthetic data models for detailed data sets (*i.e.*, those with thousands of categories, time series, and non-normally distributed dependent variables) with informative prior distributions.

This paper proposes a method that would allow disclosure avoidance procedures at statistical agencies to be applied to the release of detailed micro-data, using a formal protection method that is incorporated into the estimation of Bayesian statistical model of the confidential data. The proposed approach may be more efficient compared than current rule-based procedures, which require small and sometimes extremely large values in micro-data to be suppressed even after other protections have been applied. Our proposed modeling approach randomly generates small counts (including zero) using synthetic data from a zero-inflated mixed model, thereby obviating the need to suppress the data observation. Mixed models with a zero-inflated structure have demonstrated better fits than fixed-effects models with zero-inflation or mixed-effect models without (Hall 2000).

We also propose a new measure called the Probability of Range Identification (PoRI), which protects lumpy zeros and large counts by ensuring the magnitude of released synthetic data is not usually close to the true data. An alternative approach in the literature measures the probability of record linkage in a data set (Reiter 2005a), but our approach differs in that we use a real-valued dependent variable. An agency can use the PoRI along with the existing measure of empirical differential privacy to decide the desired degree of protection.

Although the method is designed to be applied directly to the confidential micro-data, and then used to create a synthetic version of those data for publication, we test all of our procedures on the public-use Quarterly Workforce Indicators, which are actually tabulations of the underlying micro-data. Our reasons are two-fold. First, this allows us to test and refine the methods in a scientific forum where we can freely share both the input and output data products. To this end, the scientific inputs and programming have been deposited in a curated data repository run by the Cornell University Libraries http://digitalcommons.

ilr.cornell.edu/ldi/XX/.

Methodologically, we present a solution to the disclosure avoidance problem by using a single probability model with a privacy-preserving prior distribution whose hyper-parameters are fixed before there is any model estimation. Our use of prior distributions on the variance components of a Bayesian GLMM with zero inflation is a new disclosure limitation method. Further, unlike previous research we neither add noise to parameter estimates ex-post model estimation (Abowd and Schneider 2011) nor to the data directly (Abowd et al. 2012). We are also the first to apply empirical differential privacy to the class of Bayesian Generalized Linear Mixed Models (BGLMM), which accommodates non-normally distributed dependent variables. We also propose a new privacy metric (PoRI) that can be used to control the posterior probability of attribute disclosure.

This paper proceeds by introducing the model and its corresponding prior distributions in Section II. In Section III, we explain empirical differential privacy in relation to our disclosure avoidance methodology, and define PoRI. Then, in Section IV, we apply our disclosure avoidance methodology to the Census Bureau's QWI and discuss the results. Finally, in Section V, we conclude.

II. Model Specification

A. The Bayesian Zero-Inflated Poisson Mixed Model

Rubin (1993) proposed the use of fully synthetic data constructed using multiple draws from the Bayesian posterior predictive distribution fitted to the underlying confidential data. We use synthetic data produced from a Bayesian Zero-Inflated Poisson Mixed Model, which has several advantages in our application. The first is that mixed effect models are currently used by many statistical agencies for estimating small geographical areas and therefore, our privacy routines can be readily incorporated into procedures designed to protect such data. The second is that our model is an extension of a Generalized Linear Mixed Models (GLMMs) and ,as a result, the distribution of the dependent variable can be altered with little difficulty.The third is that as the prevalence of zeros in our dependent variable increases, the zero-inflation parameter mixes more quickly and computation time is decreased (Hadfield 2010).

Following specification in Hadfield (2010), we model the dependent variable, y_i, for $i = 1, 2, ...N$, using a zero-inflated Poisson (ZIP) likelihood. The data have been stacked so that the first time series, relating to a particular industry and geography, is in the first T rows, followed by the next time series, which relates to the next industrial and geographic unit in rows $T + 1$ to $2T$, and so forth. We construct the appropriate design matrices below to incorporate the industrial and geographic structure into the estimation.

The ZIP likelihood is formed using two latent variables. The first latent variable is the Poisson random variable ($\dot{\eta}$) with link function is on the log scale. The second latent variable is the zero inflation random variable ($\ddot{\eta}$) with link function on the logit scale. A link function establishes the relationship between the mean of the dependent variable and the latent variable, which is modeled using a linear mixed-effect predictor. The two latent variables have the structure

$$\dot{\eta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{u} + \boldsymbol{\xi}_{1i} \tag{1}$$

$$\ddot{\eta}_i = \boldsymbol{\xi}_{2i} \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects in the Poisson linear predictor, $\boldsymbol{u}$ is the vector of random effects in the Poisson linear predictor, $\boldsymbol{x}_i$ is a row of the design matrix for the fixed effects, $\boldsymbol{z}_i$ is a row of the design matrix for the random effects, $\boldsymbol{\xi}_{1i}$ is the linear prediction error for the Poisson linear predictor, and $\boldsymbol{\xi}_{2i}$ is the linear prediction error for the zero-inflation latent variable. Notice that there are neither fixed nor random effects, only a prediction error, for the latent zero-inflation random variable in equation (2). The parametric structure is completed by specifying the joint distribution of $\left[\begin{array}{cccc} \boldsymbol{\beta}^T & \boldsymbol{u}^T & \boldsymbol{\xi}_1^T & \boldsymbol{\xi}_2^T \end{array} \right]^T$ as

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \\ \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{R}_2 \end{bmatrix} \right) \tag{3}$$

$$\begin{aligned} \boldsymbol{B} &= \boldsymbol{I}_{\dim(\beta)} \\ \boldsymbol{G} &= \begin{bmatrix} \sigma^2_{c_1} \boldsymbol{I}_{\dim(c_1)} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sigma^2_{c_m} \boldsymbol{I}_{\dim(c_m)} \end{bmatrix} \\ \boldsymbol{R}_1 &= \sigma^2_{\xi_1} \boldsymbol{I}_N \\ \boldsymbol{R}_2 &= \sigma^2_{\xi 2} \boldsymbol{I}_N \end{aligned} \tag{4}$$

We have assumed that the random effects are independent with constant variances $\sigma^2_{c_1}, \sigma^2_{c_2}, \ldots, \sigma^2_{c_m}$. The number of random effect categories, m, is determined by our model selection procedures. Serial dependence among the individual time series is modeled by vectors of lagged regressors, and is incorporated

into the fixed effects. In the Poisson process, $\mathbf{X}$ is the fixed effects design matrix and $\mathbf{Z}$ is the random effects design matrix. The rows of $\mathbf{X}$ and $\mathbf{Z}$ are designated $\boldsymbol{x}_i^T$ and $\boldsymbol{z}_i^T$, respectively.

The distribution of the dependent variable relies is derived directly from the link functions and the latent random variables $\dot{\eta}_i$ and $\ddot{\eta}_i$. Specifically, the total probability that $y_i = 0$ is the sum of the probability that $y_i = 0$ from the zero-inflation process and the probability that $y_i = 0$ from the Poisson process, given that it did not come from the zero-inflation process. The probability that $y_i > 0$ depends on the probabilities from the Poisson process, given that it was not zero from the zero-inflation process. Formally, the equations are

$$\Pr\left[y_i = 0 \left| \boldsymbol{x}_i, \boldsymbol{z}_i \right.\right] = \frac{\exp(\ddot{\eta})}{1+\exp(\ddot{\eta})} + \left(1 - \frac{\exp(\ddot{\eta})}{1+\exp(\ddot{\eta})}\right) f_{Pois}\left(y_i = 0 \left| \exp(\dot{\eta}_i)\right| \boldsymbol{x}_i^T, \boldsymbol{z}_i^T\right) \tag{5}$$

$$\Pr\left[y_i = 1, 2, .. \left| \boldsymbol{x}_i^T, \boldsymbol{z}_i^T \right.\right] = \left(1 - \frac{\exp(\ddot{\eta})}{1+\exp(\ddot{\eta})}\right) f_{Pois}(y_i \left| \exp(\dot{\eta}_i)\right| \boldsymbol{x}_i^T, \boldsymbol{z}_i^T \tag{6}$$

$$E\left[y_i \left| \boldsymbol{x}_i^T, \boldsymbol{z}_i^T, \boldsymbol{\beta}, \boldsymbol{u} \right.\right] = \mu_i = \exp(\dot{\eta}_i)\left(1 - \frac{\exp(\ddot{\eta})}{1+\exp(\ddot{\eta})}\right) \tag{7}$$

To complete the Bayesian specification of the model, we state the prior distributions of the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{B}$, $\boldsymbol{G}$, $\boldsymbol{R}_1$, and $\boldsymbol{R}_2$, respectively. The prior mean, $\boldsymbol{\beta}_0$, of the fixed effects is always zero, and its prior covariance matrix is the identity, as specified above, so there are no hyper-parameters for the fixed effects. The priors for all of the variance components in the random effects and prediction errors are Inverse Gamma. The specification of these prior distributions is used to control the formal privacy properties of the model. We discuss these in the next subsection.

B. The Privacy-preserving Prior Distributions

We use the prior distributions on the variance components to control the privacy properties of our protected synthetic data. We focus primarily on the prior distributions of the random effects, $\boldsymbol{u}$, because estimated random effects often rely on only a few observations, and our prior work suggests that it is the contribution of influential observations to an estimated random effect that cause the empirical differential privacy limits to move. The random effects model an observation's deviation from the global mean of the model. Therefore, the estimated random effects are likely to have a larger disclosure risk than the fixed effects, each of which usually depends on a large number of observations, none of which is very influential. The Bayesian GLMM controls the amount of information contained in the estimated random

effects by shrinking the estimated of the random effects toward the global mean of our model.

The prior distributions of the variances of the random effects, $\sigma^2_{c_1}, \sigma^2_{c_2}, ...\sigma^2_{c_m}$, in the $\boldsymbol{G}$ matrix are each assumed independent with an Inverse-Gamma distribution. The Inverse-Gamma random variable is non-negative with hyper-parameters ν and V and its density is

$$\sigma^2_{c_1} \sim IG(V_{c_1}, \nu_{c_1}), .., \sigma^2_{c_m} \sim IG(V_{c_m}, \nu_{c_m}) \tag{8}$$

$$p\left(\sigma^2_{c_r}|V_{c_r}, \nu_{c_r}\right) = \frac{|\nu_{c_r} V_{c_r}|^{\frac{\nu_{c_r}}{2}}}{2^{\frac{\nu_{c_r}}{2}}\Gamma\left(\frac{\nu_{c_r}}{2}\right)} |\sigma^2_{c_r}|^{-\frac{\nu_{c_r}+2}{2}} \exp\left(-\frac{1}{2}\frac{\nu_{c_r} V_{c_r}}{\sigma^2_{c_r}}\right) \tag{9}$$

for $r = 1, \dots m$. As the variances of the random effects are scaled toward zero,the linear predictor shrinks toward the global mean. Consequently, the estimated random effects are less informative about the specific category (in our data a geographic unit), affording more protection in a manner that we formalize below. The choices of ν_{c_r} and V_{c_r} define the strength of the informative prior, P_m, which we then compare to a non-informative prior P_0 in our empirical application. To increase the protective strength of the prior, we set the prior mean (Hadfield 2010), $(\nu_{c_r} V_{c_r})/(\nu_{c_r} - 2)$, of the variance components near zero with small variance, $(\nu^2_{c_r} V^2_{c_r})/((\nu_{c_r} - 2)^2\left(\frac{\nu_{c_r}}{2} - 2\right))$. Specifically, as the degree of belief parameter $\nu_{c_r} \to \infty$, the prior mean approaches V_{c_r} and the prior variance approaches zero, given $\nu_{c_r} > 4$. The two hyper-parameters interact in determining the location and spread of the prior distribution on the variance component variances. Our approach varies both parameters so that the mean of the prior distribution stays fixed near zero and the variance of the prior distribution tightens.

The prior distributions on the fixed effects, $\boldsymbol{\beta}$, and error variances, $\sigma^2_{\xi_1}$ and $\sigma^2_{\xi_2}$, are kept relatively diffuse, but proper, because their corresponding posterior estimates depend on many observations, and influential points move them relatively little. These prior distributions and densities are

$$\sigma^2_{\xi_1} \sim IG(V_{\xi_1}, \nu_{\xi_1}), \sigma^2_{\xi_2} \sim IG(V_{\xi_2}, \nu_{\xi_2}), \boldsymbol{\beta} \sim MVN(\mathbf{0}, \boldsymbol{I}) \tag{10}$$

$$p(\boldsymbol{\beta}\ \ |\boldsymbol{I}) = (2\pi)^{-\dim(\boldsymbol{\beta})/2} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \tag{11}$$

$$p\left(\sigma^2_{\xi_s}|V_s, \nu_s\right) = \frac{|\nu_s V_s|^{\frac{\nu_s}{2}}}{2^{\frac{\nu_s}{2}}\Gamma\left(\frac{\nu_s}{2}\right)} |\sigma^2_{\xi_s}|^{-\frac{\nu_s+2}{2}} \exp\left(-\frac{1}{2}\frac{\nu_s V_s}{\sigma^2_{\xi_s}}\right) \tag{12}$$

for $s = 1, 2$.

III. Privacy Measures

We define two measures of privacy in this section. The first is ϵ-empirical differential privacy, which bounds the posterior odds ratios of a Bayesian model with and without the most influential observation. The second is a new privacy metric called the probability of range identification (PoRI), which measures the posterior probability of inferring a close range of the confidential dependent variable given the synthetic data. Our empirical section investigates how these change when we vary the hyper-parameters in the prior distributions.

A. Empirical Differential Privacy for Bayesian Models

Empirical differential privacy for Bayesian Models was originally defined by Abowd, Schneider, and Vilhuber (2013), and details for calculation can be found in that paper. The value ϵ represents a bound across all observations in a given data set and is therefore an worst-case measure of privacy breeches. It is a highly sensitive to outliers and influential data points. From an inference point of view, ϵ measures the bounds of the logarithm of the posterior odds ratios of all released estimates produced from two models. The first model includes all observations and the second model excludes the most influential observation. Candidate influential observations are selected by calculating the posterior mean residuals on the response scale. We expect ϵ to change based upon the selection of prior distribution. Empirical differential privacy can be used on any Bayesian GLMM; however, we focus on the Bayesian ZIP mixed model in our empirical section since it is the best fitting model for the our application.

B. Probability of Range Identification (PoRI)

PoRI measures the posterior probability of inferring a sensitive range of a real-valued y_i given the synthetic data and knowledge of the design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$. We expect this measure to be a major concern at statistical agencies using noise infusion currently multiply the input variables by a noise factor formed such that no input value is within a certain percent of its true value (Abowd et al. 2012). Although multiplicative noise protects the data at face value, regression-based methods could possibly recover true values.

Suppose the statistical agency wishes to protect the confidential dependent variable, $\boldsymbol{y}$, by releasing synthetic data $\boldsymbol{y}^s$. Further suppose that y_i follows the Zero-inflated Poisson process defined above. The fixed effects design matrix $\boldsymbol{X}$ and the random effects design matrix $\boldsymbol{Z}$ are assumed known. This is

a reasonable assumption because design matrices in our empirical application consist of industries and geographies which are public knowledge. The synthetic data is assumed known, but the confidential values of $\boldsymbol{y}$ (*i.e.*, inside the firewall, not released) are not known. The agency chooses the sensitive range of disclosure of y_i for each i, y_{i,b_1} to y_{i,b_2} and measures PoRI for each observation as

$$PoRI_i = \int_{y_{i,b_1}}^{y_{i,b_2}} \int p(y_i|\theta, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}^s) p(\theta = (\boldsymbol{\sigma}^2, \boldsymbol{\beta}, \boldsymbol{u})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}^s) d\theta dy_i \tag{13}$$

It is important for the agency to set meaningful thresholds. We suggest a rule-based framework (*e.g.*, $y_i \pm 10\%$ of y_i or $y_i \pm 10\%$ of group standard deviation). To simulate PoRI for each i, first we recover estimates of the model parameters given the synthetic data and design matrices. Then, we generate T draws of each i from the posterior predictive distribution of y_i which is conditional on our estimates of the model parameters and the known design matrices. We count the proportion of posterior samples from y_{i,b_1} to y_{i,b_2} and set it equal to $PoRI_i$. If there are no posterior samples in that range, $PoRI_i = 0$, and there is no chance of inferring a close range of y_i.

There are several differences between $PoRI_i$ and ϵ. First, ϵ is found by empirically searching the maximal change over all parameters and influential observations, but $PoRI_i$ is calculated for every observation. Second, ϵ was measured from posterior samples of model parameters, whereas PORI was measured from the posterior predictive distribution of the dependent variable. However, both a low value of ϵ and a low value of $PoRI_i$ are associated with higher degrees of privacy.

IV. Empirical Application

A. The Quarterly Workforce Indicators

The Quarterly Workforce Indicators (QWIs) are a collection of detailed local labor market indicators comprising 32 economic measures (*e.g.*, the number of jobs created, number of jobs destroyed, number of workers hired, number of workers separated, earnings, etc.) categorized by gender, age, race, ethnicity, education level, U.S. county, U.S. state, NAICS industry classification, and owner. These data are published online at http://qwiexplorer.ces.census.gov/ and are available to the public. At the U.S. county by NAICS sector aggregation level, there are over 50,000 time series (two million observations) for each indicator. Further disaggregation by ethnicity, age, gender, or education level result in tens of millions of time series. Currently, much of these data at the disaggregated levels are either not released because one of the indicators fails a publication quality standard that applies to indicators based on just a few jobs. One

economic indicator is the number of job creations. Researchers can request the number of job creations by selecting from the available categories. For example, a researcher may request the number of job creations last quarter for Cuyahoga County, Ohio in the manufacturing sector.

Job creations are measured on an establishment level and are defined as $JC_t = \max(0, E_t - B_t)$ where E_t is the number of people employed at the end-of-quarter. A person is considered end-of-quarter employed if the wage record submitted for that job (person, employer combination) has earnings greater of at least \$1.00 in quarters t and $t+1$ for a given employer. B_t is the number of beginning-of-quarter employed persons. A person is considered beginning-of-quarter employed if the wage record submitted for that job has earnings of at least \$1.00 in quarters $t-1$ and t for a given employer. Since each establishment belongs to a county c and industry j, B_t and E_t are aggregated to the county by industry level for all establishments in the indicated industry that operate in that, sometimes very small, geographic area.

When the current disclosure avoidance protocol would have suppressed a value, thus restricting the disaggregated information available to the public, our method releases protected synthetic value. In addition, even when the current protocol would have released a noisy value, our proposed method releases the synthetic value.

We evaluate the release of the synthetic data, based on a Bayesian GLMM with a privacy-protecting prior, by comparing it to data that would be released from best fitting Bayesian model with a very diffuse prior. Thus, we examine the trade-off between the goodness of fit (*e.g.*, the Deviation Information Criterion (DIC) and the correlation of fitted vales to true values) and disclosure risk (*e.g.*, empirical differential privacy and PoRI) of the synthetic data. For small values including zero, our synthesizer generates a small value, which can be zero, using synthetic data from the appropriate estimated ZIP model, thereby obviating the need to suppress the observation. For data with large magnitudes, we randomly generate a different value, again using the appropriate posterior predictive distribution. Note that in the application to actual micro-data, this would be equivalent protecting the value of a large business. In the published QWI, this protection isn't necessary because the data do not relate to a single establishment, but since we are using the QWI to simulate the relevant micro-data, we apply the protection to demonstrate its effectiveness.

We evaluate our proposed approach using job creations data for the state of Ohio. We use these data at the NAICS sector by county level. Although the published data are clearly not confidential, we use them as proxies for the actual confidential data because they have a similar statistical structure to the

underlying confidential data.

Our dependent variable $\mathbf{y}$, consists of elements y_{jct}, the count of job creations in NAICS sector j for county c in quarter t, where $t = 1, 2, ..., 80$ from 1990:q2 to 2010:q1), and N is the total number of observations. $\boldsymbol{X}$ is the design matrix for the fixed effects (industries and lagged regressors) and $\boldsymbol{Z}$ is the design matrix for the random effects (counties and interactions chosen by model selection). At the current level of aggregation, we treat the observed job creations ($\mathbf{y}$) as confidential and the synthetic values ($\mathbf{y}^s$) as the proposed non-confidential release-data produced by our model. Since the structure at the county-by-industry level of aggregation is the same as in the micro-data, we expect our results to be generalizable for the actual confidential micro data. In Ohio, at the current aggregation level, there are $N = 51,582$ observations, 88 counties, 19 industries, and approximately a 3.4% prevalence ratio of zeros which quickly increases with additional levels of disaggregation.

B. Model Selection

We performed model selection procedures across three levels of hierarchical interactions. Table I displays the DIC and goodness of fit statistics between $\boldsymbol{y}$ and the model fitted values ($\hat{\boldsymbol{y}}$) for seven candidate models. DIC is defined as the sum of negative two times the log likelihood, the effective number of parameters, and a fixed constant (see Hadfield 2010 for the ZIP-specific formulas). The conditional correlation is the correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = E(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})$,where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the estimated posterior means of $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. The truncated correlation is also the correlation between $\boldsymbol{y}$ and $E(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})$ except that $E(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})$ is truncated at 8,000. The synthetic fit measures the correlation between $\boldsymbol{y}^s$ and $\boldsymbol{y}$, where $\boldsymbol{y}^s$ is a draw from the posterior predictive distribution of $\boldsymbol{y}$ given the original $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$.

For all models, we set a particular proper, relatively non-informative prior P_1 as the reference prior that would be used in an unfettered analysis of the original, confidential data. For P_1, we set the prior mean of $\boldsymbol{\beta}$ to $\mathbf{0}$ with a variance of $\boldsymbol{I}$. We set the prior of $\sigma^2_{\xi_1}$ to have a degree of belief parameter $\nu = 10$ and a centrality parameter of $V = 1$, whereas, $\nu = 10$ and $V = 1$ was set for the rest of the variance components of the random effects. One issue in estimation is that the variance component related to the zero inflation parameter, $\sigma^2_{\xi_2}$, cannot be factored out from $\sigma^2_{\xi_1}$ when there is no data information to distinguish zeros sampled form the Poisson from zeros sampled from the inflation process (Hadfield, 2010). Therefore, we fixed $\sigma^2_{\xi_2} = 1$.

We selected the model with three-way interactions because it had the lowest DIC and highest correlations. The chosen model results in the estimation of 8,356 fixed and random effects, which means that

Goodness of Fit Statistics for Candidate Models

Model	DIC	Conditional Fit	Truncated Fit	Synthetic Fit	Number of Cells
No Interactions	$372,944$	3.9%	75.0%	5.1%	113
County by Quarter	$372,969$	5.9%	75.7%	7.3%	465
County by Industry	$371,403$	4.3%	77.3%	4.7%	$1,684$
Quarter by Industry	$372,532$	58.2%	77.2%	51.4%	189
All 2-Way	$370,104$	79.7%	81.5%	75.0%	$2,112$
3-Way and All 2-Way	$369,350$	82.5%	82.9%	78.1%	$8,356$

Notes: Models in column (1) refer to the specification of the design of the fixed and random effects in equation (1). County is the geographic county identifier. Quarter is the quarterly seasonal. In the interaction models, Industry is the NAICS sub-sector. 2-way interactions involving industry are random effects. All specifications are hierarchical.

TABLE I

there is about one parameter for every six observations. Additionally, there are seven variance components. The large differences between the truncated correlations and the conditional correlations are due to the 12 observations in Ohio with more than 8,000 job creations in a quarter. These were also the observations that produced highest values of ϵ, the empirical differential privacy.

For the best fitting model with prior P_1, we searched over all posterior mean residuals and gathered 10 candidate influential observations to use for the measurement of ϵ. The deletion of the most influential observation produced an ϵ as high as 2.8, but results for the other 9 influential observations ranged from 0.7 to 1.5. Since ϵ is defined as the maximum across all observations, the empirical $\epsilon-$differential privacy was 2.8, which is a bound on the posterior odds of 16.4:1. The observation, $i = 41,399$, which determined ϵ was a large magnitude observation and not a zero. It had 52,019 job creations in a county-industry sector (county: Cuyahoga, NAICS sector: Health Care and Social Assistance) which typically averaged 6,600. This observation occurred in the last quarter of 2009. For comparison, all the county-NAICS sectors in Ohio averaged 167 job creations per quarter with a standard deviation of 569.

Synthetic data only slightly lowered the fit for the very best fitting models with non-informative priors, but would permit the release of protected data with similar variation and fit as the real data. The introduction of synthetic data also allowed the release of values of different magnitudes for true zeros in our data and for high magnitude observations. According to Table I, the synthetic data have a correlation of 78.1% with the actual data.

We analyzed one hundred draws of synthetic data from the posterior predictive distribution of the model in the last row of Table I. This was the best-fitting model on the DIC criterion. We found the following

key differences between the synthetic and actual data. The overall prevalence of zeros in the actual data was 3.4% compared to an average overall prevalence in the synthetic data of 2.2%. Most of the zeros in the best fitting model were picked up by the Poisson process instead of the Zero-inflation parameter. If this model were applied to data with a higher prevalence of zeros, we would expect the Zero-inflation parameter to contribute more to the generation of zeros than the Poisson process.

The maximum number of job creations in the actual data was 52,019 compared to an average maximum number of job creations in the synthetic data of 15,820. The mean and standard deviation of the number of actual job creations were 167 and 569, respectively, compared to 162 and 509 in the synthetic data.

To further explore potential limitations of the synthetic data, we regressed the actual and synthetic job creation data on a constant and time trend, deliberately suppressing all of the other effects, including seasonal effects. The trend coefficient estimated from the actual data is -0.62 (± 0.25). In the synthetic data, the average trend coefficient is -1.17 (± 0.21). This result suggests that a more sophisticated modeling of the time effects might be of interest, although there is still substantial overlap in the 95% confidence intervals for the trend coefficient.

C. Model Protection

We protected the model and its resulting synthetic data by shrinking the variance components toward zero. Tables II and III demonstrate the trade-off between privacy and model fit for protective prior distributions. We set $V = .001$ and varied our degree of belief hyper-parameter ν. All prior distributions on the variances of the random effects were set equal. For PoRI, we considered a 10% range around each y_i as a sensitive range. PoRI was approximated by generating 1,000 synthetic values of y_i for each observation and summing an indicator variable on the sensitive range of y_i.

Results indicate that as the priors become more precise, and therefore shrank the released data more towards the global mean, ϵ and the median PoRI decreased. The results for $V = 0.001$ and $\nu = 8,000$ showing ranges of ϵ and the median PoRI for four different models appear in Table III.

Choosing a level of protection always involves trade-offs between the privacy parameters and the goodness of fit of the released model. If the agency chose to release the best-fitting model based on prior P_1, then the released synthetic data would have goodness of fit shown in Row (1) of Table II and the privacy properties shown in Column (2) of Table III. Many agencies might consider ϵ in the range $(2.1, 2.8)$ and PoRI in the range $(14\%, 15\%)$ to be acceptable. The decision to truncate would not very much affect the data quality.

Protective Priors on the Variance Components of the Random Effects

V	ν	DIC	Conditional Fit	Truncated Correlation	Synthetic Correlation	Median PoRI
1.000	10	369,350	82.5%	82.9%	78.1%	14%
.0010	2000	370,494	51.0%	79.7%	50.8%	12%
.0010	4000	372,809	18.8%	77.0%	18.8%	4%
.0010	6000	373,030	13.0%	76.5%	13.0%	4%
.0010	8000	373,193	8.1%	76.2%	8.1%	4%
.0010	10000	373,123	6.6%	75.8%	6.5%	3%
.0001	3000	376,868	3.1%	67.6%	3.1%	0%

Notes: Row (1) refers to the best fitting model from Table I. Rows (2)-(7) refer to the same statistical model with successively tighter prior distributions.

TABLE II

Privacy Measures for Baseline and Best Models for QWI

Measure/Model	Best Model with P_1	Best Model with P_5
ϵ	$(2.1, 2.8)$	$(0.72, 0.83)$
Median PoRI	$(14\%, 15\%)$	$(3.9\%, 4.0\%)$
Average DIC	$369,350$	$373,193$

Notes: The best model with P_1 is Row (1) of Table II. The best model with P_5 is Row (5) of Table II. The influential observation determining the bound on ϵ is row 41,399 in the archival data set.

TABLE III

If, on the other hand, the agency wanted tighter privacy protection, then it could use the following strategy. Using the best fitting Bayesian ZIP mixed model with prior P_5 $(V = .001, \nu = 8,000)$, we were able to increase privacy compared to the same model with a non-informative prior. This can be seen by comparing Row (5) of Table II to Row (1). However, degradation in model fit appeared severe, as can be seen in the same comparison of rows. The degradation in fit was due to the 12 observations that had actual values of y_i greater than 8,000 job creations. The strength of the prior required to mitigate the effects of these observations pulls all of the variances of the random effects towards zero with a high prior degree of belief (controlled by the ν hyper-parameters). Table II shows that after truncating these observations, the correlation is over 76% when $V = 0.001, \nu = 8,000$. Compared to the model that only included counties as random effects (Row (1) of Table I which had an ϵ of 1.7, a median PoRI of 5%, and a truncated correlation of 75%), we were able to increase privacy and maintain about the same fit quality. Overall, our proposed methodology greatly increased privacy while maintaining a decent fit after excluding the 12 highest observations (less than 0.03% of observations).

V. Conclusion

We implemented a protection strategy based on Zero-inflated Poisson extensions of Bayesian Generalized Linear Mixed Models with privacy-preserving prior distributions to produce synthetic data for related time series from thousands of small groups. Our method demonstrated the trade-off between producing synthetic data with analytical features close to the original data (effectively the synthetic data that would be produced from the best Bayesian analysis of the actual confidential data) and synthetic data produced using privacy-preserving prior distributions. We applied our methodology to the strict privacy measure of empirical differential privacy and a newly defined privacy measure, PORI, which controls the attribute disclosure risk of each observation in the data set. We found that as the prior distributions of the variance components become more precise toward zero, privacy increased on all accounts as compared to a model with a non-informative prior. As an alternative to current rule-based procedures, which include the suppression of data, our research allows agencies to use disclosure avoidance procedures to release their detailed data using a protection method incorporated into the estimation of a formal probability model. Agencies could, therefore, balance "the level of protection provided and the effects on the ability of users to draw valid inferences" (Duncan et al. 1993) by adjusting their preference for privacy within a prior distribution.

VI. References

Abowd, J., Gittings, R. K., McKinney, K., Stephens, B., Vilhuber, L., & Woodcock, S. (2012) "Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time series." Federal Committee on Statistical Methodology, Office of Management and Budget, 2012 Research Conference Papers, available at fcsm.sites.usa.gov/files/2014/05/Vilhuber_2012FCSM_VIII-C.pdf (cited September 7, 2014).

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., & Woodcock, S. (2009) "The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators." In *Producer Dynamics: New Evidence from Micro Data* (Chicago: University of Chicago Press for the National Bureau of Economic Research) pp. 149-230, available at www.nber.org/chapters/c0485.pdf (cited September 7, 2014).

Abowd, J. M., & Schneider, M. J. (2011) "An Application of Differentially Private Linear Mixed Modeling," in Data Mining Workshops (ICDMW) 2011, IEEE 11th International Conference on Data Mining (December): 614-619, DOI:10.1109/ICDMW.2011.26.

Abowd, J. M., Schneider, M. J., & Vilhuber, L. (2013) "Differential Privacy Applications to Bayesian and Linear Mixed Model Estimation." *Journal of Privacy and Confidentiality*, Vol. 5: 1, Article 4, available at repository.cmu.edu/jpc/vol5/iss1/4/ (cited September 7, 2014).

Drechsler, J., & Reiter, J. P. (2010) "Sampling with synthesis: a new approach for releasing public use census microdata." *Journal of the American Statistical Association*, 105(492), 1347-1357.

Duncan, G. T., Jabine, T. B., de Wolf, V. A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Research Council (Washington, DC: National Academies Press), pp. 288, available at http://www.nap.edu/catalog/2122.html (cited September 7, 2014).

Duncan, G. T., Elliot, M. J. & Salazar-Gonzalez, J-J. (2011) *Statistical Confidentiality: Principles and Practice* (New York: Springer).

Dwork, C., (2006) "Differential Privacy," International Colloquium Automata, Languages and Programming 2006, Proceedings Part II, Lecture Notes in Computer Science, Volume 4052, pp. 1-12, DOI: 10.1007/11787006_1.

Federal Committee on Statistical Methodology (FCSM) (2005) "Report on Statistical Disclosure Limitation Methodology," Statistical Policy Working Paper 22, Second version, published online at http://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf (cited on September 21, 2014).

Hadfield, J. D. (2010) "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package." *Journal of Statistical Software*, 33(2), 1-22.

Hall, D. B. (2000) "Zero-inflated Poisson and binomial regression with random effects: a case study," *Biometrics*, 56(4), 1030-1039.

Hotz, V. J., Goerge, R., Balzekas, J., & Margolin, F. (1998) "Administrative data for policy-relevant research: Assessment of current utility and recommendations for development," *Report of the Advisory Panel on Research Uses of Administrative Data*, Chicago: Northwestern University/University of Chicago Joint Center for Poverty Research.

Kohnen C. N. and Reiter, J. P. (2009) "Multiple imputation for combining confidential data owned by two agencies," *Journal of the Royal Statistical Society, Series A,* 172, 511 - 528.

Little, R. J. A. (1993) "Statistical Analysis of Masked Data," *Journal of Official Statistics*, Vol. 9, No. 2, 407-426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008) "Privacy: Theory meets practice on the map." In IEEE 24th International Conference on Data Engineering, ICDE 2008 (April):

277-286).

OECD (2012) “Public-use Files: Practices and Methods to Increase Quality of Released Microdata,” Report of the Expert Group for International Collaboration on Microdata Access, Statistics Directorate.

Reiter, J. P. (2005a) “Estimating risks of identification disclosure in microdata,” *Journal of the American Statistical Association*, 100(472).

Reiter, J. P. (2005b) “Releasing multiply-imputed, synthetic public-use micro-data: An illustration and empirical study,” *Journal of the Royal Statistical Society, Series A*, 168, pp. 185-205.

Rubin, D. B. (1993) “Statistical disclosure limitation.” *Journal of Official Statistics*, 9(2), 461-468.