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

This document includes online supplementary material for the main text. In Section 1, we provide a formal description of the three sub-models used to model the employee's career. In Section 2, we discuss the modeling strategies used to synthesize several variables and to deal with some of the modeling challenges in the SF data. In Section 3, we formally describe the verification measures for longitudinal trends in regression coefficients. In Section 4, we provide the full list of the synthesized variables along with a brief description of each of them. In Section 5, we present the analyses of wage gaps conditional on six broad categories of occupation rather than the 803 used in the main text.

1 Model for Employees' Careers

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 during all 24 years, these sequences do not always have the same length. This poses an additional challenge to be addressed when modeling this variable. To avoid this issue, we create an additional level for this variable. This level corresponds to the status not working. With this additional level, we only have to work with sequences of length equal to 24. Thus, the career of the ith employee is represented by

$$V_1^i = (V_{1,1}^i, \ldots, V_{1,24}^i)$$

where $V_{1,t}^i$ denotes the agency where the ith employee worked in year t. To model these sequences, we create three additional variables: G^i, Z^i, and W^i, where

- G^i is the number of agencies where the ith employee worked during the 24 years,
- $Z^i = (Z_1^i, \ldots, Z_{G^i-1}^i)$ represents the list of years (minus one) when the ith employee moved to a new agency, and

- $W^i = (W_1^i, \ldots, W_{G^i}^i)$ is the ordered sequence of agencies where the ith employee has worked.

Since that $(G^i, Z^i, W^i) \mapsto V_1^i$ is a one-to-one mapping, we can equivalently define a model for either (G^i, Z^i, W^i) or V_1^i. Thus, we define a model for V_1^i by using (G^i, Z^i, W^i) and an appropriate conditional representation. That is,

$$\begin{aligned}\mathbb{P}\left[V_1^i = v\right] &= \mathbb{P}\left[(G^i, Z^i, W^i) = (g, z, w)\right] \\ &= \mathbb{P}\left[W^i = w | (G^i, Z^i) = (g, z)\right] \mathbb{P}\left[Z^i = z | G^i = g\right] \mathbb{P}\left[G^i = g\right],\end{aligned}$$

with the following particular case,

$$\begin{aligned}\mathbb{P}\left[V_1^i = (v, v, \ldots, v)\right] &= \mathbb{P}\left[(G^i, W^i) = (1, v)\right] \\ &= \mathbb{P}\left[W^i = v | G^i = 1\right] \mathbb{P}\left[G^i = 1\right].\end{aligned}$$

We propose to estimate $\mathbb{P}\left[G^i = g\right]$ and $\mathbb{P}\left[V_1^i = (v, v, \ldots, v)\right]$ by using the corresponding observed frequencies. The estimators for the conditional models of Z^i and W^i are explained below.

Model for $\mathbb{P}[Z^i = z | G^i = g]$

Henceforth we suppress the superscript i for ease of notation. For sake of generality, we denote by T the length of the employees' career. Conditional on G, we define the range of Z as

$$\mathcal{P}_T^G := \left\{(a_1, \ldots, a_G) : a_1 < \ldots < a_G,\ (a_1, \ldots, a_G) \in \{1, \ldots, T-1\}^G\right\}.$$

For instance, if $T = 4$ and $G = 2$, then the range of Z corresponds to the set

$$\mathcal{P} = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}. \tag{1}$$

Notice that we can identify the space $\mathcal{P}_T^G$ with elements in the simplex space by considering the mapping,

$$\mathcal{P}_T^G \ni Z = (Z_1, Z_2, \ldots, Z_G) \mapsto S := \left(\frac{Z_1 - 1}{T-1}, \frac{Z_2 - Z_1 - 1}{T-1}, \ldots, \frac{Z_G - Z_{G-1} - 1}{T-1}\right) \in \Delta_G,$$

where Δ_G is the G-dimensional simplex space; that is,

$$\Delta_G = \left\{(a_1, \ldots, a_G) \in [0,1]^G : \sum_{j=1}^{G} a_j \leq 1\right\}.$$

In the example above, under this mapping, the set given in (1) is identified with the following set

$$\Delta = \left\{ (0,0), \left(0, \frac{1}{3}\right), \left(0, \frac{2}{3}\right), \left(\frac{1}{3}, 0\right), \left(\frac{1}{3}, \frac{1}{3}\right), \left(\frac{2}{3}, 0\right) \right\}.$$

Let $\tilde{S} = (\tilde{S}_1, \ldots, \tilde{S}_G)$ be a continuous latent random vector such that, conditioned on G, is Δ_G-valued and

$$S = \left(\frac{\lfloor (T-1)\tilde{S}_1 \rfloor}{T-1}, \ldots, \frac{\lfloor (T-1)\tilde{S}_G \rfloor}{T-1} \right).$$

Thus, we define a probability model for Z, conditional on G, of the form

$$\begin{aligned}
&\mathbb{P}\left[Z = z \middle| G = g \right] \\
&= \mathbb{P}\left[S = \left(\frac{z_1 - 1}{T-1}, \frac{z_2 - z_1 - 1}{T-1}, \ldots, \frac{z_g - z_{g-1} - 1}{T-1} \right) \middle| G = g \right], \\
&= \mathbb{P}\left[\tilde{S} \in \left(\frac{z_1 - 1}{T-1}, \frac{z_1}{T-1} \right] \times \left(\frac{z_2 - z_1 - 1}{T-1}, \frac{z_2 - z_1}{T-1} \right] \times \ldots \times \right. \\
&\qquad \left. \left(\frac{z_g - z_{g-1} - 1}{T-1}, \frac{z_g - z_{g-1}}{T-1} \right] \middle| G = g \right].
\end{aligned}$$

Since $\tilde{S}$ is a continuous random vector on the simplex, we assume that its corresponding probability measure is absolutely continuous with respect to the Lebesgue measure with density $f_{\tilde{S}}$. This probability density function is assumed to be a mixture of Dirichlet densities of the form,

$$f_{\tilde{S}}(\tilde{s} | G = g, \mathbf{p}, \boldsymbol{\alpha}) = \sum_{\mathbf{j} \in \mathcal{P}_T^g} p_{\mathbf{j}} \mathrm{dir}_g\left(\tilde{s} \,|\alpha_{\mathbf{j}}\right)$$

where $\mathbf{p} = (p_{\mathbf{j}})_{\mathbf{j} \in \mathcal{P}_T^g}$ denotes the weights, $\mathrm{dir}_g(\cdot|a)$ stands for a g-dimensional Dirichlet density with parameters a, and $\boldsymbol{\alpha} = (\alpha_{\mathbf{j}})_{\mathbf{j} \in \mathcal{P}_T^g}$. The number of components of this mixture is equal to the cardinality of $\mathcal{P}_T^g$, i.e., equal to the number of elements in the range of Z.

Models designed to generate synthetic datasets should have parameters that allow users to control the trade-off between privacy and statistical usefulness. For this reason, we propose to estimate the parameter $\mathbf{p}$ and use a deterministic definition for $\boldsymbol{\alpha}$. Specifically, we propose to estimate $\mathbf{p}$ using the empirical frequencies, i.e.,

$$\hat{p}_{\mathbf{j}} = \hat{p}_{\{j_1, j_2, \ldots, j_g\}} \propto \sum_{i \in \{l : G_i = g\}} \mathbb{I}_{\left\{ Z_1^i = j_1, \ldots, Z_g^i = j_g \right\}}.$$

Instead of using frequentist estimation of $\mathbf{p}$, we could also place a prior distribution on $\mathbf{p}$ and provide an estimation through the posterior distribution. Regarding the parameter $\boldsymbol{\alpha}$, we propose the following definition.,

$$\alpha_{\mathbf{j}} = \theta \left(2j_1 - 1, 2(j_2 - j_1) - 1, \ldots, 2(j_{g-1} - j_g) - 1, 2(T-1) - 2j_g + g \right),$$

where θ is a positive constant. If the parameters $\mathbf{p}$ and $\boldsymbol{\alpha}$ are defined as before, then the overfitting and underfitting of the model can be controlled by θ. Notice that if a model overfits the confidential dataset, then the synthetic dataset obtained from that model should be very similar to the confidential dataset. Hence, overfitting implies high statistical usefulness but low privacy level. Under a analogous reasoning, we can associate the underfitting with low statistical usefulness but high privacy level. Thus, under this parametrization, the mean and the variance of a Dirichlet distribution with parameter $\alpha_{\mathbf{j}}$ are equal to

$$\left(\frac{2j_1-1}{2(T-1)}, \frac{2(j_2-j_1)-1}{2(T-1)}, \ldots, \frac{2(j_g-j_{g-1})-1}{2(T-1)}\right),$$

and

$$\left(\frac{2(j_l-j_{l-1})\theta[2(T-1)\theta-g-2(j_l-j_{l-1})\theta]}{[2(T-1)\theta-g]^2[2(T-1)\theta-g+1]}\right)_{l=1}^{g},$$

respectively. Notice that this mean always belongs to the inner of the hypercube defined by

$$\mathcal{I}_{\mathbf{j}} := \left(\frac{j_1-1}{T-1}, \frac{j_1}{T-1}\right] \times \left(\frac{j_2-j_1-1}{T-1}, \frac{j_2-j_1}{T-1}\right] \times \ldots \times \left(\frac{j_g-j_{g-1}-1}{T-1}, \frac{j_g-j_{g-1}}{T-1}\right],$$

and notice too that the variance goes to 0 when $\theta \to \infty$. The above statements imply that

$$\mathbb{P}\left[\, Z=z \middle|\, G=g, \theta, \text{Data}\right] = \sum_{\mathbf{j}\in\mathcal{P}_T^g} \hat{p}_{\mathbf{j}} \int_{\mathcal{I}_z} \text{dir}_g\left(\tilde{s}\,|\alpha_{\mathbf{j}}\right) d\tilde{s} \underset{\theta\to\infty}{\longrightarrow} \hat{p}_z,$$

meaning that if θ increases, the latent model assigns probability equal to $\hat{p}_{\mathbf{j}}$ to the hypercube $\mathcal{I}_{\mathbf{j}}$ (i.e., overfitting). On the other hand, when θ decreases, the model borrows information across all the hypercubes $\{\mathcal{I}_{\mathbf{j}}\}_{\mathbf{j}\in\mathcal{P}_T^g}$ (i.e., underfitting). The information related to an specific hypercube $\mathcal{I}_{\mathbf{j}}$ is borrowed across the other hypercubes. The amount of borrowed information depends on the distance between $\mathcal{I}_{\mathbf{j}}$ and the other hypercubes. The smaller the distance, the larger the amount of borrowed information. In particular, if θ is small enough, most of the information will be transferred to the other hypercubes. This increases the probability of generating new careers for the synthetic employees that do not match with the careers of the confidential employees. In other words, if θ is small, it can lead to a degradation of the statistical information contained in a synthetic data, but also an improvement in terms of privacy. Under this model, θ is the parameter controlling the trade-off between statistical usefulness and privacy level.

Model for $\mathbb{P}[W = w|(G, Z) = (g, z)]$

Under some assumptions of independence, we simplify the definition of this model. Specifically, we assume that

$$\begin{aligned}
\mathbb{P}[W = w|(G, Z) = (g, z)] \quad &\propto \quad \mathbb{P}[W = w, G = g, Z = z] \\
&:= \mathbb{P}[W_1 = w_1|G = g, Z_1 = z_1] \\
&\quad \times \mathbb{P}[W_2 = w_2|W_1 = w_1, G = g, Z_1 = z_1, Z_2 = z_2] \\
&\quad \times \prod_{j=3}^{g} \mathbb{P}[W_j = w_j|W_{j-1} = w_{j-1}, G = g, Z_j = z_j, Z_{j-1} = z_{j-1}, Z_{j-2} = z_{j-2}] \\
&\quad \times \mathbb{P}\left[W_{g+1} = w_{g+1}|W_g^i = w_g, G = g, Z_g = z_g, Z_{g-1} = z_{g-1}\right], \\
&= \mathbb{P}[W_1 = w_1|G = g, Z_1 = z_1] \\
&\quad \times \prod_{j=2}^{g+1} \mathbb{P}[W_j = w_j|W_{j-1} = w_{j-1}, G = g, Y_j = y_j] \qquad (2)
\end{aligned}$$

where $Z \mapsto (Y_2, \ldots, Y_{g+1})$ is a one-to-one transformation and

$$Y_j = (Y_{j,1}, Y_{j,2}, Y_{j,3}) = \begin{cases} (0, Z_1 - 1, Z_2 - Z_1 - 1) & \text{if } j = 2, \\ (Z_{j-2} - 1, Z_{j-1} - Z_{j-2} - 1, Z_j - Z_{j-1} - 1) & \text{if } j = 3, \ldots, g, \\ (Z_{g-1} - 1, Z_g - Z_{g-1} - 1, T - Z_g) & \text{if } j = g + 1. \end{cases}$$

Notice that Y_j is related to the moments where the transition $w_{j-1} \to w_j$ is made. The first component of this random vector represents the year (minus one) when the employee started to work for agency w_{j-1}. The second and third components indicate for how many years (minus one) the employee worked for agency w_{j-1} and w_j, respectively. Since the terms in (2) can be re-written of the form,

$$\begin{aligned}
\mathbb{P}[W_1 = w_1|G = g, Z_1 = z_1] \quad &\propto \quad \mathbb{P}[W_1 = w_1, G = g, Z_1 = z_1], \\
&= \sum_{z_2, w_2} \mathbb{P}[Z_1 = z_1, Z_2 = z_2|W_2 = w_2, W_1 = w_1, M = m] \, \mathbb{P}[W_2 = w_2, W_1 = w_1, G = g], \\
&= \sum_{(\cdot, z_2) \mapsto y_2, w_2} \mathbb{P}[Y_2 = y_2|W_2 = w_2, W_1 = w_1, G = g] \, \mathbb{P}[W_2 = w_2, W_1 = w_1, G = g],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}[W_j = w_j|W_{j-1} = w_{j-1}, G = g, Y_j = y_j] \propto \mathbb{P}[W_j = w_j, W_{j-1} = w_{j-1}, G = g] \\
\times \mathbb{P}[Y_j = y_j|W_j = w_j, W_{j-1} = w_{j-1}, G = g],
\end{aligned}$$

we propose to estimate the terms $\mathbb{P}[W_j = w_j, W_{j-1} = w_{j-1}, G = g]$ using the observed frequencies. We could also estimate these probabilities through a multinomial-Dirichlet Bayesian model. For the term $\mathbb{P}[Y_j = y_j|W_j = w_j, W_{j-1} = w_{j-1}, G = g]$, we use the same latent model defined de-

scribed in the previous subsection. This is possible because $Y_j/(T-G) = (Y_{1,j}, Y_{2,j}, Y_{3,j})/(T-G)$ lies in the three-dimensional simplex space. Thus, let $\tilde{Y}_j = (\tilde{Y}_{1,j}, \tilde{Y}_{2,j}, \tilde{Y}_{3,j})$ be a Δ_3-valued continuous latent random vector such that, conditional on G,

$$Y_j = \left(\lfloor (T-M)\tilde{Y}_{1,j} \rfloor, \lfloor (T-M)\tilde{Y}_{2,j} \rfloor, \lfloor (T-M)\tilde{Y}_{3,j} \rfloor \right).$$

We define a probability model for Y_j, conditional on W_j, W_{j-1}, and G, of the form

$$\begin{aligned}
&\mathbb{P}\left[Y_j = (y_{1,j}, y_{2,j}, y_{3,j}) \middle| \, W_j = w_j, W_{j-1} = w_{j-1}, G = g \right] \\
&= \mathbb{P}\left[\tilde{Y}_j \in \left(\frac{y_{1,j}}{T-1}, \frac{y_{1,j}+1}{T-1} \right] \times \left(\frac{y_{2,j}}{T-1}, \frac{y_{2,j}+1}{T-1} \right] \times \left(\frac{y_{3,j}}{T-1}, \frac{y_{3,j}+1}{T-1} \right] \middle| \, W_j = w_j, W_{j-1} = w_{j-1}, G = g \right].
\end{aligned}$$

We also assume that the law of $\tilde{Y}_i$ has a probability density function and is defined as a mixture of Dirichlet densities of the form,

$$f_{\tilde{Y}_j}(\tilde{y} | W_j = w_j, W_{j-1} = w_{j-1}, G = g, \mathbf{p}', \boldsymbol{\alpha}') := \sum_{\mathbf{j} \in \mathcal{P}^3_{T-g}} p'_{\mathbf{j}} \text{dir}_3\left(\tilde{y} \, | \alpha'_{\mathbf{j}} \right)$$

where $\mathbf{p}' = (p'_{\mathbf{j}})_{\mathbf{j} \in \mathcal{P}^3_{T-g}}$ and $\boldsymbol{\alpha}' = (\alpha'_{\mathbf{j}})_{\mathbf{j} \in \mathcal{P}^3_{T-g}}$. Since our goal is still the same, i.e., to propose models that provide control over the trade-off between privacy and statistical usefulness, the parameters $\mathbf{p}'$ and $\boldsymbol{\alpha}'$ are estimated and defined in a similar manner to the one proposed in the previous subsection. Specifically,

$$\hat{p}'_{\{j_1, j_2, j_3\}} \propto \sum_{\substack{j = 2, \ldots, g+1 \\ i \in \{l : G^i_l = g, W^i_j = w_j, W^i_{j-1} = w_{j-1}\}}} \mathbb{I}_{\left\{ Y^i_{1,j} = j_1 - 1, Y^i_{2,j} = j_2 - j_1 - 1, Y^i_{3,j} = j_3 - j_2 - 1 \right\}}.$$

and

$$\alpha'_{\mathbf{j}} = \theta \left(2j_1 - 1, 2(j_2 - j_1) - 1, 2(j_3 - j_2) - 1, 2(T-g) - 2j_3 + 3 \right).$$

Here, the implications of increasing or decreasing the value of θ remain the same as in the previous subsection.

2 General Strategies for Synthesizing the OPM Dataset

The modeling of the SF dataset requires us to deal with many non-trivial problems. For each of these problems, we design different strategies that take run time and computational resources into account. In this section, we provide a more detailed description of the most relevant strategies proposed during the modeling of the SF dataset.

2.1 Deriving predictors from employees' careers

After generating the synthetic careers, we create a set of variables that are functions of the employees' careers. These variables serve as predictors in the modeling of the remaining variables in the SF dataset. Specifically, we create the following variables.

- *Initial year*: year when the employee was included in the dataset.
- *Last year*: year in which the employee stopped working in the last agency.
- *Total years*: number of years that the employee worked.
- *Initial agency*: agency in which the employee started working.
- *Number of moves*: number of times that the employee changed agency during her/his career.
- *Number of gaps*: number of times that the employee stopped working for at least one year and then started working again.

2.2 General strategy for static variables

Static variables are those variables whose values remain the same across time. We model these variables using classification and regression trees (CART), as described in Reiter (2005). For each static variable, we use as predictors the variables derived from the employees' careers along with all the original values of the variables previously synthesized. The sex and a binary variable associated with months of military service are classified in this category.

2.3 General strategy for longitudinal variables

Longitudinal variables are those variables that do not change deterministically across time. Let $t_1^i < \ldots < t_{n_i}^i$ be the years when ith employee is observed and $V_j^i := (V_{j,t_1^i}^i, \ldots, V_{j,t_{n_i}^i}^i)$. If V_j^i is a longitudinal variable, $j > 1$, then we consider the following conditional representation of p_j,

$$p_j\left(V_j^i \,|V_1^i, \ldots, V_{j-1}^i\right) := \prod_{l=1}^{n_i} p_{j,t_l^i}\left(V_{j,t_l^i}^i \,\middle|\, V_1^i, \ldots, V_{j-1}^i, V_{j,t_1^i}^i, \ldots, V_{j,t_{l-1}^i}^i\right) \tag{3}$$

where p_{j,t_l^i} denotes the distribution V_{j,t_l}^i which is conditioned on the values of the previous variables and the past values of V_j^i, i.e., $V_{j,t_1^i}^i, \ldots, V_{j,t_{l-1}^i}^i$. In order to simply the modeling of p_{j,t_l^i}, we assume that

$$p_{j,t_l^i}\left(V_{j,t_l^i}^i \,\middle|\, V_1^i, \ldots, V_{j-1}^i, V_{j,t_1^i}^i, \ldots, V_{j,t_{l-1}^i}^i\right) = p_{j,t_l^i}\left(V_{j,t_l^i}^i \,\middle|\, V_{1,t_l^i}^i, \ldots, V_{j-1,t_l^i}^i, V_{j,t_{l-1}^i}^i\right), \tag{4}$$

This assumption implies that the conditional distribution of V^i_{j,t^i_j} only dependents of current values of V^i_l, $1 < l < j$, and the nearest past value of V^i_j. We estimate these conditional probabilities using CART models.

2.4 General strategy for variables with a high proportion of constant sequences

There are variables whose values do not change across time for most employees. Specifically, race and educational level show this pattern. For this reason, we create an auxiliary binary variable that indicates whether the values of the variable remain the same across time or not. After imputing this binary variable to the synthetic employees using CARTs, we divide the dataset into two groups. The first group represents those employees whose values remain the same across time. This group is modeled using the general strategy for static variables. The second group represents those employees whose values change across time. We model this group using the general strategy for longitudinal variables.

2.5 General strategy for oddities

Some observations have values that are theoretically impossible. For example, for any given employee, we expect the values associated with educational level are not decreasing. However, we observe that there are some employees whose educational level drops at some point. We assume that a drop in the educational level should be considered as an oddity. The SF synthetic dataset represents a methodological tool for those researches that will only have access to the original dataset for a limited period of time. For those researchers, the synthetic dataset can be used to define which models they plan to run when they have access to the original dataset. Hence, fitting a model to the synthetic data should lead to similar challenges to the ones the researches will face when they access the original dataset. For this reason, we define a model able to generate those oddities. To do so, we create a binary variable that indicates whether the employee contains an oddity or not. Thus, we use this binary variable to fit a CART model that allows us to classify the synthetic employees in two groups. The first is the group that is synthesized with a model that does not generate oddities, i.e., a model that only generates non-decreasing sequences. The second is a group that is synthesize with a model that generates oddities with positive probability.

2.6 General strategy for bucketed continuous variables

Age and yrsdegrng—years since the employee earned the degree mentioned in educational level—are classified in this category. The levels of these variables are reported in 5-year buckets. For this reason, we model age and yrsdegrng as categorical variables using the first reported

bucket as a response variable. Thus, we synthesize these variables using the general strategy for static variables. Once we impute the first age and yrsdegrng to the synthetic employee, we deterministically impute the values of the next years using the middle of the range of the buckets. For example, we impute a_t—age in year t—by adding one to the mid-range value in the previous year a_{t-1}; that is, $a_t = a_{t-1} + 1$. Finally, we bucket the imputed values back into 5-year buckets.

2.7 General strategy for variables with a large number of levels

Fitting a CART using a response variable with a large number of levels requires a high computational cost. In fact, the R function `tree` only allows a response variable with at most 32 levels. To wave this issue, we create an auxiliary variable that is a copy of the original variable with only 32 levels, where the first 31 levels correspond to those levels with the highest observed frequencies and the last level groups the remaining levels. Then we fit a CART model to this auxiliary variable and predict the values for the synthetic employees. Thus, there are some synthetic employees having the value associated with the last level of the auxiliary variable. For those employees, we re-impute their values using a CART fitted to a new auxiliary variable. This CART is fitted to a subset of the original dataset that does not contain those employees whose values correspond to one of the 31 levels with most data points. This new auxiliary variable also has 32 levels. The first 31 levels correspond to those levels with most data points in such subset and the last level grouped the rest of the levels. We repeat this process until we reach those levels with the smallest observed frequencies.

2.8 General strategy for variables with low observed frequency levels

This strategy is used for occupation. This variable has over eight hundred levels each year. Some of these occupations have a very low observed frequency. Therefore, the probability that we impute one of these occupations to a synthetic employee is also small. In fact, we observe that, after having used the general strategy for variables with a large number of levels, there are some occupations in the synthetic data with an observed frequency equal to zero. The absence of these occupations is problematic for those researchers interested in occupations with small frequency. In other words, it would be useless for them to have a synthetic dataset where some of these occupations do not appear. To deal with this issue, we start modeling this variable by using the general strategy for variables with a large number of levels. Thus, we guarantee that those occupations with a large number of data points are reasonably represented in the synthetic dataset. For those occupations with a small number of observations, we use propensity score matching. Specifically, we combine the synthetic and original datasets into one dataset. Then, we fit a logistic regression model using whether the employee is a synthetic one or not as a response variable. We use the predicted probabilities to match synthetic employees with

authentic employees having an occupation that has a small observed frequency. Thus, the synthetic employee is assigned to the same occupation of the corresponding matched authentic employee.

2.9 General strategy for structural zeros

We define a structural zero as an impossible combination of levels of different variables. For example, if we have age and educational level, the combination one-year-old and college degree should occur with a probability equal to zero. Defining models for a categorical response that deal with structural zeros is a difficult task. This task can be even more difficult if there is no exhaustive list of these impossible combinations. Structural zeros can occur if we do not carefully model some of the variables in the SF dataset. For example, there are occupations absent in certain agencies or some grades that only make sense within a particular pay plan. The strategy we use here is to split the dataset into subsets such that if we fit a CART model to that subset, the CART model will not impute values that produce structural zeros. Specifically, we always split the data at least into agency. However, for some variables, we require dividing the dataset considering other variables. For example, we know that the levels that grade can take are restricted by the pay plan. In this case, we have to divide the dataset not only by agency but also by pay plan.

2.10 General strategy for missing values

Almost all variables of the SF dataset have missing values. To deal with this aspect, we assume that the missing status is an additional level that each variable can take. Hence, under this strategy, models fitted to the confidential dataset are able to generate synthetic employees with missing values. This leads to synthetic datasets more similar to the confidential one regarding the presence of missing values. Thus, users can design modeling strategies that account for missing data using the synthetic dataset. These strategies could potentially be implemented in the confidential dataset if the user plans to access it in the future.

Another strategy to deal with missing data is imputation. In this work, we impute a missing value if a deterministic rule can provide a reasonable approximation of the unobserved value. We can think of this strategy more as a cleaning step than as a formal statistical procedure to deal with missing values. This strategy is used in only one variable, educational level. The rules that we consider for this specific variable are:

- If the initial educational levels are missing, we impute those with the first reported educational level.

- If the last educational levels are missing, we impute those with the last reported educational level.

- If the educational level is reported at year t_1 and t_2, with $t_1 < t_2$, and the values of this variable are missing for every year $t \in \{t_1 + 1, \ldots, t_2 - 1\}$, then we impute those missing values with the educational level reported at year t_1.

Notice that, under these rules, some employees can still have missing values in this variable. Specifically, those employees for whom all the educational level values are missing will not be assigned any values for educational level after the imputation. To deal with this issue, as before, we model this variable assuming that the missing status corresponds to an additional educational level.

2.11 General strategy for allowances

For each year, we create a binary variable to predict whether or not the synthetic employee receives an allowance. We model these binary sequences using the general strategy for longitudinal variables. Then, for each year and using those authentic employees that have received an allowance, we compute how much this allowance is as a percentage of basic pay. This creates a population of percentages related to the allowances received by the employees each year. If a synthetic employee is classified as receiving an allowance, then we compute this allowance by multiplying her basic pay by a percentage randomly drawn from the percentage population of the corresponding year.

3 Longitudinal Verification Measure

In addition to verifying whether a given coefficient exceeds some threshold, analysts can also be interested in studying how β_{jt} changes across time, where β_{jt} is the regression coefficient of interest at year t. To study how β_{jt} changes across time, we assume that $\boldsymbol{D}$ can be divided into nonempty subsets $\boldsymbol{D}^1, \ldots, \boldsymbol{D}^{24}$, where $\boldsymbol{D} = \{(x_i, y_i)\}_{i=1}^n$, $\boldsymbol{D}^t$ denotes all the data points in $\boldsymbol{D}$ observed at year t, $y_i \in \mathbb{R}$ is the response variable, and $x_i = (1, x_{i,1}, \ldots, x_{i,p})^T \in \mathbb{R}^{p+1}$ are the set of predictors. We also assume that, for every $(y_{it}, x_{it}) \in \boldsymbol{D}^t$, $E(y_{it}|x_{it}) = \beta_t^T x_{it}$, where $\beta = (\beta_{0t}, \ldots, \beta_{pt})^T \in \mathbb{R}^{p+1}$. To formally state our goal here, let $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}}\right)$ be a $\{0,1\}$-valued function which returns a zero if the OLS line passing through the points $\{(t, \beta_{jt})\}_{t\in\mathcal{T}}$ has negative slope and returns a one if the slope is positive, where $\mathcal{T}$ is a period of years. Here, we presume the analyst is specifically interested in checking whether β_{jt} has an increasing or decreasing trend in a given period of years $\mathcal{T}$, i.e., whether $m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}}\right)$ equals zero or one. More generally, we assume the analyst can consider K periods of the form $\mathcal{T}_k = [t_{k-1}, t_k]$, $k = 1, \ldots, K$, where $1 = t_0 < t_1 < \ldots < t_K = 24$, and check whether the trend of β_{jt} is increasing or decreasing within each $\mathcal{T}_k$. Hence, for a given sequence $(\tau_1, \ldots, \tau_K) \in \{0,1\}^K$, we can think of the analyst's interest as an inference problem where the parameter to infer is defined by $\theta_0 = \prod_{k=1}^K \mathbb{I}_{\{\tau_k\}}\left(m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}\right)\right)$. Notice that θ_0 is a binary parameter such that it is equal

to one when $\tau_k = m\left(\{(t, \beta_{jt})\}_{t\in\mathcal{T}}\right)$, for every $k = 1, \ldots, K$, and is equal to zero otherwise. For example, an analysis of whether the trend of β_{jt} is decreasing during the first 9 years and is increasing during the last 15 years would examine whether $\theta_0 = 1$ when $(\tau_1, \tau_2) = (0, 1)$, $\mathcal{T}_1 = [1, 9]$, and $\mathcal{T}_2 = [10, 24]$.

Similar to the DP verification procedure for the threshold, and because of the large sample sizes in the SF data, approximated inferences for θ_0 can be made by using a pseudo parameter θ_N. This pseudo parameter is a function of the sampling distribution of the MLE of β_{jt}, $t = 1, \ldots, 24$. We define the pseudo parameter θ_N by

$$\theta_{N_1,\ldots,N_{24}} = \begin{cases} 1 & \text{if } P[m(\{(t, \hat{\beta}_{jt}^{N_t})\}_{t\in\mathcal{T}_k}) = \tau_k, k = 1, \ldots, K] \geq \gamma_1, \\ 0 & \text{if } P[m(\{(t, \hat{\beta}_{jt}^{N_t})\}_{t\in\mathcal{T}_k}) = \tau_k, k = 1, \ldots, K] < \gamma_1. \end{cases}$$

where $\hat{\beta}_{jt}^{N_t}$ is the MLE of β_{jt} based on a sample with N_t individuals and, again, $\gamma_1 \in (0, 1)$ reflects the degree of certainty the user requires before she decides there is enough evidence to conclude that $\theta_0 = 1$. In this case, if $\hat{\beta}_{jt}^{N_t}$ is a consistent estimator of β_{jt}, we have that $\lim_{\forall t, N_t\to\infty} \theta_{N_1,\ldots,N_{24}} = \theta_0$.

Since making inferences about $\theta_{N_1,\ldots,N_{24}}$ is equivalent to making inferences about $r = P[m(\{(t, \hat{\beta}_{jt}^{N_t})\}_{t\in\mathcal{T}_k}) = \tau_k, k = 1, \ldots, K]$, we focus on providing a DP procedure for releasing inferences for r. This procedure is based on the subsample and aggregate method. We start by randomly splitting each $\boldsymbol{D}^t$ into M disjoint subsets, $\boldsymbol{D}_1^t, \ldots, \boldsymbol{D}_M^t$, of the same size (or approximately the same size when n_t/M is not an integer with $n_t = |\boldsymbol{D}^t|$), where M is specified by the user. Then, in each D_l^t, we compute the MLE b_{jtl} of β_{jt}. We assume that, for each t, $b_{jt1}, \ldots, b_{jtM}$ is a random sample from the sampling distribution of $\hat{\beta}_{jt}^{N_t}$, where $N_t = n_t/M$. Let $W_l = \prod_{k=1}^{K} \mathbb{I}_{\{\tau_k\}}(m\left(\{(t, b_{jtl})\}_{t\in\mathcal{T}_k}\right)$ and $S = \sum_{l=1}^{M} W_l$. Since $W_1, \ldots, W_M$ are independent Bernoulli distributed random variables with parameter r, we can provide inferences for r by using the Binomial random variable S. Unfortunately, inferences directly based on S can lead to leakage of information. Hence, we propose to make inferences for r based on a DP version of S, say $S^R + S + \eta$, where η is drawn from a Laplace distribution with mean zero and variance $1/\epsilon$.

Based on S^R, we make inferences for r by using the following model,

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

Under this model, the verification server can report back any graph or summary of the posterior distribution of r to the analyst. Then, she can compare any of those outputs with her degree of certainty represented by γ_1 and decides whether or not $\theta_0 = 1$. She can alternatively interpret this posterior distribution for r as an asymptotic approximation of the Bayesian posterior probability, $\pi(m(\{(t, \beta_{jt})\}_{t\in\mathcal{T}_k}) = \tau_k, k = 1, \ldots, K|S^R)$. In our previous example where $(\tau_1, \tau_2) = (0, 1)$, $\mathcal{T}_1 = [1, 9]$, and $\mathcal{T}_2 = [10, 24]$, if the mode of the posterior probability for r equals 0.93, we could say that the posterior probability that the trend of β_{jt} is decreasing during the

first 9 years and is decreasing during the last 15 years is approximately equal to 0.93.

4 List of Synthesized Variables

In this section, we provide a full list of the synthesized variables. The variables in this list are ordered from the first to last to be synthesized.

1. **Agency**. Each entry of personnel data from the Central Personnel Data File (CPDF) is accompanied by a distinct agency identifier (e.g., AG13 or HUAA). These 4-digit codes are a combination of letters and numbers. The first two digits signify the overarching agency (e.g., AG=Department of Agriculture) and the last two digits signify a sub-element within the agency if there is one (e.g., Forest Service).

2. **Sex**. An employee's sex.

3. **Race**. Race or National Origin - An employee's race or national origin. Employees of mixed race or national origin should be identified with the race or national origin with which they most closely associate themselves. This data standard is only applicable to an employee whose accession occurs prior to July 1, 2006. See the ETHNICITY AND RACE IDENTIFICATION data standard for an employee whose accession occurs on or after January 1, 2006.[1]

4. **Eribridge**. The data standard is applicable to accessions occurring on or after January 1, 2006, and is required for accessions occurring on or after July 1, 2006. The data standard consists of one ethnicity category (Hispanic or Latino) and five race categories.

5. **Educ_lvl**. The extent of an employee's educational attainment from an accredited institution.[2]

6. **Agerange**. The age of the employee in the year observed within a particular range. The method to generate the variable is as follows: OPM took the real age, randomly added error, which is uniformly distributed around 0 and goes from -2 to +2. That generates a predicted age, which is then bucketed into 5-year buckets. The variable agerange is the predicted age in the 5-year bucket. This is generated year-by-year, not person-by-person.

7. **Yrsdegrng**. Years since the employee earned the degree mentioned in educ_lvl. OPM took the real year, randomly added error, which is uniformly distributed around 0 and goes from -2 to +2. That generates a predicted year, which is then bucketed into 5-year buckets. The variable yrsdegrng is the predicted number of years in the 5-year bucket. This is generated year-by-year, not person-by-person.

[1] The Guide to Data Standards, Update 16, November 15, 2014, A-420.
[2] The Guide to Data Standards, Update 16, November 15, 2014, A-130.

8. **Milmonths**. The months of military service that are creditable for annual leave accrual purposes. This variable was generated using the milserve (same as CREDIBLE MILITARY SERVICE) variable provided to us by OPM.[3]

9. **Occ**. Occupation - An employees occupational series. Occupational Series 0001 through 2299 represent white collar occupations and occupational series 2501 through 9999 represent blue collar occupations.[4]

10. **Instrctpgm**. Instructional Program - an employees field of study.[5]

11. **Occ_cat**. Occupational Category - The occupational category to which an occupational series belongs.[6]

12. **Funcclas**. Functional Class - An employee's primary work function as a scientist or engineer.

13. **Flsa**. The status of a Federal civilian employee under the authority of Section 13 of the Fair Labor Standards Act (29 U.S.C. 213), as amended.

14. **Appttype**. Type of Appointment the type of appointment under which an employee is serving.[7]

15. **Polappttype**. Political Appointment Type. - Political appointee is a generic term that is not defined in OPM staffing policy. For purposes of our analyses a political appointee, non-career SES employee, or Schedule C employee who can be identified as such in OPMs Central Personnel Data File (CPDF) or Enterprise Human Resources Integration-Statistical Data Mart (EHRI-SDM).

16. **Position**. Position Occupied an employee's position in the Competitive Service, Excepted Service, or the Senior Executive Service.[8]

17. **Tenure**. For purposes of reduction in force, the retention group in which an employee is placed based on the employee's type of appointment.[9]

18. **Svsrstat**. Supervisory status - The nature of managerial, supervisory, or non-supervisory responsibility assigned to an employee's position.[10]

[3]The Guide to Data Standards, Update 16, November 15, 2014, A-86.
[4]The Guide to Data Standards, Update 16, November 15, 2014, A-307.
[5]The Guide to Data Standards, Update 16, November 15, 2014, A-173-A-236.
[6]The Guide to Data Standards, Update 16, November 15, 2014, A-343.
[7]The Guide to Data Standards, Update 16, November 15, 2014, A-510.
[8]The Guide to Data Standards, Update 16, November 15, 2014, A-396.
[9]The Guide to Data Standards, Update 16, November 15, 2014, A-506 A-507.
[10]The Guide to Data Standards, Update 16, November 15, 2014, A-504 A-505.

19. **Bargunit**. An employee's bargaining unit. Bargaining unit names and codes can be found in the Office of Personnel Managements Federal Labor Management Information System (FLIS) website (https://apps.opm.gov/flis/start.aspx).[11]

20. **Pay_plan**. A particular table or array of pay rates prescribed by law or other authoritative source that establishes the basic pay rates for certain employees. In most cases, a pay plan (system) is a two dimensional matrix of pay rates: one dimension providing a series of different pay rates or ranges corresponding to differences in grade (or level, class, rank, or pay band of work) and the other dimension providing a series of pay rates or a range of rates within a grade. These rates may be a function of length of service in the grade or of performance ratings.[12]

21. **Grade**. An indicator of hierarchical relationships among positions covered by the same pay plan or system.[13]

22. **Steprate**. An indicator of a specific salary within a grade, level, class, rate, or pay band.

23. **Paybasis**. The principal condition in terms of time, production, or other criteria that, along with salary rate, determines the compensation paid to an employee.

24. **Workschd**. Work Schedule - The time basis on which an employee is scheduled to work.

25. **Payrated**. A designation of any special factors that help determine an employee's rate of basic pay or adjusted basic pay.[14]

26. **Localpay**. Locality Pay Area - the identification of an area for purposes of locality-based comparability payments.

27. **Paybasic**. The employees rate of basic pay. Exclude supplements, adjustments, allowances, differentials, incentives, or other similar additional payments.

28. **Retallow**. This variable comes from Nature of Action code 827, Retention Incentive.[15]

29. **Svsr_diff**. Supervisory Differential - The annual total dollar amount paid, over and above paybasic, to a General Schedule supervisor who otherwise would be paid less than one or more of the civilian employees supervised.[16]

[11]The Guide to Data Standards, Update 16, November 15, 2014, A-59.
[12]The Guide to Data Standards, Update 16, November 15, 2014, A-352 A-386.
[13]The Guide to Data Standards, Update 16, November 15, 2014, A-168 A-169.
[14]The Guide to Data Standards, Update 16, November 15, 2014, A-387 A-392.
[15]The Guide to Data Standards, Update 16, November 15, 2014, A-302.
[16]The Guide to Data Standards, Update 16, November 15, 2014, A-503.

	Males' Regression		Females' Regression	
Variable	Synthetic	Confidential	Synthetic	Confidential
AI/AN	-.016 (9)	-.034 (17)	-.017 (11)	-.036 (21)
Asian	-.014 (11)	-.027 (20)	.004 (4)	.014 (12)
Black	-.058 (86)	-.083 (116)	-.009 (20)	-.011 (23)
Hispanic	-.016 (20)	-.033 (37)	-.014 (17)	-.018 (22)
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 1: Coefficients from overall regression models. 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. These models include an indicator for an individual's occupational category code – professional, administrative (omitted category), technical, clerical, other white collar, or blue collar. Models also include controls for age, age squared, and educational attainment.

5 Results for Model With Broad Occupation Classification

Table 1 includes estimates of the racial wage gap over the 1988-2011 time period for men and women from the synthetic and authentic datasets. These estimates are from models similar to those reported in the main text of the paper. However, rather than including indicators for disaggregated occupational information, we use indicators for broad occupational indicators created by the Office of Personnel Management. There are six possible categories: administrative (omitted group), blue collar, clerical, other white collar, professional, and technical.

In general, administrative jobs require a college degree and are primarily white collar positions with management functions. This is the largest category of positions in the federal government. Professional positions tend to be complex and technical in function, requiring advanced degrees and training. Examples would include engineers and scientists. Clerical positions and technical positions tend to be supportive roles within agencies that do not require a bachelor's degree and/or individuals may be able to be trained on the job. These categories include positions like administrative assistants (clerical) and nursing assistants (technical). Blue collar positions have shrunk in number over time, but they include trades and craft workers in both supervisory and non-supervisory roles. Finally, other white collar positions include those that do not comfortably fall into any of the other white collar categories, such as student trainees for a variety of white collar positions or border patrol enforcement (Office of Personnel Management, 2014).

If individuals' races play an important role in the specific occupations that they work in, then conditioning on occupational information could induce post-treatment bias in our estimates of the racial wage gap. Race may play a role, for example, if there is significant racial discrimination

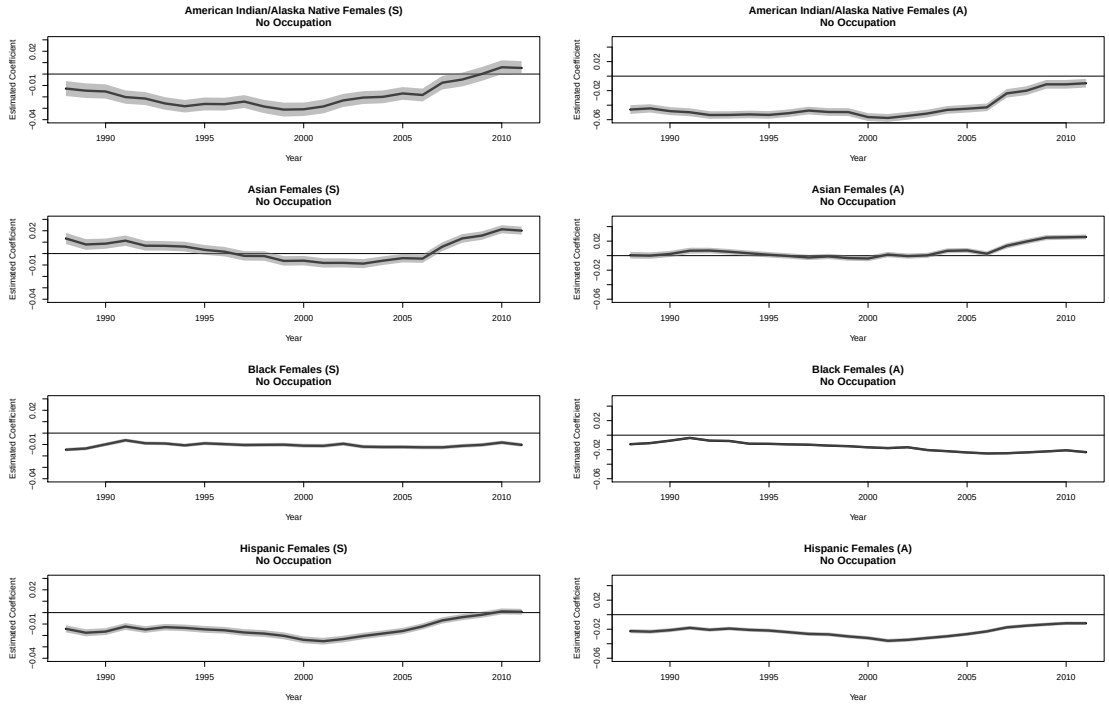

Figure 1: Estimated racial wage gaps (coefficients of race indicators) for yearly females' regressions in synthetic data (left) and confidential, authentic data (right). These models include an indicator for an individual's occupational category code – professional, administrative (omitted category), technical, clerical, other white collar, or blue collar.

or legacies of discrimination in particular occupations or classes of occupations. In that case, conditioning on occupation would not give a full picture of racial pay disparities. However, completely excluding occupational information may lead to incorrect inferences if occupational sorting rather than discrimination is at play. Here, we "split the difference," by including only broad occupational information. Results with no occupational indicators in models are available upon request.

As can be seen in Table 1, the estimated wage gaps are in general larger than those reported in the main text of the paper when including more aggregated occupational indicators. In all cases, except for Asian female employees, we see an estimated negative coefficient for each group, indicating that they are paid less than comparable white employees. These signed relationships are found in both the synthetic and the confidential datasets, although the coefficients estimated from the confidential dataset tend to be of larger magnitude.

Figure 1 displays the over-time trends in estimated racial pay gaps with 95% confidence intervals for female employees in the synthetic and confidential datasets, respectively. Both figures show similar trends both to one another, as well as to those reported in the main text of the paper. The key place in which the synthetic and confidential datasets depart in their results

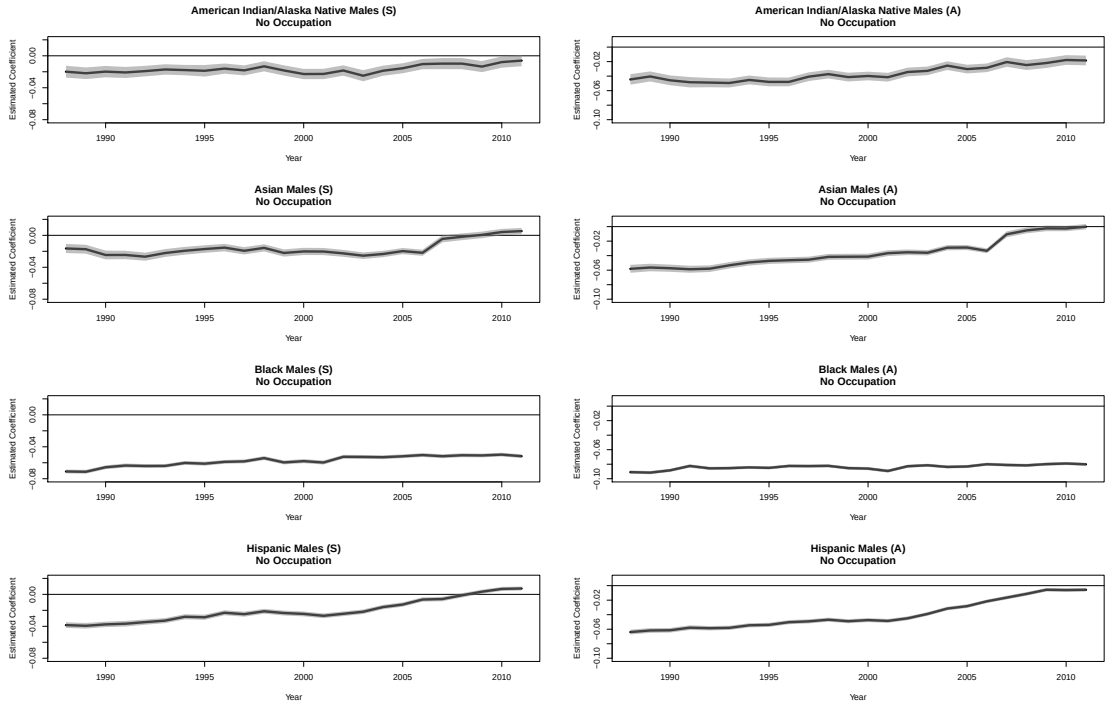

Figure 2: Estimated racial wage gaps (coefficients of race indicators) for yearly males' regressions in synthetic data (left) and confidential, authentic data (right). These models include an indicator for an individual's occupational category code – professional, administrative (omitted category), technical, clerical, other white collar, or blue collar.

is for black female employees, a trend which is replicated in models with more disaggregated occupational information as well. In the confidential dataset, we see a decline in black female earnings relative to whites over time, whereas we see no clear trend in the synthetic dataset estimates. Again, we observe relatively larger magnitude estimates in the confidential datasets.

Figure 2 displays over-time trends and 95% confidence intervals for the racial pay gap for male employees in the federal government. As can be seen, there are relatively similar trends to those reported in the main text of the paper and the results are fairly consistent across the synthetic and the confidential datasets as well. The only sign discrepancy appears to be for Hispanic males, who are estimated to earn more than white males in the latter years of the dataset in the synthetic dataset. This is not replicated in the confidential dataset. There are a couple of minor differences between these results and those reported in the main text. With disaggregated occupational information, we see that Asian male employees appear to reach parity with white male employees in the latter years of the dataset. We also see that in these models, employees who identify as American Indian or Alaska Native also appear to fare less well, never reaching parity, while they do when using disaggregated occupational data.

References

Office of Personnel Management (2014). The guide to data standards, part a: Human resources. Tech. rep., Office of Personnel Management.

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