ASSESSING STATISTICAL DISCLOSURE RISK FOR DIFFERENTIALLY PRIVATE, HIERARCHICAL COUNT DATA, WITH APPLICATION TO THE 2020 U.S. DECENNIAL CENSUS

Zeki Kazan and Jerome P. Reiter

Duke University

Supplementary Material

This supplementary material contains a proposition on the sums of discrete Gaussian random variables, derivations of the full conditionals for the Gibbs sampler in the main text, and discussion of extensions of targeting non-unique individuals and other types of disclosure attacks.

S1 Other Attacks

This section examines other attacks an adversary could perform in addition to the attack focused on in the main text. One class of attacks considers the same attack as in the main text, but with a target who is not unique in group g_1. Another class considers different attacks that are mathematically equivalent to the attack from the main text. A final class considers an attack

by an adversary with substantially more information than the adversaries examined in the main text.

S1.1 Non-unique Individuals

The empirical analysis in Section 4 of the main text focuses on the case where the targeted individual is unique at the lowest level of the hierarchy (the block-level in the 2020 decennial census application). This assumption is not necessary for the methodology. If rather than $x_{1,-t} = 0$ and $x_1 = 1$, we had $x_{1,-t} = m$ and $x_1 = m + 1$—i.e., the adversary knows there are m individuals other than the target with characteristics c—the analysis would be identical to what is presented in the main text. All the probabilities plotted in Section 4 of the main text would stay the same, and all plots involving X_1^* and X_2^* would be shifted by m.

S1.2 Equivalent Attacks

The methodology and empirical evaluations in the main text consider the scenario where an adversary is interested in determining whether individual t has characteristics c. In this section we describe how several other attacks map onto our notation with only the meaning of the prior probability, p, changing. The results from the main text thus can be applied to these other

attacks directly.

An adversary may seek to determine whether the target filled out the census at all. In this setting, we assume the adversary possesses the complete information for all n_1 individuals in g_1, and believes a priori with probability $p_f \in (0, 1)$ that individual t actually filled out the census. The adversary assumes that the other $n_1 - 1$ individuals filled out the census accurately. Thus, the data holder can simply replace p with p_f in the main text and examine the risk from this attack.

Another adversary may seek to determine whether a census respondent lied or made a mistake when completing the census. In this setting, we assume the adversary possesses complete information for all n_1 individuals in g_1, and believes a priori with probability $p_\ell \in (0, 1)$ that individual t reported the correct information. The adversary assumes that the other $n_1 - 1$ individuals filled out the census accurately. Thus, the data holder can simply replace p with p_ℓ in the main text and examine the risk from this attack.

Finally, an adversary may seek to determine an unknown variable. In this setting, we assume the adversary possesses the complete information for all n_1 individuals in g_1, except that they do not know one of the variables for individual t. For example, the adversary could know individual t's race,

HHGQ status, and whether they are of voting age, but not their ethnicity. Let c_e be the true ethnicity of the individual as reported on the decennial census, and let p_v be the prior probability the adversary assigns to individual t having ethnicity c_e. The data holder can simply replace p with p_v in the main text and examine the risk from this attack.

S1.3 Adversaries with Additional Information

The settings in Sections S1.1 and S1.2 presume the adversary only has information on individuals in g_1. Another class of attacks presumes the adversary has information at higher levels of the hierarchy as well. For example, and as suggested by a reviewer, consider an adversary who seeks to determine an unknown variable for target t, say the individual's ethnicity. Let c_e be the true ethnicity of the target (unknown to the adversary), and let c_{-e} be the true characteristics of the target for the other three variables (known to the adversary). We now include the additional assumption that the adversary knows a priori that the target's value of c_{-e} is unique at ℓ levels of the hierarchy. For example, if $\ell = 2$, the target is the only individual in their block group with characteristics c_{-e}; if $\ell = 4$, the target is the only individual in their county with characteristics c_{-e}. Because the target is so distinct, the released noisy counts $X_1^*, \ldots, X_\ell^*$ can be combined

to improve the adversary's posterior, which now has the form

$$\begin{aligned}&\mathbf{P}[X_1 = x_1 \mid X_1^* = x_1^*, \ldots, X_\ell^* = x_\ell^*] \\ &\qquad = \frac{p \prod_{i=1}^{\ell} e^{-\rho_i (x_i^* - x_1)^2}}{p \prod_{i=1}^{\ell} e^{-\rho_i (x_i^* - x_1)^2} + (1-p) \prod_{i=1}^{\ell} e^{-\rho_i (x_i^* - x_{1,-t})^2}}. \qquad \text{(S1.1)}\end{aligned}$$

We can marginalize over the ℓ noisy counts, as in the main text, to compute the marginal posterior the adversary makes the correct conclusion and the corresponding disclosure risk.

Figure 1 plots the marginal posterior and disclosure risk as a function of ℓ for adversaries with different prior beliefs, where the prior parameters are set as in the analysis in the main text (levels 3-6 all have $\rho_i \approx 0.05$; see Table 2 in the main text for details). As expected, the risk increases as a function of ℓ due to the increasing amount of information available to the adversary. For all priors, the increase is most substantial between $\ell = 1$ and $\ell = 2$, since ρ_2 is the largest privacy parameter and thus provides the most accurate release. Overall and in contrast to the findings from the main text, we conclude that, for this type of attack, the hierarchical information can sharpen the adversary's estimates substantially.

To carry out this attack, this adversary requires detailed information across geographical hierarchies. Whether this is a realistic adversary or not is a matter for policymakers to determine in their particular scenarios.

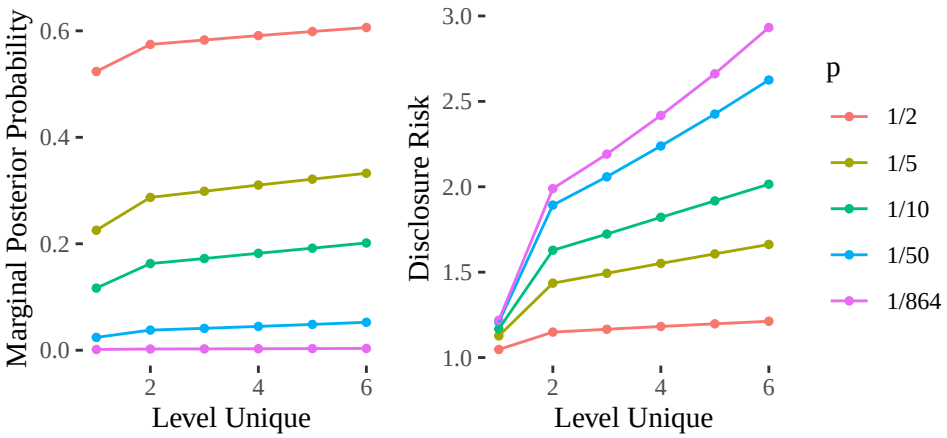

Figure 1: The left panel plots the marginal posterior probability the adversary makes the correct decision that $X_1 = 1$ as a function of the number of levels of the hierarchy at which the target is known to be unique. The right panel plots the corresponding implied disclosure risk. The colors correspond to different adversary prior beliefs; the ρ at each level is the value used by the U. S. Census Bureau in 2020.

S2 Sums of Discrete Gaussian Random Variables

This section focuses on the following proposition.

Proposition 1. *Let* $Z_1, \ldots, Z_n \overset{iid}{\sim} \mathsf{DG}(0, s = 1/(2\rho))$. *Then, for* $\rho < 1$ *and* n *large,* $\sum_{i=1}^{n} Z_i$ *is well approximated by* $\mathsf{DG}(0, s = n/(2\rho))$.

We present an informal proof of this fact, based on empirical results.

To begin, we denote the variance of each Z_i as σ^2. Figure 2 plots σ^2 as a function of both the scale parameter, s, and $\rho = 1/(2s)$. We see that for

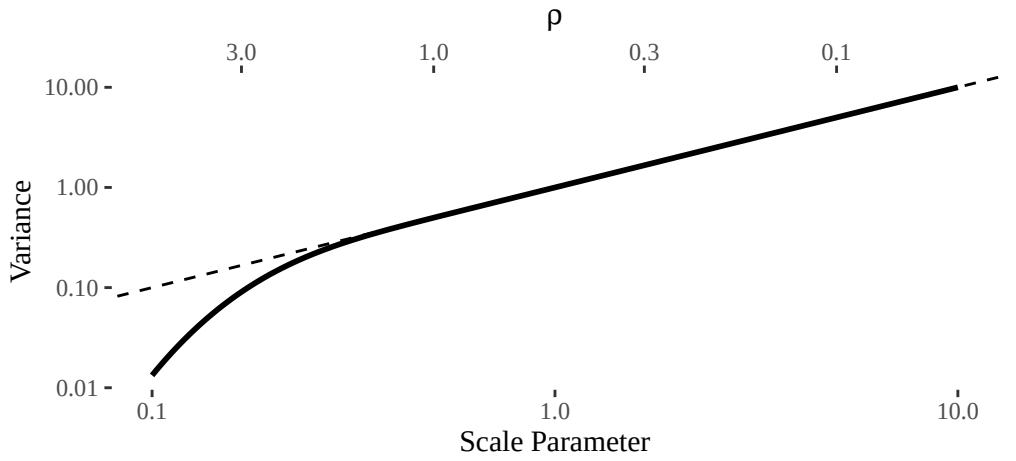

Figure 2: Plot of σ^2, the variance of Z_i (computed to very high precision), as a function of the scale parameter of the discrete Gaussian, s. The dashed line is the line $\sigma^2 = s$; the upper axis presents $\rho = \frac{1}{2s}$ for comparison. Note the log scales.

$\rho < 1$, which corresponds to $s > 0.5$, the approximation $\sigma^2 \approx s = 1/(2\rho)$ is quite accurate. Empirically, $|\sigma^2 - s| < 0.002$ for all $s > 0.5$ and $|\sigma^2 - s| < 10^{-6}$ for all $s > 1$. Thus, we are able to approximate the variance of the Z_i in this range of ρ with s.

For n sufficiently large, we can apply the Central Limit Theorem, which gives the approximation

$$\sum_{i=1}^{n} Z_i \approx \mathcal{N}(0, n\sigma^2). \tag{S2.1}$$

As this distribution is discrete and $n\sigma^2 = n/(2\rho) \gg 1$, it makes sense intuitively to instead use the approximation,

$$\mathcal{N}(0, n\sigma^2) \approx \mathsf{DG}(0, n\sigma^2) = \mathsf{DG}\left(0, \frac{n}{2\rho}\right). \tag{S2.2}$$

Combining the two approximations gives

$$\sum_{i=1}^{n} Z_i \approx \mathsf{DG}\left(0, \frac{n}{2\rho}\right). \tag{S2.3}$$

This approximation is quite accurate in practice. Figure 3 compares the probability mass function of $\mathsf{DG}(0, n/(2\rho))$ to $\sum_{i=1}^{n} Z_i$ for $\rho \in \{1, 0.099\}$ and for $n \in \{5, 27\}$. We note that 0.099 is the value of ρ_1 used in the 2020 census application, and $n = 27$ is used in Section 4.3 of the main text. The approximation does extremely well for these values. Even when n is small and ρ is large, the approximation remains quite accurate.

S3 Full Conditionals for Gibbs Sampler

This section provides the derivations for and forms of the full conditionals for the Gibbs Sampler described in Section 3.2 of the main text. We start with the expression for the posterior distribution of (X_1, X_2) given $\mathcal{D} =$

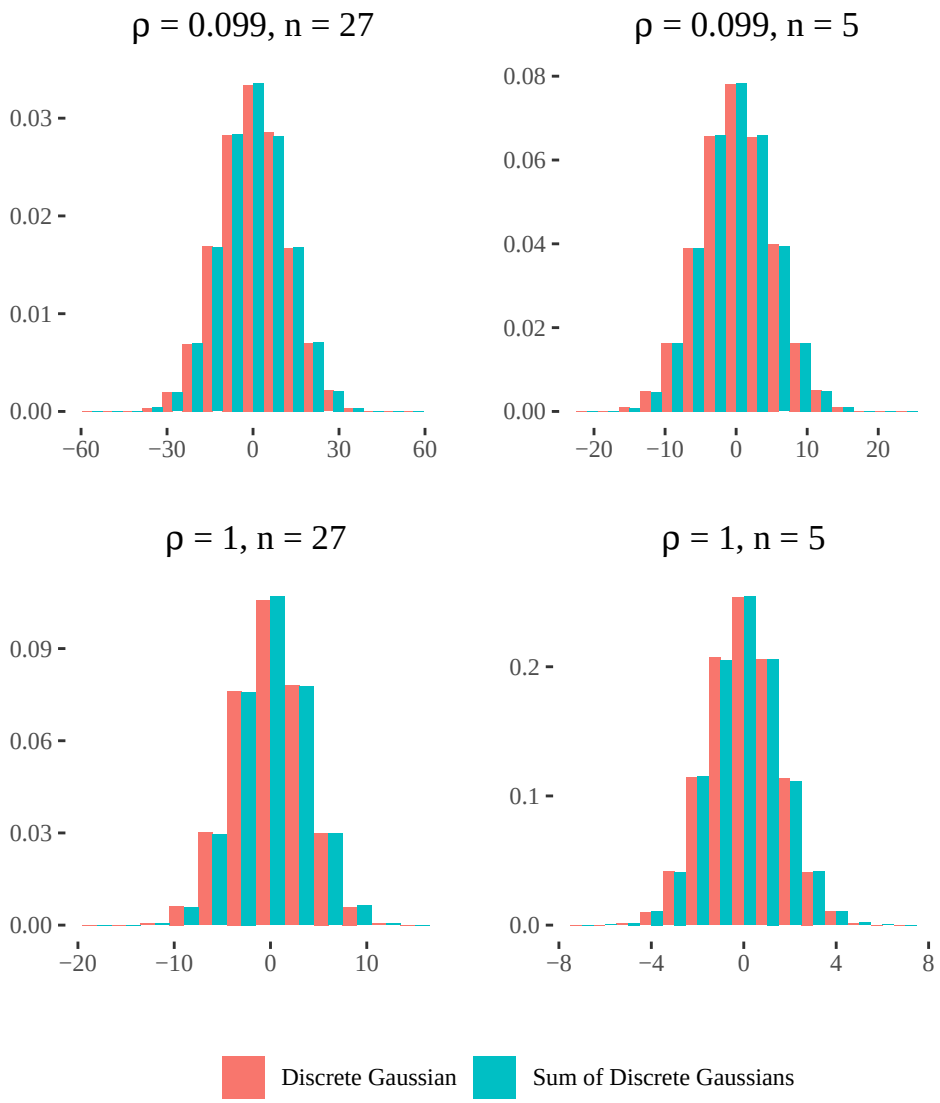

Figure 3: Histograms comparing the probability mass function of $\mathsf{DG}(0, n/(2\rho))$ (in red) to the probability mass function of $\sum_{i=1}^{n} Z_i$ (in blue). Histograms are included for $\rho = 0.099$ on the top row, $\rho = 1$ on the bottom row, $n = 27$ on the left column, and $n = 5$ on the right column.

$(X_1^* = x_1^*, X_2^* = x_2^*, Y_1^* = y_1^*)$. We have

$$\mathbf{P}[X_1 = k_1, X_2 = k_2 \mid X_1^* = x_1^*, X_2^* = x_2^*, Y_1^* = y_1^*] \tag{S3.1}$$

$$\propto \mathbf{P}[X_1^* = x_1^*, X_2^* = x_2^*, Y_1^* = y_1^* \mid X_1 = k_1, X_2 = k_2]$$

$$\mathbf{P}[X_2 = k_2 \mid X_1 = k_1]\, \mathbf{P}[X_1 = k_1] \tag{S3.2}$$

$$\propto \mathbf{P}[X_1^* = x_1^* \mid X_1 = k_1]\, \mathbf{P}[X_2^* = x_2^* \mid X_2 = k_2]$$

$$\mathbf{P}[Y_1^* = y_1^* \mid X_1 = k_1, X_2 = k_2]\, \mathbf{P}[X_2 = k_2 \mid X_1 = k_1]\, \mathbf{P}[X_1 = k_1] \tag{S3.3}$$

$$\propto \exp\left\{-\rho_1 (x_1^* - k_1)^2\right\} \exp\left\{-\rho_2 (x_2^* - k_2)^2\right\}$$

$$\cdot \exp\left\{-\frac{\rho_1}{d}(y_1^* - (k_2 - k_1))^2\right\} \mathbf{1}[k_2 \geq k_1]\, \mathbf{P}[X_1 = k_1]. \tag{S3.4}$$

The full conditional for X_1 is then, for $k_1 \in \{x_{1,-t}, x_{1,-t} + 1\}$ and $k_1 \leq k_2$,

$$\mathbf{P}[X_1 = k_1 \mid X_2 = k_2, X_1^* = x_1^*, X_2^* = x_2^*, Y_1^* = y_1^*] \tag{S3.5}$$

$$\propto \exp\left\{-\rho_1 (x_1^* - k_1)^2\right\} \exp\left\{-\frac{\rho_1}{d}(y_1^* - k_2 + k_1)^2\right\} \mathbf{P}[X_1 = k_1] \tag{S3.6}$$

$$\propto \exp\left\{-\rho_1 (k_1^2 - 2k_1 x_1^*) - \frac{\rho_1}{d}(k_1^2 - 2k_1(k_2 - y_1^*))\right\} \mathbf{P}[X_1 = k_1] \tag{S3.7}$$

$$\propto \exp\left\{-\frac{d+1}{d}\rho_1 k_1^2 + 2\rho_1 (x_1^* + \frac{1}{d}(k_2 - y_1^*))k_1\right\} \mathbf{P}[X_1 = k_1] \tag{S3.8}$$

$$\propto \exp\left\{-\frac{d+1}{d}\rho_1 \left[k_1 - \frac{d x_1^* + (k_2 - y_1^*)}{d+1}\right]^2\right\} \mathbf{P}[X_1 = k_1]. \tag{S3.9}$$

This full conditional is straightforward to sample from.

The full conditional for X_2 is, for $k_2 \in \{k_1, k_1+1, \ldots\}$,

$$\mathbf{P}[X_2 = k_2 \mid X_1 = k_1, X_1^* = x_1^*, X_2^* = x_2^*, Y_1^* = y_1^*] \tag{S3.10}$$

$$\propto \exp\left\{-\rho_2(x_2^* - k_2)^2\right\} \exp\left\{-\frac{\rho_1}{d}(y_1^* + k_1 - k_2)^2\right\} \tag{S3.11}$$

$$\propto \exp\left\{-\rho_2(k_2^2 - 2k_2x_2^*) - \frac{\rho_1}{d}(k_2^2 - 2k_2(y_1^* + k_1))\right\} \tag{S3.12}$$

$$\propto \exp\left\{-\left(\rho_2 + \frac{\rho_1}{d}\right)k_2^2 + 2\left(\rho_2 x_2^* + \frac{\rho_1}{d}(y_1^* + k_1)\right)k_2\right\} \tag{S3.13}$$

$$\propto \exp\left\{-\left(\rho_2 + \frac{\rho_1}{d}\right)\left[k_2 - \frac{\rho_2 x_2^* + \frac{\rho_1}{d}(y_1^* + k_1)}{\rho_2 + \frac{\rho_1}{d}}\right]^2\right\}. \tag{S3.14}$$

This is a truncated discrete Gaussian distribution centered at $\frac{\rho_2 x_2^* + \frac{\rho_1}{d}(y_1^* + k_1)}{\rho_2 + \frac{\rho_1}{d}}$. It can be easily sampled over a grid, since the tails of the distribution decay rapidly. Using these full conditionals, the adversary can sample from the posterior distribution and examine the marginal posterior distribution for X_1.

S4 Prior Sensitivity

This section examines how sensitive the analysis producing Figure 4 in Section 4.3 of the main text is to the choice of the adversary's prior on $X_2 \mid X_1$. In particular, since the prior

$$(X_2 \mid X_1 = k_1) \sim \mathsf{Unif}(\{k_1, k_1+1, \ldots\}), \qquad k_1 \in \{0, 1\}, \tag{S4.1}$$

is an improper probability distribution with unbounded support, it may unduly favor values that are practically implausible. To determine whether this is the case, we re-do the analysis producing Figure 4 with a selection of other priors and examine how the conclusions change. We assume throughout that the prior probability for X_1 is $p = 1/2$, the number of other blocks is $d = 27$, and the true counts are $x_1 = x_2 = 1$. Figure 4 from the main text is reproduced as the top panel of Figure 4, for ease of comparison.

We begin by examining a variation on the uniform prior used in the main text. Suppose that an adversary, utilizing information from auxiliary data sources, knows that the number of individuals in block group g_2 with characteristics c is at most 10. A reasonable prior might then be

$$(X_2 \mid X_1 = k_1) \sim \mathsf{Unif}(\{k_1, \ldots, 10\}), \qquad k_1 \in \{0, 1\}. \tag{S4.2}$$

This prior has bounded support and does not place any prior probability on very extreme values for X_2. The results for this prior are presented on the bottom panel of Figure 4. We do not observe a substantial change between the truncated and non-truncated priors; for both, the adversary makes the correct decision 59% of the time. The lack of change is likely due to the fact that, as suggested by Table 8 in the main text, the unbounded uniform prior allows the data to rule out implausible values away from x_2.

Another interesting comparison is to the case where the adversary

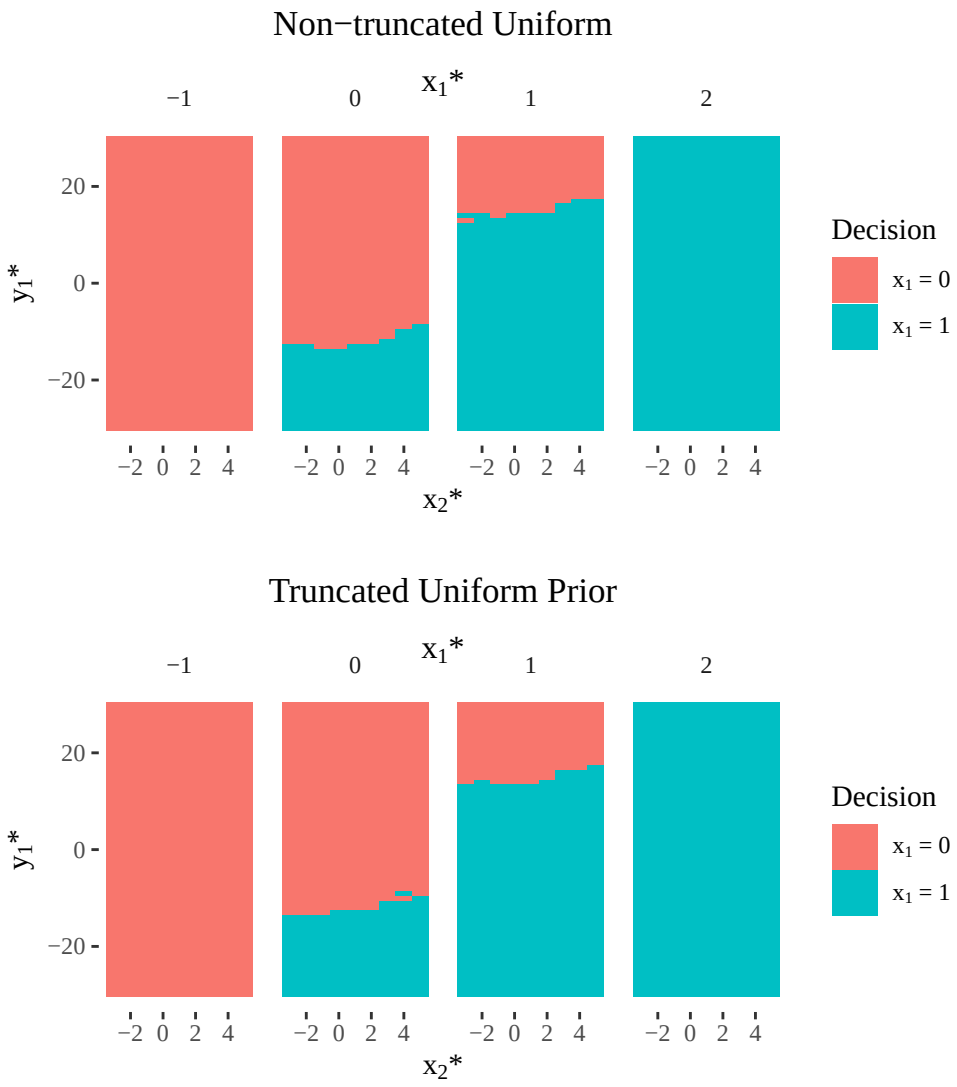

Figure 4: Adversary's decision under 0-1 loss for each combination of x_1^*, x_2^*, and y_1^*. The top plot reproduces Figure 4 from the main text, while the bottom plot uses the prior $X_2 \mid X_1 \sim \mathsf{Unif}(\{k_1, \ldots, 10\})$. Privacy parameters are set as in the census application, $p = 1/2$, and $d = 27$. 10^3 MCMC draws are taken for each combination in most cases. When the posterior probability $X_1 = 1$ is close to 0.5, the number of MCMC draws is increased to 2.5×10^5.

knows a priori that $x_2 = 1$. This corresponds to a prior with all the probability mass on $X_2 = 1$. As a consequence, the adversary's decision about x_1 does not depend on x_2^*. The results for this prior are presented in the top panel of Figure 5. We do not observe a substantial change from the previous two figures; the adversary still makes the correct decision 59% of the time, even with perfect knowledge at the second level. The agreement between this result and the uniform priors suggests that the uniform priors are not biasing the results to any substantial degree.

One might take the above as evidence that the choice of prior for $X_2 \mid X_1$ is of little importance. We demonstrate that this is not the case by examining the results under a poorly specified prior. Suppose that the adversary incorrectly believes that $x_2 = 25$ and places a prior with all the probability mass on $X_2 = 25$. The results for this prior are presented in the bottom panel of Figure 5. We observe a substantial change between this plot and the previous three: the adversary now makes the correct decision 74% of the time. But consider the counterfactual where in truth $x_1 = 0$ (and $x_2 = 0$). Now the misspecified prior leads the adversary astray, and they make the correct decision only 42% of the time (the distribution of X_1^* and X_2^* change in the counterfactual, so the probability is not simply $100\% - 74\%$). Evidently, an inaccurate prior can impact the results, possibly

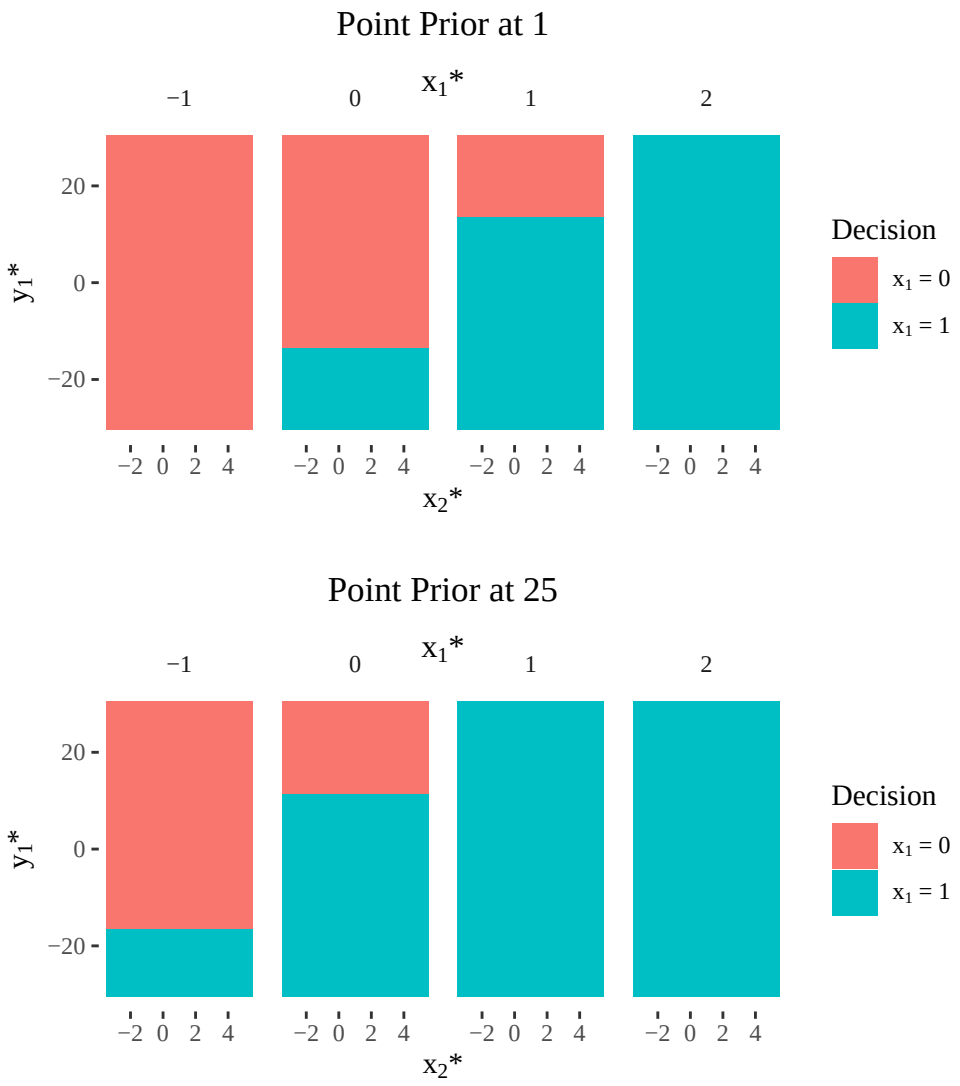

Figure 5: Adversary's decision under 0-1 loss for each combination of x_1^*, x_2^*, and y_1^*. Priors are of the form $\mathbf{P}[X_2 = k_2 \mid X_1 = k_1] = 1$ for $k_2 = 1$ (top) and $k_2 = 25$ (bottom). Privacy parameters are set as in the census application, $p = 1/2$, and $d = 27$. 10^3 MCMC draws are taken for each combination in most cases. When the posterior probability $X_1 = 1$ is close to 0.5, the number of MCMC draws is increased to 10^6.

to the detriment or benefit of the adversary depending on the value of x_1. Of course, in practical contexts the adversary does not know whether they benefit or suffer from an informative prior. Given that the uniform prior (with support that includes the true count) allows the distributions of the noisy counts to fully determine the posterior probability computations, it appears to be a sensible choice when evaluating statistical disclosure risks.

S5 Composition of Risk

In this section, we briefly examine how the risk measures from the main text behave under composition. That is, if the Census Bureau were to perform a second data release, how would the risks from the two releases combine? Let X^*_{1i} be the released noisy count from the i^{th} release and x^*_{1i} be the corresponding observed value. Recall that the disclosure risk from the first release is

$$R'(x^*_{11}) = \frac{\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]}{\mathbf{P}[X_1 = x_1]}. \tag{S5.1}$$

We can similarly examine the disclosure risk from the second release. Assuming the releases are sequential, the adversary will have already observed x^*_{11}, so their prior probability for the second release corresponds exactly to

their posterior from the first release. That is,

$$R'(x^*_{12} \mid x^*_{11}) = \frac{\mathbf{P}[X_1 = x_1 \mid X^*_{12} = x^*_{12}, X^*_{11} = x^*_{11}]}{\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]}. \tag{S5.2}$$

This quantity is analogous to $R'(x^*_{11})$ and, in practice, the posterior in the numerator can be decomposed as follows via Bayes Theorem:

$$\begin{aligned}
&\mathbf{P}[X_1 = x_1 \mid X^*_{12} = x^*_{12}, X^*_{11} = x^*_{11}] \\
&\quad = \frac{\mathbf{P}[X^*_{12} = x^*_{12} \mid X_1 = x_1, X^*_{11} = x^*_{11}]\,\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]}{\mathbf{P}[X^*_{12} = x^*_{12} \mid X^*_{11} = x^*_{11}]} && \text{(S5.3)} \\
&\quad = \frac{\mathbf{P}[X^*_{12} = x^*_{12} \mid X_1 = x_1]\,\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]}{\sum_{k_1 = x_{1,-t}}^{x_{1,-t}+1} \mathbf{P}[X^*_{12} = x^*_{12} \mid X_1 = k_1]\,\mathbf{P}[X_1 = k_1 \mid X^*_{11} = x^*_{11}]}. && \text{(S5.4)}
\end{aligned}$$

The latter equality assumes that the mechanism for releasing X^*_{12} does not depend on the observed x^*_{11}. (S5.4) has a form identical to the form of $\mathbf{P}[X_1 = x_1 \mid X^*_1 = x^*_1]$ in the main text, except that the prior is conditional on the observed x^*_{11} from the first release. This means that the analysis of the second release can proceed exactly as the first with the only difference being an "updated" prior.

The total risk from the two releases is then

$$\begin{aligned}
R'(x^*_{11}, x^*_{12}) &= \frac{\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}, X^*_{12} = x^*_{12}]}{\mathbf{P}[X_1 = x_1]} && \text{(S5.5)} \\
&= \frac{\mathbf{P}[X_1 = x_1 \mid X^*_{12} = x^*_{12}, X^*_{11} = x^*_{11}]}{\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]} \cdot \frac{\mathbf{P}[X_1 = x_1 \mid X^*_{11} = x^*_{11}]}{\mathbf{P}[X_1 = x_1]} && \text{(S5.6)} \\
&= R'(x^*_{12} \mid x^*_{11}) R'(x^*_{11}). && \text{(S5.7)}
\end{aligned}$$

This argument generalizes to an arbitrary number of releases. Letting m be the total number of releases, the total risk composes as

$$R'(x^*_{11}, \ldots, x^*_{1m}) = R'(x^*_{1m} \mid x^*_{11}, \ldots, x^*_{1,m-1}) \cdots R'(x^*_{11}). \tag{S5.8}$$

Thus, the cumulative risk is simply the product of the risk from each release. The generalized marginal risk and generalized probability the adversary makes the correct decision are straightforward to compute from the generalized R'.

S6 The Effect of Post-Processing

In this section, we illustrate how a post-processing step could affect the risk analysis in this article. Our intent is not to give a complete treatment of this but rather to provide a rough intuition. Thus, in this analysis, we make a substantial number of simplifying assumptions about the adversary and the way the post-processing is performed compared to the TopDown algorithm used for the 2020 decennial census data.

To begin, we outline our illustrative post-processing algorithm. Let $\tilde{X}_1, \tilde{X}_2$ be the post-processed counts corresponding to X_1, X_2 and $\tilde{x}_1, \tilde{x}_2$ be their observed values. Similarly, let $\tilde{Y}_1^{(1)}, \ldots, \tilde{Y}_1^{(d)}$ be the post-processed counts corresponding to $Y_1^{(1)}, \ldots, Y_1^{(d)}$ and $\tilde{y}_1^{(1)}, \ldots, \tilde{y}_1^{(d)}$ be their observed

values. We define the post-processing algorithm at the block level as follows. Taking $\tilde{x}_2$ as fixed, we enforce the aggregation constraint $\tilde{x}_2 = \tilde{x}_1 + \sum_{i=1}^{d} \tilde{y}_1^{(i)}$, while minimizing the sum of squared deviations from the noisy counts:

$$\underset{\tilde{x}_1, \tilde{y}_1^{(1)}, \ldots, \tilde{y}_1^{(d)}}{\operatorname{argmin}} \left\{ (\tilde{x}_1 - x_1^*)^2 + \sum_{i=1}^{d} (\tilde{y}_1^{(i)} - y_1^{(i)*})^2 \right\}. \tag{S6.1}$$

The post-processing algorithm used in the TopDown algorithm enforces several aggregation constraints and minimizes a weighted sum of squared deviations involving more quantities, so this is a substantial simplification, but one that we expect to roughly approximate the effects of the true algorithm. Letting x_1^* and y_1^* be the observed noisy counts corresponding to X_1 and Y_1, this simplified problem has a closed form solution, which we denote $\bar{x}_1$:

$$\bar{x}_1 = \frac{d x_1^* + (\tilde{x}_2 - y_1^*)}{d+1}. \tag{S6.2}$$

It is possible for $\bar{x}_1$ to be outside the range $[0, \tilde{x}_2]$ or to be non-integer valued. To correct for this, we truncate the solution to be in the correct range and round to the nearest integer. The complete post-processing algorithm includes a non-negativity constraint in the optimization and performs a second controlled rounding step, although we expect this change to have a limited effect for our illustration. We denote the final solution to the

optimization as $f(x_1^*, y_1^*, \tilde{x}_2)$, which is given by

$$f(x_1^*, y_1^*, \tilde{x}_2) = \begin{cases} 0, & \text{if } \bar{x}_1 < 0; \\ \tilde{x}_2, & \text{if } \bar{x}_1 > \tilde{x}_2; \\ [\bar{x}_1], & \text{otherwise.} \end{cases} \tag{S6.3}$$

We now return to the perspective of the adversary. From the above, the likelihood from the post-processing step is simply an indicator variable

$$\mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1^* = x_1^*, Y_1^* = y_1^*, \tilde{X}_2 = \tilde{x}_2] = \mathbf{1}[\tilde{x}_1 = f(x_1^*, y_1^*, \tilde{x}_2)]. \tag{S6.4}$$

The full likelihood is then, assuming that the adversary considers $\tilde{x}_1, \tilde{x}_2$ and not $\tilde{y}_1^{(1)}, \ldots, \tilde{y}_1^{(d)}$,

$$\begin{aligned} &\mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1 = k_1, \tilde{X}_2 = \tilde{x}_2] \\ &\quad = \sum_{x_1^*=-\infty}^{\infty} \sum_{y_1^*=-\infty}^{\infty} \mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1^* = x_1^*, Y_1^* = y_1^*, \tilde{X}_2 = \tilde{x}_2] \\ &\qquad\qquad \mathbf{P}[X_1^* = x_1^* \mid X_1 = k_1]\, \mathbf{P}[Y_1^* = y_1^* \mid X_1 = k_1] \qquad \text{(S6.5)} \\ &\quad = \sum_{x_1^*=-\infty}^{\infty} \sum_{y_1^*=-\infty}^{\infty} \mathbf{1}[\tilde{x}_1 = f(x_1^*, y_1^*, \tilde{x}_2)]\, \mathbf{P}[X_1^* = x_1^* \mid X_1 = k_1] \\ &\qquad\qquad \mathbf{P}[Y_1^* = y_1^* \mid X_1 = k_1]. \qquad \text{(S6.6)} \end{aligned}$$

To simplify $\mathbf{P}[Y_1^* = y_1^* \mid X_1 = k_1]$, we assume the adversary knows x_2 exactly a priori, in addition to making the approximation from Section S2. Assuming as in the main text that the true $x_1 = x_2 = 1$, the known count

is $x_{1,-t} = 0$, and the adversary's prior on X_1 is Bernoulli with parameter p, the posterior probability the adversary makes the correct decision is

$$\mathbf{P}[X_1 = 1 \mid \tilde{X}_1 = \tilde{x}_1, \tilde{X}_2 = \tilde{X}_2] = \frac{\mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1 = 1, \tilde{X}_2 = \tilde{x}_2]\, p}{\mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1 = 1, \tilde{X}_2 = \tilde{x}_2]\, p + \mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid X_1 = 0, \tilde{X}_2 = \tilde{x}_2](1-p)}. \tag{S6.7}$$

Finally, for comparison to the results from the main text, we can marginalize out $\tilde{X}_1$ from the posterior:

$$\mathbf{P}[X_1 = 1 \mid x_1 = 1, \tilde{X}_2 = \tilde{x}_2] = \sum_{\tilde{x}_1=0}^{\tilde{x}_2} \mathbf{P}[X_1 = 1 \mid \tilde{X}_1 = \tilde{x}_1, \tilde{X}_2 = \tilde{x}_2] \mathbf{P}[\tilde{X}_1 = \tilde{x}_1 \mid x_1 = 1, \tilde{X}_2 = \tilde{x}_2]. \tag{S6.8}$$

Note that the result will vary with $\tilde{x}_2$.

We now examine whether, on average, releasing the counts with post-processing will have lower disclosure risk than releasing the counts without post-processing. The first panel of Figure 6 compares the marginal disclosure risks in the case where $\rho_1 \approx 0.099$ for various values of $\tilde{x}_2$. We find that the marginal risk with post-processing is bounded above by the marginal risk without post-processing, as expected. Larger values of $\tilde{x}_2$ give risks closer to the bound, which makes sense intuitively; larger values of $\tilde{x}_2$ allow for a larger range of possible observed $\tilde{x}_1$, which will make it easier for the adversary to "work backward" to x_1^*. The second panel of Figure 6 com-

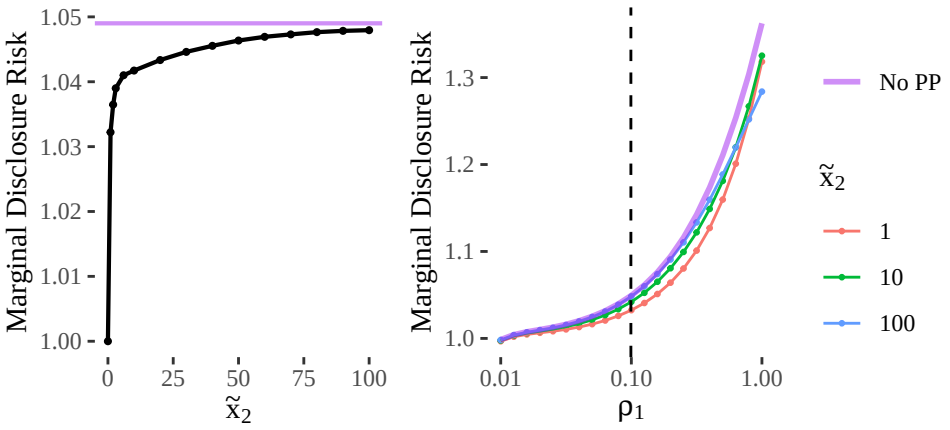

Figure 6: The left panel plots the marginal disclosure risk from (S6.8) as a function of $\tilde{x}_2$ when $\rho_1 \approx 0.099$. The right panel plots the marginal disclosure risk from (S6.8) as a function of ρ_1 colored by $\tilde{x}_2 \in \{1, 10, 100\}$. The dashed line represents $\rho_1 \approx 0.099$. In both panels, the purple line represents the marginal disclosure risk without post-processing. Both set $p = 1/2$ and assume the adversary knows that $x_2 = 1$.

pares the marginal disclosure risks as a function of ρ_1 for a selection of $\tilde{x}_2$. We observe a similar effect, with the marginal risk without post-processing providing an upper bound on the marginal risk with post-processing. In general, we find that the bound is fairly tight; the reduction in disclosure risk due to the post-processing is minor (given the simplifications and assumptions we make).