NBER WORKING PAPER SERIES

A FRAMEWORK FOR SHARING CONFIDENTIAL RESEARCH DATA, APPLIED TO INVESTIGATING DIFFERENTIAL PAY BY RACE IN THE U. S. GOVERNMENT

Andrés F. Barrientos
Alexander Bolton
Tom Balmat
Jerome P. Reiter
John M. de Figueiredo
Ashwin Machanavajjhala
Yan Chen
Charles Kneifel
Mark DeLong

Working Paper 23534
http://www.nber.org/papers/w23534

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2017

This research was supported by NSF grants ACI-14-43014 and SES-11-31897, as well as the Alfred P. Sloan Foundation grant G-2-15-20166003. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Framework for Sharing Confidential Research Data, Applied to Investigating Differential Pay by Race in the U. S. Government

Andrés F. Barrientos, Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charles Kneifel, and Mark DeLong

NBER Working Paper No. 23534
June 2017
JEL No. C51,C53,C55,C81,J15,J45

ABSTRACT

Data stewards seeking to provide access to large-scale social science data face a difficult challenge. They have to share data in ways that protect privacy and confidentiality, are informative for many analyses and purposes, and are relatively straightforward to use by data analysts. We present a framework for addressing this challenge. The framework uses an integrated system that includes fully synthetic data intended for wide access, coupled with means for approved users to access the confidential data via secure remote access solutions, glued together by verification servers that allow users to assess the quality of their analyses with the synthetic data. We apply this framework to data on the careers of employees of the U. S. federal government, studying differentials in pay by race. The integrated system performs as intended, allowing users to explore the synthetic data for potential pay differentials and learn through verifications which ″findings in the synthetic data hold up in the confidential data and which do not. We find differentials across races; for example, the gap between black and white female federal employees' pay increased over the time period. We present models for generating synthetic careers and differentially private algorithms for verification of regression results.

Andrés F. Barrientos
Department of Statistical Science
Duke University, Box 90251
Durham, NC 27708
afb26@stat.duke.edu

Alexander Bolton
Emory University
201 Dowman Drive
Atlanta, GA 30322
alexander.bolton@duke.edu

Tom Balmat
Social Science Research Institute
Gross Hall 235
140 Science Drive
Durham, NC 27708
thomas.balmat@duke.edu

Jerome P. Reiter
Department of Statistical Science
Duke University, Box 90251
Durham, NC 27708
jerry@stat.duke.edu

John M. de Figueiredo
The Law School and Fuqua
School Duke University
210 Science Drive, Box 90360
Durham, NC 27708
and NBER
jdefig@duke.edu

Ashwin Machanavajjhala
Department of Computer Science
Duke University, Box 90129
Durham NC 27709
ashwin@cs.duke.edu

Yan Chen
Department of Computer Science
Duke University, Box 90129
Durham NC 27709
yanchen@cs.duke.edu

Charles Kneifel
Duke University
Office of Information Technology
334 Blackwell Street
Durham, NC 27701
Charley.Kneifel@duke.edu

Mark DeLong
Duke University
Office of Information Technology
334 Blackwell Street
Durham, NC 27701
mark.delong@duke.edu

Supplemental material is available at
http://www.nber.org/data-appendix/w23534

A Framework for Sharing Confidential Research Data, Applied to Investigating Differential Pay by Race in the U. S. Government

Andrés F. Barrientos, Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo
Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, Mark DeLong*

Abstract

Data stewards seeking to provide access to large-scale social science data face a difficult challenge. They have to share data in ways that protect privacy and confidentiality, are informative for many analyses and purposes, and are relatively straightforward to use by data analysts. We present a framework for addressing this challenge. The framework uses an integrated system that includes fully synthetic data intended for wide access, coupled with means for approved users to access the confidential data via secure remote access solutions, glued together by verification servers that allow users to assess the quality of their analyses with the synthetic data. We apply this framework to data on the careers of employees of the U. S. federal government, studying differentials in pay by race. The integrated system performs as intended, allowing users to explore the synthetic data for potential pay differentials and learn through verifications which findings in the synthetic data hold up in the confidential data and which do not. We find differentials across races; for example, the gap between black and white female federal employees' pay increased over the time period. We present models for generating synthetic careers and differentially private algorithms for verification of regression results.

*Andrés F. Barrientos is Postdoctoral Associate, Department of Statistical Science, Duke University, Durham, NC 27708 (email: afb26@stat.duke.edu); Alexander Bolton is Assistant Professor, Department of Political Science, Emory University, Atlanta, GA 30322 (abolton@emory.edu); Tom Balmat is Statistician, Social Science Research Institute, Duke University, Durham, NC 27708 (thomas.balmat@duke.edu); Jerome Reiter is Professor, Department of Statistical Science, Duke University, Durham, NC 27708 (jerry@stat.duke.edu); John de Figueiredo is Edward and Ellen Marie Schwarzman Professor, Law School and Fuqua School of Business, Duke University, Durham NC 27708 (jdefig@duke.edu); Ashwin Machanavajjhala is Assistant Professor, Department of Computer Science, Duke University, Durham, NC 27708 (ashwin@cs.duke.edu); Yan Chen is Graduate Student, Department of Computer Science, Duke University, Durham, NC 27708 (yanchen@cs.duke.edu); Charley Kneifel is Senior Technical Director, Office of Information Technology, Duke University, Durham, NC 27701 (charley.kneifel@duke.edu); and, Mark DeLong is Director of Research Computing, Office of Information Technology, Duke University, Durham, NC 27701 (mark.delong@duke.edu). This research was supported by NSF grants ACI-14-43014 and SES-11-31897, as well as the Alfred P. Sloan Foundation grant G-2-15-20166003.

Key Words: Access, Disclosure, Privacy, Synthetic, Verification.

1 Introduction

Widespread access to large-scale social science datasets greatly enhances the work of evidence-based policy makers, social scientists, and statisticians. Yet, widespread dissemination of large scale social science data also carries a significant social cost: it puts data subjects' privacy and confidentiality at risk. Simply stripping unique identifiers like names and exact addresses, while necessary, generally does not suffice to protect confidentiality. As is well documented (e.g., Sweeney, 1997, 2013; Narayanan and Shmatikov, 2008; Parry, 2011), ill-intentioned users may be able to link records in the released data to records in external files by matching on variables common to both sources. These threats are particularly serious for large-scale social science data. Such data often come from administrative or privately collected sources so that, by definition, someone other than the organization charged with sharing the data knows the identities and (a large number of) attributes of data subjects. Large-scale social science data also typically include many variables that, since the data arguably are known by others, could serve as matching variables.

As the size, richness, and quality of social science data have increased, so too have the threats to confidentiality. Confronted with these risks, responsible data stewards face a difficult dilemma: how can they provide access to confidential social science data while protecting confidentiality of data subjects' identities and sensitive attributes? Often data stewards—whether in academia, government, or industry—default to restricting access to carefully vetted and approved researchers via licensing arrangements or physical data enclaves. This is only a partial solution. It denies the benefits of data access to broad subsets of society including, for example, students who need data for learning the skills of data analysis and citizen scientists seeking to understand their society.

In this article, we describe and illustrate a general framework for providing access to large-scale social science data. The framework integrates three key ideas from the literature on confidentiality protection and data access. The first idea is to provide synthetic public use files, as proposed by Rubin (1993) and others (e.g., Little, 1993; Fienberg, 1994; Raghunathan *et al.*, 2003; Reiter, 2005a; Reiter and Raghunathan, 2007; Drechsler, 2011; Callier, 2015). Such files comprise individual records with every value replaced with simulated draws from an estimate of the multivariate

distribution of the confidential data. When generated appropriately, synthetic data can preserve many, but certainly not all, important associations in the confidential data. They also should carry low disclosure risks, since the released data are not actual records. This largely eliminates the kinds of record linkage attacks that have broken typical disclosure control methods, as it is nonsensical for ill-intentioned users to match synthetic records to external files.

While synthetic data have been used to release public use versions of several high profile social science datasets (Abowd *et al.*, 2006; Hawala, 2008; Machanavajjhala *et al.*, 2008; Drechsler *et al.*, 2008; Kinney *et al.*, 2011), at present they have a critical weakness. Users of synthetic data cannot determine how much their analysis results have been impacted by the synthesis process. This limitation leads to the second idea that we integrate in the framework: provide users access to verification servers (Reiter *et al.*, 2009). A verification server is a query-based system that (i) receives from the user a statistical query that enables comparison of results from the synthetic and confidential data, and (ii) returns an answer to the query without allowing the user to view the confidential data directly (Karr and Reiter, 2014). With the output from a verification server, users can decide whether or not results based on the synthetic data are of satisfactory quality for their particular purposes.

Undoubtedly, some analyses will not be adequately preserved. The verification server will help users learn this, thereby reducing the chances of false findings based on the synthetic data. These users may desire access to the confidential data, which motivates the third prong of our integrated data access system: provide remote access to confidential data to approved users via virtual machines on a protected research data network (PRDN). We do not describe how to set up the architecture for a PRDN in this article, although we use one in our illustrative application.

Integrating all three ideas in a single system creates synergies. Users start with the synthetic data, which has low barriers to access. They can use the synthetic data to investigate distributions and relationships, determine what questions might be answerable with the data (e.g., are there enough cases of interest to support accurate modeling?), examine the need for transformations and recoded variables, and develop appropriate code. The verification server can enable researchers and policy makers to know when to trust and act on their results, and when perhaps not to do so. Even users who are not satisfied with the quality of the results can benefit from starting with the synthetic data. Storage and processing of large-scale data are costly to data stewards, who likely

will pass some costs to users. Analysts who have an informed analysis plan—for example, they know the approximate marginal distributions of the data and have a sense of the data structure—can improve their efficiency when using the server, thereby saving their own time and money. Crucially, by performing their data explorations outside the PRDN, these analysts will use up fewer cycles on the protected systems and open opportunities for more efficient use of those systems. Finally, with a system that allows analysts to make informed decisions and a convenient means to access confidential data remotely, data stewards can facilitate the highest quality analyses to trusted individuals in ways that are (i) more cost-effective to establish and sustain, and more readily modified and updated to incorporate new technologies, than physical enclaves, and (ii) more secure than distributing licensed datasets to researchers for use on their own machines.

We apply and illustrate the framework on data comprising the workforce of the United States federal government from 1988 to 2011. Specifically, we generate an entirely synthetic federal workforce, including new methodology for modeling and generating career trajectories. Using the synthetic data, we estimate regression models that assess systematic differences in employee salaries by race and gender, and investigate how such differences change over time. We present several new verification measures that satisfy differential privacy (Dwork, 2006), a criterion with strong and formal guarantees of data privacy. We apply these measures to the synthetic data results, and ultimately verify the final models using the confidential data inside a PRDN at Duke University (`https://oit.duke.edu/what-we-do/services/protected-network`). The findings are remarkable for both methodological and substantive reasons. For the former, the integrated system performs as advertised, allowing us to see the validity, and shortcomings, of the synthetic data results. For the latter, the confidential data suggest that (i) the differential in pay for white and black female employees has been increasing over time, and that (ii) Asian male employees make substantially less on average than white male employees over much of the time frame we analyzed. As far as we know, neither of these findings have been previously documented in the public sector at this magnitude or detail.

The remainder of this article is organized as follows. In Section 2, we describe the data in more detail and outline the procedures used to generate the synthetic data, including the approach for generating synthetic careers. In Section 3, we describe the verification measures. In Section 4, we mimic a usage of the integrated system to analyze the differential pay gap: we start with the

synthetic data, verify findings using the differentially private measures, and repeat the analysis on the confidential data. In Section 5 we discuss some issues on implementation of this framework.

2 Description of Data

2.1 Overview of the confidential data

The Office of Personnel Management (OPM) maintains the personnel records for all civil servants in the United States. We work with a subset of the data from the OPM's Central Personnel Data File (CPDF) and Enterprise Human Resources Integration system (EHRI), which we jointly refer to as the Status File (SF). The SF we use is a snapshot of the civil service on every September 30 (the end of the fiscal year), comprising approximately 2 million employees per year from 1988 to 2011. For each employee, the file includes annual data on characteristics like age, agency, education level, pay grade, occupation, supervisory status, entry and departure, and other background characteristics. The data are longitudinally linked. We exclude employees from the armed services, the Department of Defense, the U. S. Postal Service, and individuals who work in classified roles, sensitive agencies, and sensitive occupations as defined by OPM. The final analysis file includes personnel records from 3,511,824 employees.

The OPM data are valuable because they allow researchers to investigate many key questions in the study of human capital in large organizations and government organizations in particular. For example, researchers can use the OPM data to examine government agencies' ability to recruit high quality individual talent, to develop their employees' expertise within those agencies, and to retain the best and brightest in government service (e.g., Lewis and Durst, 1995; Bolton and de Figueiredo, 2016). These are important and complicated challenges; public agencies must cope with episodic turnover of political appointees, limited ability to adjust worker compensation in response to outside market pressures, difficulty in performance measurement due to the nature of governmental tasks, and constraints on frictionless alterations to the government workforce because of employment terms for civil servants (Borjas, 1980; Bolton *et al.*, 2016). Ultimately, research with these data can shed light on the relative costs and benefits of human capital management strategies.

The OPM data files are currently housed in a PRDN at Duke University with strict access and confidentiality standards under an IRB-approved protocol. The data have not been released

by OPM to the general public in as comprehensive form as we propose, primarily because of confidentiality concerns. OPM is required by regulation to disclose the name, agency, location, position, grade, and salary of each civil service employee upon request. In fact, it produces annually an identifiable list of all federal employees with these six fields. This poses a disclosure risk problem. If the SF data were "anonymized" by standard techniques, an ill-intentioned user might be able to reverse engineer a large percentage of the OPM database by matching the six known fields to the same fields in the anonymous data. They subsequently could retrieve the private and confidential data of a large percentage of the federal personnel.

2.2 Overview of synthetic data creation

To reduce these re-identification risks, the OPM could create a fully synthetic version of the SF, as we do for this article. In this section, we provide an overview of our process for generating the synthetic data, mainly to provide context for the framework of synthetic data coupled with verification. Our emphasis for this article is the big picture—the concept of an integrated system for data access—rather than the specifics of creating or evaluating synthetic data for this file. We do include more details on the methods for generating synthetic careers, races, and wages in this section; these variables are central to our illustrative analysis of wage differentials. The synthesis models for other variables are described in the online supplementary material. We do not discuss additional evaluations of the usefulness or disclosure risks in the synthetic data, as these are ancillary to the objectives of this article. We note that, as of this writing, the OPM has not yet determined whether or not to make the synthetic data available to a broader set of researchers.

The SF data are complicated, making the task of generating useful synthetic data challenging. The employees are measured on 29 variables over the course of 24 years. They work across 607 agencies, some of which have only a handful of employees and some of which have thousands of employees. The variables are mostly nominal with levels ranging from 2 (sex) to 803 (occupation), and also include a small number of numerical variables. For any employee, most variables can change annually, although a few demographic variables remain constant or change deterministically (age). Many pairs of variables have theoretically impossible combinations, such as certain occupations being restricted to certain education and degree types. Some variables should be non-decreasing over time, such as months of military service and educational levels, although the confidential data

have records that violate those restrictions, presumably due to reporting errors.

Broadly, we use sequential conditional modeling to make the synthetic data, as done in Kinney *et al.* (2011). We order the variables from the first to last to be synthesized. Let V_{ij} be the value of the j-th ordered variable for the i-th employee, where $j = 1, \ldots, 29$. We seek the joint distribution,

$$p(V_{i1}, \ldots, V_{i29}) = p_1(V_{i1}) \times p_2(V_{i2}|V_{i1}) \times \ldots \times p_{29}(V_{i29}|V_{i1}, \ldots, V_{i28}), \tag{1}$$

where each p_j denotes the conditional distribution of V_j given $V_1, \ldots, V_{j-1}$. We let V_1 correspond to the sequence of agencies where the employee has worked, which defines the employee's career. Nearly all other variables depend on when and where the employee works, so modeling this variable first facilitates the synthesis process. We let $(V_2, \ldots, V_7)$ be, in order, gender, race, educational level, age in years, years since the employee earned the degree mentioned in educational level, and an indicator for ever having served in the military. These demographic variables are, for the most part, straightforward to model because either (1) they remain constant across time or change in a deterministic manner after the initial year, or (2) change with only low probabilities. We let $(V_8, \ldots, V_{29})$ include the remaining variables, which depend on the characteristics of the employee's job that year. Examples of these variables include occupation, part-time or full-time status, grade and step classification, supervisory status, and pay. A full list of variables is in the online supplementary material.

For V_j that can change annually, where $j > 1$, we generally apply lag-one modeling strategies to simplify computation. Specifically, let V_{ijt} be the j-th variable at year t for the i-th employee. Let $t_{i1} < \ldots < t_{in_i}$ be the n_i years when employee i has values (is working), and set $V_{ij} = (V_{ijt_{i1}}, \ldots, V_{ijt_{in_i}})$. For longitudinal V_j, we use the conditional representation,

$$p_j\left(V_{ij}\,|V_{i1}, \ldots, V_{ij-1}\right) = \prod_{l=2}^{n_i} p_{j,t_{il}}\left(V_{ijt_{il}}\,\middle|V_{i1}, \ldots, V_{ij-1}, V_{ijt_{i1}}, \ldots, V_{ijt_{il-1}}\right), \tag{2}$$

where $p_{j,t_{il}}$ denotes the distribution of $V_{ijt_{il}}$ conditioned on the previous $j-1$ variables and the values of V_{ij} up to time t_{il-1}. We assume that

$$p_{j,t_{il}}\left(V_{j,t_{il}}\,\middle|V_{i1}, \ldots, V_{ij-1}, V_{j,t_{i1}}, \ldots, V_{ijt_{il-1}}\right) = p_{j,t_{il}}\left(V_{j,t_{il}}\,\middle|V_{i1t_{il}}, \ldots, V_{ij-1,t_{il}}, V_{ijt_{il-1}}\right). \tag{3}$$

Employee	Employee's career										G	Z	W
e_1	0	0	A	A	0	0	C	C	C	C	3	(3,5,7)	(0,A,0,C)
e_2	0	0	0	0	0	0	0	0	0	B	2	10	(0,B)
e_3	A	0	B	C	C	A	A	A	0	0	4	(2,3,4,6,9)	(A,0,B,C,A,0)

Table 1: Illustration of how to define (G, Z, W) using three hypothetical employees and 10 years. Each column in the employee's career represents a year; a 0 means the employee did not work that year; and, A, B, and C represent three different agencies. For example, employee e_1 did not work in years 1 and 2, worked in agency A for two years, stopped working in years 5 and 6, and worked in agency C during the last four years.

Thus, the conditional distribution of $V_{ijt_{il}}$ depends only on current values of $V_{ij'}$, $1 < j' < j - 1$, and the nearest past value of V_{ij}.

2.2.1 Modeling strategy for employees' careers (V_1)

We define an employee's career as the sequence of agencies where the employee has worked throughout the 24 years. Since most employees have not worked in all 24 years, we create an additional "agency" corresponding to the status of not working. With this additional level, all employees' sequences have length 24.

To model these sequences, we create three additional variables. Let G be the number of agencies where the employee has worked over the 24 years. Let Z be the list of years in which the employee moved to a new agency, including a change in working status. Let W be the ordered sequence of unique agencies where the employee has worked. The values of (G, Z, W) completely describe the entire career of any employee, as illustrated in Table 1.

Defining a model for employees' careers is equivalent to defining a model for (G, Z, W), which we do sequentially. For G, we use a discrete distribution on $\{1, \ldots, 24\}$ with probabilities equal to the observed frequencies of each value, from which we randomly generate the values of G associated with the synthetic employees. For $Z|G$, we create a one-to-one function T_G to map Z into a space of permutations dependent on G. We model $T_G(Z)|G$ using a latent model defined on the simplex space. The latent model is defined using mixtures of Dirichlet distributions. This model allows us to borrow information across different agency patterns (given G) and, therefore, to give positive probability to unobserved values of $Z|G$. Since we use a one-to-one mapping, the model for $T_G(Z)|G$ can be easily used to generate values from $Z|G$. Finally, we model $W|Z, G$ using a Markov chain

of order one. A formal description of the three sub-models is in the online supplementary material. While targeted at modeling careers, this model can be applied more generally for other sequences of categorical variables.

2.2.2 General strategy for race (V_3)

Almost all employees report the same value of race in all 24 years. However, 2.7% of employees report different values across the years, usually changing values only once or twice. Rather than model longitudinal changes in race across time for all employees, which easily could result in far more switching than observed in the data, we instead create an auxiliary binary variable that indicates whether the values of race remain the same across all years or not. We estimate a model for this binary outcome using classification and regression trees (CART), conditioning on sex and predictors derived from the employee's career. We use this model to generate synthetic values of this variable (Reiter, 2005b). For employees whose race values do not change, we estimate a predictive model for their first observed race using CART, again conditioning on (V_1, V_2), and synthesize V_3 based on the synthetic values of careers and gender. Finally, for employees whose values change across time, we model the race at each year using (2) and (3), using CART for each $p_{3,t}$ where $t = 1, \ldots, 24$.

2.2.3 General strategy for wages (V_{27})

Federal employees' basic pay (salary before any location adjustments) is set by tables known as pay plans. For example, most government employees in white collar occupations fall under the General Schedule pay plan. For most pay plans, basic pay is a deterministic function of a combination of variables, usually including the employee's grade and step. Thus, in theory, you can find any federal employee's pay by locating their grade and step on their pay plan table. However, in the SF, some employees' basic pay is not consistent with their pay plan, grade, and step; when this happens, usually the pay coincides with a value in the pay table associated with a neighboring step.

To capture these features, we synthesize pay plan, grade, and step before basic pay, thereby allowing us to "look up" the pay for the synthetic employees. We model basic pay using (2) and (3) with a CART synthesizer, assuming that basic pay is a nominal variable. This essentially is equivalent to sampling from the values of basic pay reported in grade and step for a given pay plan,

but also allowing for other variables on the file to explain deviations from the pay plan.

We note that the government has been releasing employees' basic pay to the public in recent years (see `https://www.fedsdatacenter.com/federal-pay-rates/`), so that the values of basic pay that occur in any year can be considered publicly available information.

3 Verification Measures

The quality of inferences from synthetic data depend entirely on the quality of the models used to generate the data, as the synthetic data only can reflect distributional assumptions in the synthesis models (Reiter, 2005a). Analysts of the synthetic data need some way to assess the accuracy of their particular inferences based on the synthetic data. Verification servers provide an automated means to provide such feedback.

In designing a verification server, the agency must account for a crucial fact: verification measures leak information about the confidential data (Reiter *et al.*, 2009; McClure and Reiter, 2012). Clever data snoopers could submit queries that, perhaps in combination with other information, allow them to estimate confidential values too accurately. To reduce and quantify these risks, one approach is to require verification measures to satisfy ϵ-differential privacy (ϵ-DP), which we now explain briefly.

Let $\mathcal{A}$ be an algorithm that takes as input a database $\mathbf{D}$ and outputs some quantity o, i.e., $\mathcal{A}(\mathbf{D}) = o$. In our context, these outputs are used to form verification measures. Define neighboring databases, $\mathbf{D}$ and $\mathbf{D}'$, as databases that differ in one row and are identical for all other rows.

Definition 1 (ϵ-differential privacy.) *An algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if for any pair of neighboring databases $(\mathbf{D}, \mathbf{D}')$, and any output $o \in range(\mathcal{A})$, the $Pr(\mathcal{A}(\mathbf{D}) = o) \leq \exp(\epsilon) Pr(\mathcal{A}(\mathbf{D}') = o)$.*

Intuitively, $\mathcal{A}$ satisfies ϵ-DP when the distributions of its outputs are similar for any two neighboring databases, where similarity is defined by the factor $\exp(\epsilon)$. The ϵ, also known as the privacy budget, controls the degree of the privacy offered by $\mathcal{A}$, with lower values implying greater privacy guarantees. ϵ-DP is a strong criterion, since even an intruder who has accss to all of $\mathbf{D}$ except any one row learns little from $\mathcal{A}(\mathbf{D})$ about the values in that unknown row when ϵ is small.

Differential privacy has three other properties that are appealing for verification measures. Let $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$ be ϵ_1-DP and ϵ_2-DP algorithms. First, for any database $\mathbf{D}$, releasing the outputs of both $\mathcal{A}_1(\mathbf{D})$ and $\mathcal{A}_2(\mathbf{D})$ ensures $(\epsilon_1 + \epsilon_2)$-DP. Thus, we can quantify and track the total privacy leakage from releasing verification measures. Second, releasing the outputs of both $\mathcal{A}_1(\mathbf{D}_1)$ and $\mathcal{A}_2(\mathbf{D}_2)$, where $\mathbf{D}_1 \cap \mathbf{D}_2 = \emptyset$, satisfies $\max\{\epsilon_1, \epsilon_2\}$-DP. Third, for any algorithm $\mathcal{A}_3(\cdot)$, releasing $\mathcal{A}_3(\mathcal{A}_1(\mathbf{D}))$ for any $\mathbf{D}$ still ensures ϵ_1-DP. Thus, post-processing the output of ϵ-DP algorithms does not incur extra loss of privacy.

A common method for ensuring ϵ-DP, which we utilize for ϵ-DP verification measures, is the Laplace Mechanism (Dwork, 2006). For any function $f : \mathbf{D} \to \mathbb{R}^d$, let $\Delta(f) = \max_{(\mathbf{D_1}, \mathbf{D_2})} ||f(\mathbf{D}_1) - f(\mathbf{D}_2)||_1$, where $(\mathbf{D}_1, \mathbf{D}_2)$ are neighboring databases. This quantity, known as the global sensitivity of f, is the maximum L_1 distance of the outputs of the function f between any two neighboring databases. The Laplace Mechanism is $\mathbf{LM}(\mathbf{D}) = f(\mathbf{D}) + \eta$, where η is a $d \times 1$ vector of independent draws from a Laplace distribution with density $p(x \mid \lambda) = (1/(2\lambda)) \exp(-|x|/\lambda)$, where $\lambda = \Delta(f)/\epsilon$.

We now present verification measures that satisfy ϵ-DP and help analysts assess the importance of regression coefficients. We derive the measures for linear regression, as we use these models in the analysis of wage differentials by race. To fix notation for describing the measures, let $\mathbf{D}$ include all n individuals in the subset of the confidential data that is of interest for analysis. For any individual i belonging to $\mathbf{D}$, let $y_i \in \mathbb{R}$ be the response variable and $x_i = (1, x_{i,1}, \ldots, x_{i,p})^T \in \mathbb{R}^{p+1}$ be a set of predictors, where both are transformed as desired for regression modeling. Hence, $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^n$, where all values are from the confidential data. Let $E(y_i \mid x_i) = \beta^T x_i$, where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$.

3.1 Measures for importance of regression coefficients

In many contexts, analysts are interested in whether or not the value of some β_j exceeds some threshold, say γ_0. For example, users of the SF data (economists, policy makers, lawyers) might consider a value of β_j corresponding to 1% or larger differential in average pay to be practically significant evidence of wage discrimination. But, they might be less concerned when β_j corresponds to a differential less than 1%. Without loss of generality, assume that we want to determine if some $\beta_j < \gamma_0$. Corresponding to this decision, we define the parameter $\theta_0 = \mathbb{I}_{(-\infty,\gamma_0]}(\beta_j)$, where $\mathbb{I}_{(-\infty,\gamma_0]}(\beta_j)$ is an indicator function that equals one when $\beta_j \in (-\infty, \gamma_0]$ and equals zero otherwise.

We note that the measure can be used for any interval, for example, of the form $[l, u]$ or $(u, \infty]$.

Because of the large sample sizes in the SF data, confidence intervals for all β_j of practical interest, including those for the race dummy variables, are extremely narrow. For these β_j and most γ_0, the MLE of β_j effectively tells the analyst whether $\theta_0 = 1$ or $\theta_0 = 0$. To formalize this notion, let $\hat{\beta}_j^N$ be the MLE of β_j based on a sample with N individuals (where N stands for a generic sample size). We approximate θ_0 by using the pseudo-parameter,

$$\theta_N = \begin{cases} 1 & \text{if } P[\hat{\beta}_j^N \leq \gamma_0] \geq \gamma_1, \\ 0 & \text{if } P[\hat{\beta}_j^N \leq \gamma_0] < \gamma_1. \end{cases}$$

Here, $\gamma_1 \in (0, 1)$ reflects the degree of certainty required by the user before she decides there is enough evidence to conclude that $\theta_0 = 1$. When $\hat{\beta}_j^N$ is a consistent estimator of β_j, we can guarantee that $\lim_{N\to\infty} \theta_N = \theta_0$.

Unfortunately, we cannot release $\hat{\beta}_j^N$, nor other deterministic functions of $\mathbf{D}$, directly and satisfy ϵ-DP. Instead, we release a noisy version of the key quantity in θ_N, namely $r = P(\hat{\beta}_j^N \leq \gamma_0)$. We do so using the sub-sample and aggregate method (Nissim *et al.*, 2007). We randomly split $\mathbf{D}$ into M disjoint subsets, $\mathbf{D}_1, \ldots, \mathbf{D}_M$, of size N (with inconsequential differences when $N = n/M$ is not an integer), where M is selected by the user. We discuss the choice of M in Section 5. In each $\mathbf{D}_l$, where $l = 1, \ldots, M$, we compute the MLE b_{jl} of β_j. The $(b_{j1}, \ldots, b_{jM})$ can be treated as M independent draws from the distribution of $\hat{\beta}_j^N$, where $N = n/M$. Let $W_l = \mathbb{I}_{(-\infty,\gamma_0]}(b_{jl})$. Each W_l is an independent, Bernoulli distributed random variable with parameter r. Thus, inferences for r can be made based on $S = \sum_{l=1}^{M} W_l$. However, we cannot release S directly and satisfy ϵ-DP; instead, we generate a noisy version of S using the Laplace Mechanism with $\lambda = 1/\epsilon$, resulting in $S^R = S + \eta$. The global sensitivity equals 1, since at most one of the partitions can switch from zero to one (or vice versa).

The noisy S^R satisfies ϵ-DP; however, interpreting it directly can be tricky. First, S^R is not guaranteed to lie in $(0, M)$ nor even to be an integer. Second, alone S^R does not provide estimates of uncertainty about r. We therefore use a post-processing step—which has no bearing on the privacy properties of S^R—to improve interpretation. We find the posterior distribution of r conditional on S^R and using the noise distribution, which is publicly known. Using simple MCMC techniques, we

estimate the model,

$$S^R|S \sim \text{Laplace}(S, 1/\epsilon), \quad S \mid r \sim \text{Binomial}(M, r), \quad r \sim \text{Beta}(1, 1). \tag{4}$$

Here, we treat S as an unobserved random variable and average over it.

The verification server reports back the posterior distribution of r to the analyst, who can approximate θ_N for any specified γ_1 simply by finding the amount of posterior mass below γ_1. Alternatively, analysts can interpret the posterior distribution for r as a crude approximation to the Bayesian posterior probability, $\pi\left(\beta_j \leq \gamma_0 | S^R\right)$. For instance, if the posterior mode for r equals 0.87, we could say that the posterior probability that $\beta_j < \gamma_0$ is approximately equal to 0.87. We caution that this latter interpretation may not be sensible for small sample sizes.

3.2 Measures for longitudinal trends in regression coefficients

With longitudinal data, analysts often are interested in how the value of some β_j changes over time. For example, in the SF analysis, we want to know whether the racial wage gap is closing or growing as the years advance. Suppose for a moment that we knew the values of β_j for all years. One simple way to characterize the trend in β_j over time is to break the data into K consecutive periods and, in each time period, find the OLS line predicting β_j from year. The slopes of these lines pasted together represent a piece-wise approximation to the trend. Of course, we do not know the values of β_j; we must use $\mathbf{D}$ to learn about these slopes.

We use this idea to construct a verification measure for longitudinal trends in regression coefficients. Specifically, the analyst begins by selecting K periods of interest. In each period, the analyst posits some interval for the slope and requests an ϵ-DP verification of whether the values of β_j are consistent with that posited interval. For example, the analyst might split $\mathbf{D}$ in $K = 2$ consecutive intervals, and posit that the slope of β_j over the years is negative in the first period and positive in the second period. In the wage gap analysis, this would correspond to a growing wage gap in the first period, followed by a shrinking wage gap in the second period. The analyst can use the synthetic data to identify the periods of interest and set the intervals for the slopes, as we illustrate in Section 4.4. Effectively, this evaluates whether the trends in β_j estimated with the confidential data match the trends estimated with the synthetic data.

Formally, suppose that $\mathbf{D}$ can be divided into nonempty subsets, $\{\mathbf{D}^t\}_{t\in\mathcal{T}}$, where $\mathbf{D}^t$ denotes all the data points in $\mathbf{D}$ at year t, and $\mathcal{T}$ is some period of years under study. Further, suppose that for every $(y_{it}, x_{it}) \in \mathbf{D}^t$, $E(y_{it}|x_{it}) = \boldsymbol{\beta}_t^T x_{it}$, where $\boldsymbol{\beta}_t = (\beta_{0t}, \ldots, \beta_{pt})^T$ is the vector of coefficients at time t. Let $\mathcal{T}_k \subset \mathcal{T}$ be a subset of years. The analyst seeks to learn the overall trend in the values of β_{jt}, where $t \in \mathcal{T}_k$, during that time. To characterize this trend, let $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right)$ be a real-valued function that returns the slope of the OLS line passing through the points $\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}$. The analyst might be interested in, for example, whether $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right) < 0$ indicating a decreasing trend, $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right) > 0$ indicating an increasing trend, or more generally, $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right) \in C_k$ for some interval C_k, e.g., C_k is tight around zero for a flat trend. Hence, for any interval C_k, the analyst seeks to learn $\theta_0 = \mathbb{I}_{C_k}\left(m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right)\right)$, where $\mathbb{I}_{C_k}\left(m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right)\right)$ is an indicator function that equals one when $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}}\right) \in C_k$ and equals zero otherwise.

Because θ_0 is a binary parameter, we can use the methods in Section 3.1 to release an ϵ-DP version of it. Here we outline the procedure; formal details are in the online supplement. We split $\mathbf{D}$ into M partitions of employees. In each $\mathbf{D}_l^t$, we compute the MLE b_{jtl} of β_{jt}. We let $W_l = \mathbb{I}_{C_k}(m\left(\{(t, b_{jtl})\}_{t\in\mathcal{T}_k}\right)$ and $S = \sum_{l=1}^{M} W_l$. Following the logic of Section 3.1, we use (4) to get posterior inferences for $r = P[m(\{(t, \hat{\beta}_{jt}^{N_t})\}_{t\in\mathcal{T}_k}) \in C_k]$, where $\hat{\beta}_{jt}^{N_t}$ is the MLE of β_{jt} based on a sample with N_t individuals.

When trends over the entire $\mathcal{T}$ are of interest, analysts can partition $\mathcal{T}$ into K consecutive periods, $\mathcal{T}_k = \{t_{k-1}, t_{k-1} + 1, \ldots, t_k\}$, where $k = 1, \ldots, K$ and $t_0 < t_1 < \ldots < t_K$. For a given set of intervals $\{C_k\}_{k=1}^K$, the analyst can do the verification separately for each interval, and interpret the set of results. Alternatively, the analyst can perform a single verification across all intervals, setting the parameter of interest to $\theta_0 = \prod_{k=1}^{K} \mathbb{I}_{C_k}\left(m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right)\right)$. Here, θ_0 equals one when $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}}\right) \in C_k$ for every $k = 1, \ldots, K$, and equals zero otherwise. For example, to examine whether the trend of β_{jt} is decreasing during the first 9 years and is increasing during the last 15 years, the analyst would set $C_1 = (-\infty, 0)$, $C_2 = (0, \infty)$, $\mathcal{T}_1 = \{1, \ldots, 9\}$, and $\mathcal{T}_2 = \{10, \ldots, 24\}$. If the mode of the posterior probability for r equals 0.93, we say that the posterior probability that β_{jt} decreases during the first 9 years and increases over the last 15 years approximately equals 0.93.

When setting C_k to $(-\infty, 0)$ or $(0, \infty)$, i.e., simply estimating whether the slope of β_j over $\mathcal{T}_k$ is negative or positive, the posterior modes have predictable behavior. Values are close to one when the true slope has the sign implied by C_k and is far from zero; values are close to zero when the

true slope has opposite sign and is far from zero; and, values are close to .5 when the true slope itself is close to zero. This last feature arises when the slopes in the partitions bounce randomly around zero.

The analyst who requests a single verification for $\mathcal{T}$ spends only ϵ of the privacy budget. However, this analyst only can tell if the whole trend over $\mathcal{T}$ in the confidential data matches that in the synthetic data. In contrast, the analyst who requests K verifications, one for each $\mathcal{T}_k$, spends $K\epsilon$ of the budget. But, this analyst gets finer details of the trend. For this reason, when the privacy budget allows, we recommend using $K > 1$ periods for verification, as we do in the analysis of the racial wage gap, to which we now turn.

4 Wage Differentials in the Federal Government

We now illustrate how synthetic data, verification, and a PRDN could be used together to analyze pay differentials by race in the federal government. Section 4.1 provides background on estimating pay differentials. Section 4.2 introduces our general regression modeling approach for the SF data analysis. Section 4.3 describes an overall analysis of average differences in pay across races, pooling all years of data. Section 4.4 investigates the trends in pay differentials over time.

4.1 Prior research on pay differentials

Social scientists have spent decades measuring the race wage gap. Estimates based on data from the Current Population Survey, the National Longitudinal Survey of Youth, and the Panel Study of Income Dynamics, among other datasets, put the unconditional black wage gap over the past thirty years between 16% – 40%. When controlling for individual demographic and job-related variables, such as education, age, gender, occupation, and industry, estimates put the gap between 0% – 15%, depending upon the precise dataset, controls, and statistical methods used (Altonji and Blank, 1999; Cancio *et al.*, 1996; Card and Lemieux, 1994; Maxwell, 1994; McCall, 2001; Neal and Johnson, 2003; O'Neill, 1990). Estimates of the race wage gap have been declining or steady over the past few decades (Hoover *et al.*, 2015; Sakano, 2002). For example, Altonji and Blank (1999) estimate that the black wage gap for full time, year-round workers, controlling for education, experience, region, industry, and occupation remained steady at approximately 6.5% from 1979 to

1995. Other studies controlling for only age, education, and location have found the gap drop from 47% in 1940 to 18% in 2000 (Black *et al.*, 2013).

While most research on the race wage gap has focused on private sector labor markets, there is a steady literature measuring it in the public sector (Borjas, 1982, 1983; Kim, 2004; Charles, 2003; McCabe and Stream, 2000; Llorens *et al.*, 2007; Lewis and Nice, 1994), and particularly in the federal government. Lewis (1998) found that between 1976 and 1986, the conditional race wage gap for black men declined from 21.0% to 16.7%, but for black women declined only from 29.7% to 27.9%; for Hispanic men the move was from 17.9% to 13.0%, and for Hispanic women it was 27.9% to 23.3%. Lewis (1998) also found that black men with educations and work experiences comparable to those for white men encounter a wage gap of 4%. He concluded that minorities made substantial progress in closing the wage gap between 1975 and 1995, especially at the very senior levels of the government. More recent work on the U. S. federal government found the race wage gap, controlling for demographic and agency characteristics, between 1988 and 2007 closed slightly for blacks from 7.9% to 7.4%, closed for Asians from 1.5% to 0.5%, and closed for Hispanics from 4.5% to 2.8% (Government Accountability Office, 2009, 57). In the GAO report, the wage gaps were not broken out separately by sex.

4.2 General modeling approach

We estimate the race wage gap using linear regression techniques (Blau and Kahn, 2017), running the same models on both the synthetic and confidential SF datasets. The dependent variable is the natural logarithm of an employee's inflation adjusted basic pay in a given year. Basic pay is an individual's base salary and excludes any additional pay related to geographic location, award payments, or other monetary incentives paid out to employees. We exclude any observations with pay values of 0 or codes indicating the record is invalid, according to the OPM. We use all available cases (Little and Rubin, 2002) for regression modeling, as we have no reason to think values are systematically missing.

The central independent variable is the race with which individual employees identify. Prior to 2005, employees could choose to identify with 16 categories. The largest five utilized were American Indian or Alaska Native, Asian or Pacific Islander, black, Hispanic, and white. The other, substantially less-utilized categories were Asian Indian, Chinese, Filipino, Guamanian, Hawaiian,

Japanese, Korean, Samoan, Vietnamese, Other Asian/Pacific Islander, and Not Hispanic in Puerto Rico. We group these categories (save the last) with the Asian or Pacific Islander category and drop the (very small) category of Not Hispanic in Puerto Rico in accordance with government practice given the ambiguity in this category (Springer, 2005, footnote 9). After 2005, OPM created a new combined race and ethnicity variable that enables respondents to select both a race and a Hispanic ethnicity. Additionally, OPM collapsed the various Asian national categories into a single Asian category, and separated out Native Hawaiian and/or Other Pacific Islander into its own category.

To make races comparable across years, we follow OPM's guidance and aggregate the Asian and Native Hawaiian and/or Other Pacific Islander categories to a single Asian category that is consistent with the aggregation for the pre-2005 data. Additionally, we code individuals that report a Hispanic ethnicity as Hispanic and disregard their self-reported race (if they did report one). In the regressions, we include indicator variables for four racial groups: American Indian/Alaska Native (AI/AN), Asian, black, and Hispanic. The omitted reference category is white.

We also include other variables plausibly correlated with race and pay. These include the employee's age as well as its square, and years of educational attainment after high school. We include fixed effects for the bureau in which an individual works to account for time-constant organizational factors that may affect wages, and over eight hundred indicators for individuals' occupations to account for differences in pay structures across occupations. This is the most disaggregated occupational measure available.

Previous research on the racial wage gap in the federal government has found substantial differences between male and female employees (e.g. Lewis, 1998). We therefore perform analyses separately by gender.

There is some question as to how general the occupational information included in regression analyses of pay disparities should be. On the one hand, if individuals are systematically excluded from different occupations on the basis of race, because of discrimination or some other factor, for instance, then including information on that occupation may lead to a biased estimate of racial pay disparities. However, at the same time, there are important differences across occupations in terms of pay structures and career advancement for which analysts would like to control (Bolton and de Figueiredo, 2017). Here, we report results conditional on occupation; the online supplement includes results that condition only on six broad occupation classifications.

4.3 Overall differentials

In the overall analysis, each observation is an employee-year. Most individuals are observed for multiple years, so these observations are not independent. We therefore use robust standard errors that account for clustering at the employee level (Cameron and Miller, 2015). We also include indicators for the year in which the observation occurs, thereby accounting for year-level shocks to wages that are experienced by all employees.

Mimicking the way analysts would use the integrated system, we start by estimating separate models for male and female employees using the synthetic data. The results in Table 2 reveal important relationships between race and pay in the federal government. In particular, according to the synthetic data results, men who identify as AI/AN, Asian, black, and Hispanic are paid significantly less than comparable white male employees. The same holds for women of all race categories except black, where the effect is not distinguishable from zero in terms of both practical or statistical significance. Male (female) employees that identify as AI/AN earn approximately 0.6% (0.9%) less than similarly situated white male (female) employees. The gaps for Asian and black male employees relative to white male employees appear to be significantly larger at 2.8% and 2.1% percent, respectively. These gaps are noticeably smaller for women of these two race categories, even non-existent for black female employees. Hispanic men and women take home about 1.4% less than comparable white employees.

The analyst next would submit requests for verification of these results. For each race coefficient in Table 2, we make a separate verification query using the method in (4) with $\epsilon = 1$; hence, the total privacy loss for both regressions equals 4. We group employees into $m = 50$ partitions, so that each employee is a member of only one partition. Thus, in the language of ϵ-DP, the neighboring databases differ in one employee, as opposed to one employee-year observation. The former is more sensible for verifications of the overall regression. A data snooper with knowledge of all but one employee-year observation could figure out many, if not all, of the values for the missing observation by logical deduction, e.g., easily inferring the age of the missing year and bounding the salary between the previous and successive years. We set the threshold $\gamma_0 = -.01$, and target queries at whether $\beta_j < -.01$ or not.

As evident in Table 2, the posterior modes of the verification measure clearly indicate that the

	Males' Regression			Females' Regression		
Variable	Synthetic	$\hat{r}$	Confidential	Synthetic	$\hat{r}$	Confidential
AI/AN	-.006 (4)	.76	-.019 (12)	-.009 (7)	.97	-.027 (19)
Asian	-.028 (30)	.99	-.040 (43)	-.011 (13)	.42	-.010 (11)
Black	-.021 (39)	.99	-.036 (61)	.00013 (.3)	.003	-.003 (8)
Hispanic	-.014 (22)	.99	-.029 (42)	-.013 (19)	.99	-.021 (30)
Age	.033 (365)		.043 (480)	.023 (286)		.032 (404)
Age Sq.	-.00027 (269)		-.00036 (352)	-.00019 (205)		-.00027 (295)
Education	.013 (122)		.021 (180)	.014 (130)		.023 (198)
Employee-years	13,008,298		12,720,500	12,263,514		11,874,048
Employees	1,446,499		1,430,238	1,390,611		1,348,381

Table 2: Coefficients from overall regression models and posterior modes $\hat{r}$ of verification measures. AI/AN stands for American Indian and Alaska Native, and Asian includes individuals that identify as Native Hawaiian or Pacific Islander. Absolute values of t-statistics are in parentheses. Disparities in sample sizes arise from deletions of cases with missing values in the confidential data analyses.

wage gaps for male employees who are black, Asian, and Hispanic are all at least 1%. The evidence of at least a 1% wage gap for AI/AN men is less obvious but still suggestive, with a posterior mode around .75. Thus, the verification measures validate the findings from the synthetic data regressions of substantial racial wage gaps for black male, Asian male, and Hispanic male employees, and they suggest the synthetic data results for AI/AN male employees are close to accurate as -.006 is not far from -0.01. For women, the posterior models of the verification measure clearly indicate at least 1% wage gaps for Hispanic and AI/AN employees. They also provide strong evidence against a wage gap of at least 1% for female black employees, with a posterior mode near zero. For female Asian employees the verification measure suggests the wage gap could be almost equally likely above or below 1%, as the posterior mode equals .42. This suggests that the true coefficient is likely near -0.01. Thus, the verification measures validate the findings from the synthetic data that the wage gap for Hispanic female employees is at least 1%, but that there is not a substantial wage gap for black female employees. They also suggest that the estimate for AI/AN (-.009) could be an underestimate, since the verification measures suggest that the true coefficient is indeed less than -.01. Finally, they suggest that the synthetic data coefficient for female Asian employees is likely accurate, since it is close to -.01.

We expect that some users might be satisfied with this level of verification, and thus can publish the synthetic data results plus the verification answers. However, others may want to perform the

analysis on the confidential data via the remote access component of the system. As shown in Table 2, in the confidential data regression the estimated coefficients for all four race indicators are negative and statistically significant for both genders, suggesting that non-white employees earn less than white employees. The estimated gaps for men are at least 1.9% across races, with particularly large gaps for black men (3.6%) and Asian men (4.0%). For women, AI/AN, Asian, and Hispanic employees earn 2.7%, 1.0%, and 2.1% less than comparable white female workers. Strikingly, the coefficient estimate for black women is essentially zero, suggesting parity with similarly situated white women. The wage gaps for black women and Asian women are substantially smaller than for men of those race categories.

The effect sizes from the confidential data are fairly similar to those from the synthetic data. This is in accord with the conclusions from the verification measure. The most practically relevant difference in the synthetic and confidential data results exists for employees that identify as AI/AN: the synthetic data show gaps of less than 1% whereas the confidential data show gaps of at least 1.9%. This group of employees is the smallest racial group in the federal government, making it challenging to create accurate synthetic data for them.

4.4 Year-by-year results

We next turn to year-by-year estimates of pay gaps in order to examine potential trends over time. We estimate the same models used in Table 2, except run on each year of data separately. As before, we start with the synthetic data. The synthetic data results in Figure 1 suggest that the wage gap for men has shrunk steadily over the period of the study in all race groups but black males. For black males, the estimated gap appears to be relatively stable throughout the time period. By 2011 in the synthetic data, the wage gap appears to have disappeared for AI/AN and Hispanic men, and reduced to around -1% for Asian men.

For female employees, the story from the synthetic data is more complicated. Figure 2 suggests that AI/AN, Asian, and Hispanic women all had declining wages relative to white women until the early 2000s, when the trend largely reversed, with all three groups making progress toward parity. Indeed, the synthetic results indicate that AI/AN women actually earned more than comparable white women after 2006. For black women, the synthetic data estimates of the wage gap change only slightly, from 0.1% in 1988 to -0.2% in 2011, suggesting negligible wage gaps at any time point

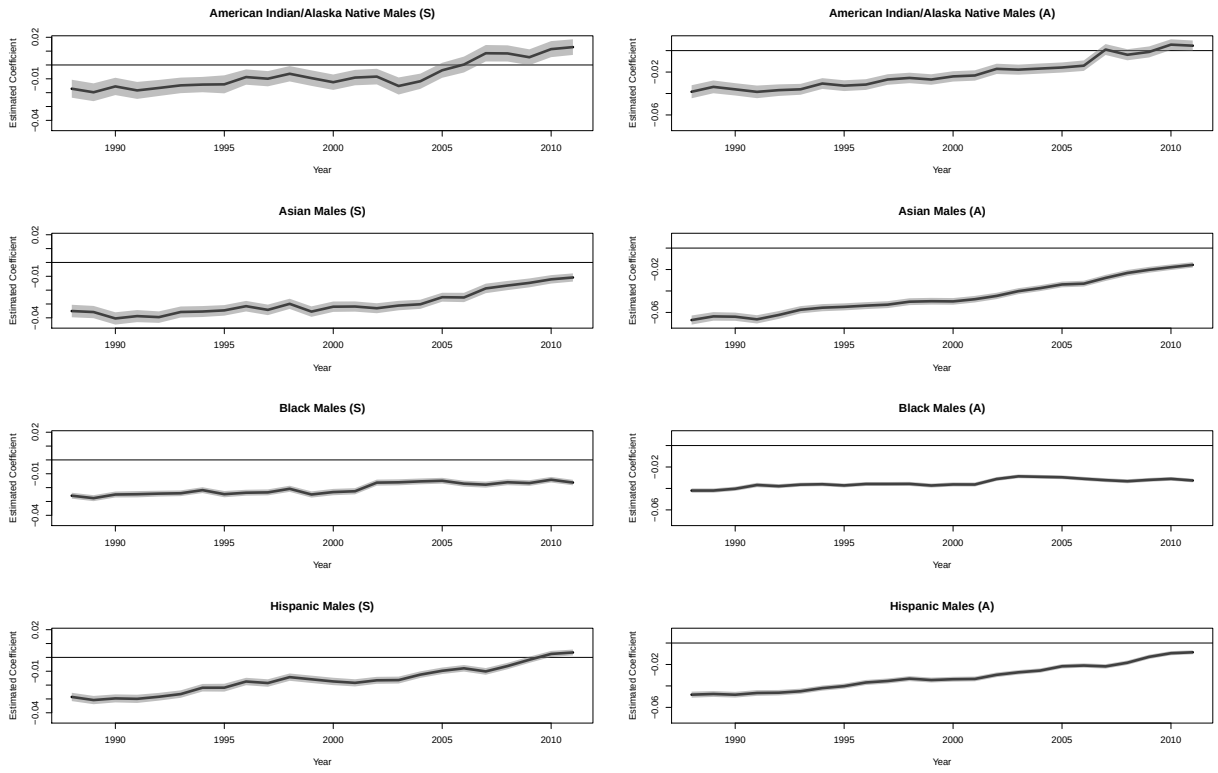

Figure 1: Estimated racial wage gaps (coefficients of race indicators) for yearly males' regressions in synthetic data (left) and confidential, authentic data (right).

in time during the study.

To verify these trends, we estimate the longitudinal measure described in Section 3.2. Looking at Figure 1 for male employees, analysts might consider two sets of time periods $\{\mathcal{T}_k\}$. The first is an overall trend, setting $K = 1$ and $\mathcal{T}_1 = \mathcal{T}$ for all races. The second uses $K = 2$ periods, with a bifurcation at a year where the pattern deviates most noticeably. These are the years 2003 for AI/AN men, 2002 for Asian men, 1998 for black men, and 1997 for Hispanic men. We do the same for female employees, using Figure 2 to identify bifurcations at 1998 for AI/AN women, and at 2000 for all other female employees. We set each C_k to indicate whether the slope is positive ($C_k = [0, \infty]$) or negative ($C_k = [-\infty, 0]$). Of course, one could examine other time periods and intervals. For each verification, we use $\epsilon = 1$ and $M = 50$ partitions, ensuring that each employee appears only once in each partition.

Table 3 displays the posterior modes of the verification measures for the two sets of periods. For men, the posterior modes are all at least 0.7 for all time periods and all races, with most

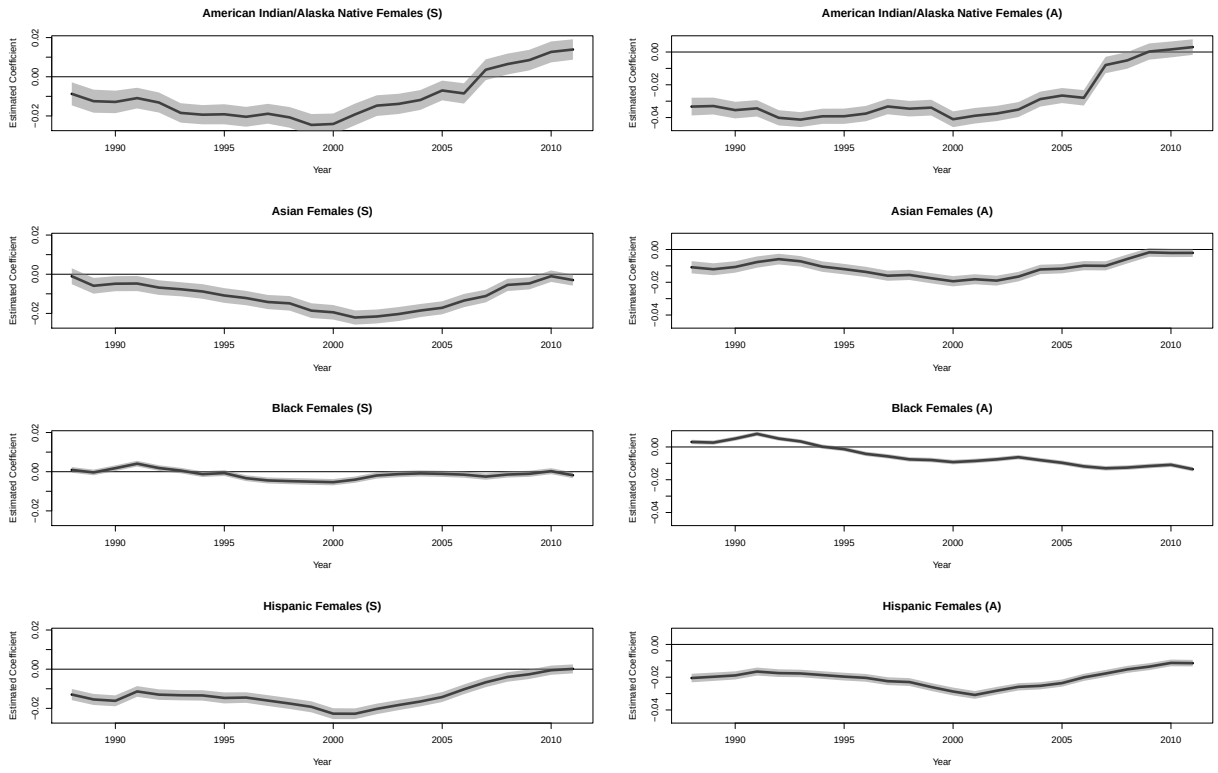

Figure 2: Estimated racial wage gaps (coefficients of race indicators) for yearly females' regressions in synthetic data (left) and confidential, authentic data (right).

above 0.9. This indicates that the trends in the synthetic data coefficients accord well with the trends in the confidential data regressions for these two sets of periods. For women, however, the verification results give reason to doubt some of the trends in the synthetic data regressions. For AI/AN women, we see strong agreement in the overall trend over all years and the trend from 1998 onward, but some uncertainty about the trend between 1998 and 2008. Verification values around .50 are consistent with a nearly flat trend in the confidential data coefficients, which is also the trend in the synthetic data. For Asian women, we see strong agreement in the synthetic and confidential regression trends over 1998 to 2011, modest agreement from 1988 to 2000, and poor agreement over the whole period. The synthetic data trend suggests the wage gap for Asian women in 2011 is nearly the same value as in 1988; however, the verification measures suggest that is not the case. For black women, the verification results confirm that the wage gap increased over the 24 years as a whole, and in particular between 1988 and 1998. However, the trend in the synthetic data coefficients—which suggests black women actually caught up to white women—is

	Males		Females	
Coefficient	Interval	$\hat{r}$	Interval	$\hat{r}$
AI/AN	1988 - 2011	.94	1988 - 2011	.94
	1988 - 2003	.82	1988 - 1998	.60
	2003 - 2011	.91	1998 - 2011	.98
Asian	1988 - 2011	.99	1988 - 2011	.33
	1988 - 2002	.89	1988 - 2000	.72
	2002 - 2011	.96	1998 - 2011	.98
Black	1988 - 2011	.89	1988 - 2011	.99
	1988 - 1998	.74	1988 - 2000	.98
	1998 - 2011	.71	2000 - 2011	.14
Hispanic	1988 - 2011	.99	1988 - 2011	.55
	1988 - 1997	.85	1988 - 2000	.74
	1997 - 2011	.99	2000 - 2011	.97

Table 3: Posterior modes $\hat{r}$ of verification measures for year-by-year trends. AI/AN stands for American Indian and Alaska Native, and Asian includes individuals that identify as Native Hawaiian or Pacific Islander.

not accurate. With a posterior mean of .14, we clearly should not trust the trend for black women in the synthetic data after 1998. For Hispanic women, we see strong agreement in the synthetic and confidential regression trends after 2000, and modest agreement from 1988 to 2000. Between 1998 and 2011, however, the verification measure is close to 0.5, suggesting that the trend line from the confidential data is nearly flat for Hispanic women.

Turning to the results on the confidential data, Figures 1 and 2 show that the race wage gap has been shrinking for all groups except black female employees. In the confidential data, we estimate a significant decline in the position of black women relative to white women in the federal service during the time period of our study. In 1988, we estimate that black women earned 0.3% *more* than similar white counterparts. By 2011, black women were earning approximately 1.4% *less* than white women with similar demographics and occupations.

The trends observed in the synthetic dataset are largely mirrored in the confidential data, with the exception of black female employees. This was apparent in the verification measures, as well, which highlighted the mismatch in the trends for black women after 2000. In both analyses, however, it is clear that black women have not experienced the gains that women identifying with other races have relative to white women. In general, estimated coefficients from the synthetic data analyses tend to be smaller in magnitude than those from the confidential data analyses. There

are some sign discrepancies for the estimated coefficients as well. For instance, in the synthetic results, Hispanic males are estimated to have higher levels of pay relative to comparable white men in the final years of the analysis. However, the coefficient estimates from the confidential data for Hispanic males never exceed zero.

5 Concluding Remarks

The integrated system is based on synthetic data plus verification, coupled with access to the confidential data via a PRDN. It seems natural to ask, why bother releasing record-level data at all? Why not only allow users to query a system for disclosure-protected outputs of analyses? This perspective is evident in some literature on differential privacy (e.g., Dwork *et al.*, 2009; Ullman and Vadhan, 2011), although it is feasible in some settings to generate synthetic data that satisfy, at least approximately, some variant of differential privacy (e.g., Barak *et al.*, 2007; Abowd and Vilhuber, 2008; Blum *et al.*, 2008; Machanavajjhala *et al.*, 2008; Charest, 2010; Hardt *et al.*, 2012; Mir *et al.*, 2013; Karwa and Slavkovic, 2015). We believe that releasing some form of record-level data has enormous benefits. Record-level data provide readily accessible testbeds for methodological researchers to evaluate their latest techniques. They help students and trainees, who may not be able to gain approval to use a PRDN or other secure data enclave, learn the skills of data analysis. Even for experienced researchers, large-scale data can be difficult to "get your head around" because of complexities and structural subtleties that are difficult to learn without seeing the data. Researchers often do not know in advance which are the right questions to ask of the data or the best modeling choices for addressing those questions. As noted by Karr and Reiter (2014), exploratory analyses dealing with the data themselves are a fruitful path to the right questions.

The key to the integrated system is the differentially private verification measures, which are based on binary variables computed on sub-samples of the confidential data. This is a generic method that can be adapted to handle many types of comparisons, making it a flexible strategy for verification. For some analyses and datasets, however, the partitioning process can result in inestimable regressions. For example, the random sub-sampling may result in partitions that have perfect co-linearities or dummy variables with all values equal to zero. Many software packages

automatically drop such variables and report coefficients from the remaining variables, making it still possible to compute the measures although complicating interpretations of the results. When errors make it impossible to obtain results, we suggest adapting the binary measure by adding a third category of counts corresponding to the number of errors. Here, the outputs of the measure include the number of ones, zeros, and errors. We can protect these counts using the Laplace Mechanism, and report posterior modes of the number of errors and the fraction of ones among cases without errors. In the wage gap analyses, fitting errors did not occur due to the large sample sizes.

The choice of the number of partitions is up to the data analyst; we used $M = 50$ in the OPM analyses. Analysts should strive to make M as large as possible to minimize the impact of the Laplace noise on the verification counts. On the other hand, users should allow r to be as close to one (or zero) as possible, as these values are easiest to interpret. Making M too large flattens the distribution of the MLEs in the partitions, thereby moving r toward 0.5 and more uncertain verification decisions. We found that $M = 50$ gave a satisfactory trade off in the OPM data. Analysts can experiment with the synthetic data to find a suitable M for their sample size. Another possibility is to spend some of the privacy budget on selecting an optimal M from a discrete set of choices, according to some loss function that depends on M and r. Developing such measures is an area for future research.

An advantage of ϵ-DP is that one can quantify the leakage from each additional release. If one follows ϵ-DP strictly, at some point the total privacy budget allowed by the owner of the data will be exhausted, at which point no new analysis results may be released. With a finite budget, one has to decide who gets access to the system and in what order. These raise complicated issues of fairness and evaluation of the importance of analyses, which have yet to be addressed in production settings. Clearly these are opportunities for research.

Finally, we conclude by noting that we are developing a verification server and associated R package that implements the verification measures from Section 3. The package also offers methods for generating differentially private plots of residuals versus predicted values for regressions, thereby helping users assess the reasonableness of the assumptions of a posited model when applied on the confidential data (Chen *et al.*, 2017). The server keeps track of the total ϵ used and ensures that the system returns the same noisy answer whenever it is asked for verification of the same query. This

package will be available on CRAN, so that data stewards and other researchers can experiment with and further develop this framework for providing access to confidential social science data.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/synth_data.html`.

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygun, eds., *Privacy in Statistical Databases*, 239–246. New York: Springer-Verlag.

Altonji, J. G. and Blank, R. M. (1999). Race and gender in the labor market. *Handbook of Labor Economics* **3**, 48, 3143–3259.

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. *Proceedings of the 27th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems* .

Black, D. A., Kolesnikova, N., Sanders, S. G., and Taylor, L. J. (2013). The role of location in evaluating racial wage disparity. *IZA Journal of Labor Economics* **2**, 2.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: extent, trends, and expectations. *Journal of Economic Literature* forthcoming.

Blum, A., Ligett, K., and Roth, A. (2008). A learning theory approach to non-interactive database privacy. *Proceedings of the 40th ACM SIGACT Symposium on Thoery of Computing* .

Bolton, A. and de Figueiredo, J. M. (2016). Why have federal wages risen so rapidly? Tech. rep., Duke University Law School.

Bolton, A. and de Figueiredo, J. M. (2017). Meausuring and explaining the gender wage gap in the U. S. federal government. Tech. rep., Duke University Law School.

Bolton, A., de Figueiredo, J. M., and Lewis, D. E. (2016). Elections, ideology, and turnover in the U.S. federal government. Tech. rep., National Bureau of Economics Research Working Paper 22932.

Borjas, G. J. (1980). Wage determination in the federal government: The role of constituents and bureaucrats. *The Journal of Political Economy* **88**, 1110–1147.

Borjas, G. J. (1982). The politics of employment discrimination in the federal bureaucracy. *Journal of Law and Economics* **25**, 2, 271–299.

Borjas, G. J. (1983). The measurement of race and gender wage differentials: Evidence from the federal sector. *ILR Review* **37**, 1, 79–91.

Callier, V. (2015). How fake data could protect real people's privacy. *The Atlantic* **July 30**.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources* **50**, 317–373.

Cancio, A. S., Evans, T. D., and Maume, D. J. J. (1996). Reconsidering the declining significance of race: Racial differences in early career wages. *American Sociological Review* **61**, 4, 541–556.

Card, D. and Lemieux, T. (1994). Changing wage structure and black-white wage differentials. *American Economic Review* **84**, 2, 29–33.

Charest, A. S. (2010). How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality* **2:2**, Article 3.

Charles, J. (2003). Diversity management: An exploratory assessment of minority group representation in state government. *Public Personnel Management* **32**, 4, 561–577.

Chen, Y., Machanavajjhala, A., Reiter, J. P., and Barrientos, A. F. (2017). Differentially private regression diagnostics. In *Proceedings IEEE International Conference on Data Mining*, 81–90. ICDM.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer-Verlag.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008). A new approach for disclosure control in the IAB Establishment Panel–Multiple imputation for a better data access. *Advances in Statistical Analysis* **92**, 439 – 458.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages, and Programming, part II*, 1–12. Berlin: Springer.

Dwork, C., Naor, M., Reingold, O., Rothblum, G., and Vadhan, S. (2009). When and how can privacy-preserving data release be done efficiently? In *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC)*, 381–390.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

Government Accountability Office (2009). Gender pay gap in the federal workforce narrows as differences in occupation, education, and experience diminish. Tech. rep., Government Accountability Office, Washington, DC.

Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. *Advances in Neural Information Processing Systems* **25**, 2348–2356.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

Hoover, G. A., Compton, R. A., and Giedeman, D. C. (2015). The impact of economic freedom on the black/white income gap. *American Economic Review: Papers & Proceedings* **105**, 5, 587–592.

Karr, A. F. and Reiter, J. P. (2014). Using statistics to protect privacy. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, eds., *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 276–295. Cambridge University Press.

Karwa, V. and Slavkovic, A. S. (2015). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Annals of Statistics* **44**, 87–112.

Kim, C.-K. (2004). Women and minorities in state government agencies. *Public Personnel Management* **33**, 2, 165–180.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review* **79**, 363–384.

Lewis, G. B. (1998). Continuing progress toward racial and gender pay equality in the federal service: An update. *Review of Public Personnel Administration* **18**, 2, 23–40.

Lewis, G. B. and Durst, S. L. (1995). Will locality pay solve recruitment and retention problems in the federal civil service? *Public Administration Review* **55**, 371–380.

Lewis, G. B. and Nice, D. (1994). Race, sex, and occupational segregation in state and local governments. *American Review of Public Administration* **24**, 4, 393–410.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Llorens, J. J., Wenger, J. B., and Kellough, J. E. (2007). Choosing public sector employment: The impact of wages on the representation of women and minorities in state bureaucracies. *Journal of Public Administration Research and Theory* **18**, 3, 397–413.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, 277–286.

Maxwell, N. L. (1994). The effect of black-white wage differences of differences in the quantity and quality of education. *Industrial and Labor Relations Review* **47**, 2, 249–264.

McCabe, B. C. and Stream, C. (2000). Diversity by the numbers: Changes in state and local government workforces, 1980-1995. *Public Personnel Management* **29**, 1, 93–106.

McCall, L. (2001). Sources of racial wage inequality in metropolitan labor markets: Racial, ethnic, and gender differences. *American Sociological Review* **66**, 4, 520–541.

McClure, D. and Reiter, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality* **4:1**, Article 8.

Mir, D., Isaacman, S., Caceres, R., Martonosi, M., and Wright, R. N. (2013). DP–WHERE: Differentially private modeling of human mobility. In *Proceedings of the IEEE Conference on Big Data*, 580–588.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy (SP)*.

Neal, D. and Johnson, W. R. (2003). The role of pre-market factors in black-white wage differences. *Journal of Political Economy* **87**, 3, 567–594.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 75–84. ACM.

O'Neill, J. (1990). The role of human capital in earnings differences between black and white men. *Journal of Economic Perspectives* **108**, 937–975.

Parry, M. (2011). Harvard researchers accused of breaching students' privacy. *The Chronicle of Higher Education* .

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

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

Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**, 1475–1482.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Sakano, R. (2002). Are black and white income distributions converging? time series analysis. *Review of Black Political Economy* **30**, 1, 91–106.

Springer, L. M. (2005). Memorandum for chief human capital officers. Office of Personnel Management, 11/09/05.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, & Ethics* **25**, 98–110.

Sweeney, L. (2013). Matching known patients to health records in Washington state data. Tech. rep., Data Privacy Lab, Harvard University.

Ullman, J. and Vadhan, S. P. (2011). PCPs and the hardness of generating private synthetic data. In *Proceedings of the 8th Theory of Cryptography Conference (TCC)*, 400–416.