Statistics-Friendly Confidentiality Protection for Establishment Data, with Applications to the QCEW

Kaitlyn Webb, Prottay Protivash, John Durrell, Aleksandra Slavković, Daniel Kifer
Pennsylvania State University

Daniell Toth
National Institute of Statistical Sciences

Abstract

Confidentiality for business data is an understudied area of disclosure avoidance, where legacy methods struggle to provide acceptable results. Standard formal privacy techniques for person-level data, like differential privacy, are designed to protect against membership inference and hence do not provide suitable confidentiality/utility trade-offs due to the highly skewed nature of business data and because extreme outlier records are often important contributors to query answers. Prior proposals, therefore, took a personalized differential privacy approach that allowed privacy parameters to degrade for the outlying records – larger establishments get weaker membership inference guarantees. However, providing guarantees to some entities that are strictly weaker than guarantees for others is problematic from a policy standpoint. In this paper, we propose a novel confidentiality framework for business data with a focus on interpretability for policy makers. Instead of protecting against membership inference, which is often not a concern in business data, we protect against attribute inferences that are too precise. In our framework, data curators specify a neighbor function that is used to define uncertainty interval bands around an establishment's attribute values and the privacy parameters govern the strength of indistinguishability between values within the same uncertainty interval. We propose two query-answering mechanisms under this framework and evaluate them on: (1) a confidential Quarterly Census of Employment and Wages (QCEW) dataset produced by the U.S. Bureau of Labor Statistics (this was done through a cooperative agreement), and (2) a substitute dataset that we created from public sources (and will publicly release).

Keywords

data privacy, formal privacy, establishment data

1 Introduction

Business data provide crucial snapshots of various sectors of a country's economy. The collection and dissemination of business statistics is of great value in both the public and private sectors. As with person-level data, there are strong confidentiality concerns, but the requirements are not the same. Protections for person-level data are designed to mask records that are outliers, as individuals with uncommon characteristics can be vulnerable to disclosure and harassment. However, the "outliers" in business data are the large companies whose contributions to employment and wage statistics are important to study. In business data, the analogue to a person is an *establishment*, which is a single physical location where business activity occurs (e.g., a specific restaurant in a restaurant chain).

Disclosure avoidance for person-level data has advanced with the rise of differential privacy (DP) [17], but much less work has focused on protecting establishment data. The dominant method in practice is cell suppression [3]. It is used, for example, to protect the Quarterly Census of Employment and Wages (QCEW), the flagship economics dataset produced by the Bureau of Labor Statistics (BLS) [11, 47]. Given a set of desired statistics, cell suppression redacts a subset of them so that a weak attacker (who can only use linear algebra and has no access to external data) cannot derive sensitive information about a business establishment. Despite this weak adversarial model, cell suppression suffers from severe loss of utility: $\approx 60\%$ out of roughly 3.6 million statistics in the QCEW tables are suppressed [11, 45, 52]. Hence there is a need for new technology with rigorous formal privacy guarantees and better utility.

This problem has received little attention from the formal privacy community [21, 23, 43, 46]. Viable solutions should (1) provide confidentiality guarantees that are **interpretable** to mathematical statisticians and policy makers in statistical agencies and (2) should provide **flexibility** that allows policy makers to specify how the confidentiality needs for large establishments differ from those of small establishments. Earlier work [23, 46] did not support any customization. Recent work [21, 43] has made progress in this direction, by applying the personalized DP philosophy [26]: every establishment should have its own privacy parameters. Specifically, [21, 43] propose schemes for protecting membership inference (i.e., whether an establishment was included in the data) where larger businesses get worse privacy parameters than smaller businesses, leading to the interpretation that some businesses get strictly weaker protections than others – a position that can be problematic for policy-makers. In many cases, membership inference, especially for large businesses, is not a concern – they need to be included in the data for reliable economic estimates, and their existence is public knowledge. For instance, the QCEW is a census and collects data about all businesses.

Instead, we design our approach based on the metric DP philosophy [10]: a entity should have plausible deniability about whether its true value is x or some "nearby" value, where the concept of "nearby" is defined using an application-specific metric. We argue that smaller establishments need stronger guarantees on the relative error of an attacker's inference about them, while large establishments need stronger guarantees on the absolute error, where absolute error refers to the standard error of the attacker's estimate and relative error is the absolute error over the confidential value. We derive mathematical conditions on metrics that will guarantee such an interpretation of the confidentiality semantics. The result is

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
Proceedings on Privacy Enhancing Technologies 2026(3), 646–687
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.56553/popets-2026-0100

that the specification of *what* needs plausible deniability becomes customized to an establishment's size and is separated from the quantification of *how strong* the plausible deniability is (i.e., how hard is it to distinguish x from a nearby value y).

We then propose two query answering mechanisms for this framework. The first mechanism provides query answers that have a data-dependent variance that cannot be released. The second mechanism piggybacks on the first one to answer additional queries with a data-dependent variance that is completely safe to release. This property is very useful to statisticians since they need to assess how much they can trust a noisy query answer. We evaluate our techniques on a real confidential high-stakes dataset, the QCEW, by leveraging a cooperative agreement with the Bureau of Labor Statistics. We also evaluate our techniques on a business establishment dataset that we synthesized from public data sources. We plan to share this dataset after publication in order to help support future research on this topic. To summarize, our contributions are:

- Inspired by Gaussian Differential Privacy [15] and Metric DP [10], we propose a privacy definition for business establishment data that we call Gaussian Establishment DP (or gedp for short). Within the Metric DP framework, we identify a class of metrics suitable for business data and which formalize the plausible deniable needs of establishments of different sizes. We show how the metrics can be interpreted as intended uncertainty intervals around an establishment's true values, emphasizing more relative error uncertainty for smaller establishments and more absolute error uncertainty for larger ones. To measure the strength of the plausible deniability, we use privacy accounting based on Gaussian Differential Privacy as it directly measures resilience to statistical attacks: how much success can the best hypothesis test achieve when trying to distinguish between two plausible values x and y that are in each other's intended uncertainty intervals?
- We advocate for a specific choice of a metric, based on the square root transform, due to its superior privacy and statistical properties compared to the traditional business data approaches that aim to give all establishments equal relative error protections.
- We introduce two mechanisms for Gaussian Establishment DP, the ψ-mechanism and pnc-mechanism, for answering group-by queries under this framework. The ψ-mechanism has data-dependent variance that cannot be revealed exactly. However, the pnc-mechanism piggybacks on top of the ψ-mechanism to provide answers to additional queries, using data-dependent Gaussian noise whose variance is completely safe to release. This is an important property for statisticians.
- We evaluate our mechanisms on real-world data (a historical, confidential QCEW dataset accessed through a cooperative agreement) and on a publicly shareable partially synthetic dataset that we reconstructed from partially redacted statistics published by the U.S. Census Bureau.

Note: our methods apply to skewed data in general. We focus on business data for concreteness to better explain the confidentiality requirements and to spur interest in the topic – it is an untapped subfield with huge practical and research potential.

The paper is organized as follows. We provide background information on the QCEW in Section 2 since the design of formal privacy definitions for establishment data require an understanding of the confidentiality concerns and characteristics of such data. In Section 3, we present notation and technical background on Gaussian DP [15] (which motivates our framework) and discuss related work in Section 4. In Section 5, we propose the privacy definition Gaussian Establishment DP. We present two query-answering mechanisms in Section 6. In Section 7, we explain how to convert noisy query answers into privacy preserving microdata. In Section 8, we present experiments on confidential real and publicly shareable partially synthetic data . We conclude and outline future work in Section 9. Proofs can be found in the appendix.

2 Background on the QCEW

The Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW) data is an economically and politically important dataset containing wage and employment information about businesses in the United States. The data provided by QCEW are crucial for measuring labor trends and industry developments, provide a sample-frame for all BLS establishment surveys, and are used to benchmark the Current Employment Statistics survey, which is a primary economic indicator used by the federal reserve [5]. QCEW data has also been used for a wide range of analyses that have helped economists understand the effects of minimum wage, the non-profit sector and the effect of the COVID-19 pandemic on employment levels [14, 30, 41].

The QCEW is collected each quarter under a cooperative program between BLS and the State Employment Security Agencies. States collect the data from unemployment insurance (UI) administrative records which firms are legally obligated to submit quarterly. The primary unit of analysis is an *establishment*, which is a single physical location where one predominant activity, classified by a single industry code, occurs [40]. Meanwhile, a *firm* is a business made up of one or more establishments. All of the state-collected, administrative data, covering over 10 million establishments, are sent to the BLS [11, 47]. Since the UI data are collected for all U.S. workers covered by state unemployment insurance laws or by the Unemployment Compensation for Federal Employees program, they represent a virtual census of employment and wages, providing a complete overview of the national economic picture.

BLS supplements the states' UI data collection by conducting an annual refiling survey to ensure each establishment is labeled under the correct industry code and further cleans the data. The data from each state are then sent back to the state along with BLS produced aggregated estimates protected by a disclosure limitation method. **Each state retains ownership over its own data under this arrangement.** Thus, there are *54 different data owners* (50 states, the District of Columbia, Puerto Rico, Virgin Islands, and BLS), each covered by different laws and policies. In addition to the data published by BLS, each data owner can separately publish statistics about their own parts of the data. Under the current cell suppression methodology, it is feasible that data published by BLS, when combined with a separate release produced by a state, could be used to undo protections for a different state. This is known as a *composition attack* [22]. Our motivation for studying a formal privacy alternative to cell suppression is motivated by this problem, since formal privacy methods for person-level data are known to resist composition attacks.

Data Schema: Quarterly data about each establishment include (1) location (e.g., county and state), (2) total quarterly wages, (3) three monthly employment counts, (4) an ownership type (private, local government, etc.) and (5) a six-digit industry code known as the North American Industry Classification System (NAICS) code. The NAICS code [49] is hierarchically structured. For example, "construction" has the NAICS prefix 23, "construction of buildings" has the prefix 236 and the NAICS code for "new multifamily housing construction" is 236116. **The wages and employment data are considered confidential while the rest of the attributes (county, NAICS code, ownership type) are considered public.**

Dissemination: The QCEW establishment-level data are published in aggregated form, mostly as group-by sum queries, also known as marginals and are identified by an aggregation level code [48]; e.g., aggregation level 15 ("National, by NAICS 3-digit – by ownership sector") provides the national-level total quarterly wages and monthly employment totals for each combination of 3-digit NAICS code and ownership type (private, foreign government, state government, etc.) while aggregation level 55 provides the same information, but broken down by state, and 75 breaks it down by county. BLS also receives requests for custom tabulations. Thus a highly desirable property for a disclosure avoidance system is the ability to convert noisy query answers into *confidentiality-protected microdata* from which those custom tabulations can be computed.

Skewness: A major difference between establishment data and person-level data is the skewness of the data – there are outliers whose influence on query answers is important to preserve. Establishments can be very small, such as an ice-cream shop with a handful of employees, or very large, like Disneyland Resort, which employs approximately 36,000 employees in the Orange County, CA [37].

The formal privacy approach for person-level data is to inject enough noise into query answers to mask the effect of any entity. Could this idea be directly applied to establishment data by defining an entity to be a person, or by defining it to be an entire establishment? The naive idea of masking the effect of one *person* in establishment data is too weak because it may reveal employment and wage information about a business that is too precise. For example Disneyland [37] and Microsoft [31] publicly provide rough estimates of their employment counts (36,000 and 125,000, respectively), that are rounded to the thousands, indicating they may need larger protections. On the other hand, injecting enough noise to mask the presence an *establishment* like Disneyland Resort would render regional economic statistics meaningless.

Proposals for addressing this skewness issue [21, 43] took a personalized DP approach [26] in which larger establishments receive worse privacy parameters than smaller establishments and therefore are given weaker privacy semantics. This can be difficult to justify from a public policy standpoint. Our alternative proposal is given in Section 5, after discussing background and related work.

3 Notation and Background

The notation used in the paper is summarized in Table 1. A dataset $\mathcal{D}$ contains records about establishments $\mathcal{E}_1, \ldots, \mathcal{E}_n$. Establishments have **public attributes** $A_1, \ldots, A_{k_p}$ and **confidential attributes** $C_1, \ldots, C_{k_c}$. The confidential attributes are numeric and often nonnegative (e.g., number of employees). To simplify the discussion, we will assume that the domain of the confidential attributes is nonnegative, but in Section 5.2.1, we explain how this condition can be completely eliminated.

Table 1: Table of Notation

Φ:	Gaussian CDF
μ:	Gaussian DP and gedp Privacy Parameter
$\mathcal{D}$:	Data set of records for establishments $\mathcal{E}_1, \ldots, \mathcal{E}_n$
$A_1, \ldots, A_{k_p}$:	Public attributes of an establishment
$C_1, \ldots, C_{k_c}$:	Confidential attributes of an establishment
r:	Establishment record. $r[C_i]$ is the value for C_i
ψ:	Neighbor function (Definition 5.2).
γ:	Closeness parameter (Definition 5.2).
$\mathcal{M}$:	Mechanism.
$\mathcal{I}_{\psi,\gamma}(x)$:	Uncertainty interval around x induced by ψ, γ.

We use a_i (resp., c_i) to refer to specific values of attribute A_i (resp., C_i). We can represent an establishment record r as a tuple $r = (a_1, \ldots, a_{k_p}, c_1, \ldots, c_{k_c})$ and refer to attributes within a record using indexing notation; e.g., the value of attribute C_1 in r is $r[C_1]$.

A *mechanism* $\mathcal{M}$ is a randomized algorithm whose input is confidential data and whose output is intended for public distribution. Our framework for protecting establishment data builds upon ideas from the Metric DP framework [10] and Gaussian DP privacy accounting [15], which are defined as follows:

Definition 3.1 (Metric DP [10])**.** Given a metric d_X over datasets, a mechanism $\mathcal{M}$ satisfies (metric) d_X-privacy if for all $x \in range(\mathcal{M})$ and all pairs of datasets $\mathcal{D}_1, \mathcal{D}_2$,

$$P(\mathcal{M}(\mathcal{D}_1) = x) \leq e^{d_X(\mathcal{D}_1, \mathcal{D}_2)} P(\mathcal{M}(\mathcal{D}_2) = x).$$

Note: the Metric DP framework does not specify how to choose a metric d_X. Thus, users of this framework must first identify a metric that is suitable for their application and then design efficient mechanisms.

If d_H is the Hamming distance (number of records on which two datasets differ) and $\epsilon > 0$, then using the metric ϵd_H recovers the original definition of pure ϵ-DP[17]. The intuitive meaning behind this definition is that if two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are close under a metric d_X, and one of them was the input to $\mathcal{M}$, then an attacker who has the output of $\mathcal{M}$ will have difficulty determining whether the input was $\mathcal{D}_1$ or $\mathcal{D}_2$.

The current state of the art in specifying what it means for an attacker to have difficulty in determining which dataset was the input is known as Gaussian DP and f-DP. It measures difficulty in terms of the best possible hypothesis test that uses the output of $\mathcal{M}$. It is defined as follows (here Φ denote the cumulative distribution function of the standard normal distribution):

Definition 3.2 (f-DP and μ-Gaussian DP [15])**.** Let f be a continuous, concave, non-decreasing function[1] such that $f(x) \geq x$. A mechanism $\mathcal{M}$ satisfies f-DP if for all pairs of datasets $\mathcal{D}_1, \mathcal{D}_2$ that differ on the value of one record and for all $S \subseteq range(\mathcal{M})$,

$$P(\mathcal{M}(\mathcal{D}_1) \in S) \leq x \Rightarrow P(\mathcal{M}(\mathcal{D}_2) \in S) \leq f(x).$$

[1]Note, this formulation is equivalent to that of Dong et al. [15] using trade-off function $1 - f$. The formulation presented here is easier to work with in our opinion.

In particular, when $f(x) \equiv \Phi(\Phi^{-1}(x) + \mu)$ for some number $\mu \geq 0$, then $\mathcal{M}$ satisfies μ-Gaussian DP or μ-GDP for short. Note that the privacy requirement on such an $\mathcal{M}$ is equivalent to the condition:

$$\Phi^{-1}(P(\mathcal{M}(\mathcal{D}_2) \in S)) \leq \Phi^{-1}(P(\mathcal{M}(\mathcal{D}_1) \in S)) + \mu.$$

Intuitively, f-DP means that if an attacker is trying to distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$ based on the output of $\mathcal{M}$, then no matter what method the attacker uses, if its false positive rate is x then the true positive rate is $\leq f(x)$. Similarly, if the false negative rate is y then the true negative rate is $\leq f(y)$. Thus the function f is a trade-off between true and false positive rates. Note that when $f(x) = x$ then the attacker does no better than random guessing, which represents perfect confidentiality. From a statistics perspective, an attacker is using the output of $\mathcal{M}$ to perform a hypothesis test of whether the input to $\mathcal{M}$ was $\mathcal{D}_1$ (the *null hypothesis*) or $\mathcal{D}_2$ (the *alternative hypothesis*). GDP guarantees that if the significance level (probability of incorrectly rejecting the null hypothesis) of the test is at most α, then the power of the test (probability of correctly rejecting the null hypothesis) is at most $f(\alpha)$ [15].

μ-GDP also has the following interpretation: the ability of an attacker to distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$ is at least as difficult as distinguishing whether a random variable y was sampled from a $N(0, 1)$ distribution or a $N(\mu, 1)$ distribution [15].

Note: Gaussian DP is not defined in terms of a metric over datasets. However, as we show in Section 5, we can combine the two ideas to get the best of both worlds: defining plausible deniability requirements using a metric, and specifying how hard it is to distinguish between two plausible values using GDP. The main challenge is in choosing suitable metrics and designing mechanisms.

Gaussian DP for person-level data has several desirable properties that simplify the design of mechanisms. The first is *postprocessing invariance* that says if a mechanism $\mathcal{M}$ satisfies μ-GDP and $\mathcal{A}$ is a (randomized) algorithm whose domain contains the range of $\mathcal{M}$, then the combined function $\mathcal{A} \circ \mathcal{M}$ (which first runs $\mathcal{M}$ on the input dataset and then runs $\mathcal{A}$ on the output of $\mathcal{M}$) also satisfies μ-GDP. The second important property is *fully adaptive composition* [44]. Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be a sequence of mechanisms. They can be adaptively chosen, so that each $\mathcal{M}_j$ and its privacy parameter μ_j is chosen after seeing the outputs of $\mathcal{M}_1, \ldots, \mathcal{M}_{j-1}$. If $\sqrt{\sum_{i=1}^{k} \mu_j^2} \leq \mu$ in all possible situations, then releasing all of their outputs satisfies μ-GDP. Thus a larger mechanism $\mathcal{M}$ can be constructed out of smaller mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$. The parameter μ is known as the overall *privacy loss budget* and the adaptive composition result allows the privacy budget to be allocated among its components $\mathcal{M}_1, \ldots, \mathcal{M}_k$. This type of privacy accounting is a desirable feature of formal privacy definitions.

Another important feature of Gaussian DP is group privacy, which is the guarantee that is achieved for pairs of datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ that differ on the addition/removal of multiple records.

Lemma 3.3 (Group Privacy [15])**.** Let $\mathcal{M}$ be a mechanism satisfying f-DP. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be datasets that differ on the values of k people. Let $f^{(k)}$ denote application of f for k times (e.g., $f^{(3)}(x) = f(f(f(x)))$). Then, for any set S, $P(\mathcal{M}(\mathcal{D}_2) \in S) \leq f^{(k)}(P(\mathcal{M}(\mathcal{D}_1) \in S))$. In particular, if $\mathcal{M}$ satisfies μ-GDP, then $\Phi^{-1}(P(\mathcal{M}(\mathcal{D}_2) \in S)) \leq \Phi^{-1}(P(\mathcal{M}(\mathcal{D}_1) \in S)) + k\mu$.

4 Related Work

Confidentiality protection for business data has received much less research attention than confidentiality for surveys of individuals. Popular classical techniques are cell suppression [12, 13], EZS noise infusion [20] and synthetic data [27, 35]. Cell suppression is the only one with a formal attack model – an adversary without external information who is restricted to linear operations and should not be able to guess true values to within $p\%$ (where p is a privacy parameter)[12, 13]. EZS multiplicative noise [20] is another relative error protection scheme. It was evaluated by Yang et al. [53], who found that it underprotects small cells and adds significant distortions to large aggregates.

One of the earliest formal approaches to business data confidentiality was proposed by Haney et al. [23]. They modified pure- and approximate-DP to provide relative error protections for confidential aggregates using multiplicative Laplace noise and also additive noise via the smooth sensitivity framework [34]. Seeman et al. [43] generalized this idea with zero-Concentrated DP (zCDP) [6] and introduced the *splitting mechanism* that repeatedly uses any arbitrary underlying zCDP mechanisms to protect data.

The Personalized DP Framework [26] allows every entity to have its own privacy parameter but does not prescribe how to choose it. Seeman et al. [43] and Finley et al. [21] follow this framework and propose methods to protect against membership inference by setting the personalized privacy parameters for zero-concentrated DP [6] so that larger businesses get larger (worse) privacy parameters. The work of [21] is closest to ours, but with the following distinctions. They are concerned with membership inference, while our security model is different and seeks to protect attribute inference (absolute error of attribute inference increases with its size, but relative error decreases). They propose a class of mechanisms called *transformation mechanisms*. The ψ mechanisms we propose in Section 6 are a subset of the transformation mechanisms, but have extra restrictions that prevent paradoxical situation where information about small establishments receives almost no protections, while information about large establishments gets heavily distorted (see Example 6.6). Furthermore, we also introduce the pnc-mechanism, which is novel (Section 6).

5 Confidentiality for Establishments

Recent work on formal privacy for establishments [21, 43] focuses on membership inference (how well the existence of an establishment is protected). However, we note that in many applications, like the QCEW, the *existence* of establishments is intentionally not protected at all. Furthermore, establishments advertise their existence (they need customers to exist) and have public observable physical locations. In terms of utility, strong membership inference protections are also undesirable – making the contributions of Disneyland Resort indistinguishable from that of an ice-cream shop would severely distort statistics used to measure regional economies. At the other extreme, masking the effect of just one **person** does not properly address the confidentiality requirements of establishments (protecting aggregate data like total employment and wages) [23]. Hence, we consider a different security model.

5.1 A security model for establishments

Instead of membership inference, we focus on protections against attribute inference. For business data, attackers are often interested in making precise inferences about establishments in order to gain a business advantage. In our security model, an attacker starts with relatively easy-to-collect information about establishments. For example, some attributes are generally publicly known, including the existence of an establishment, its location, and its industry sector. Small establishments can be visited or observed by an attacker to estimate their size (e.g., an attacker may observe that the number of employees is approximately ≤ 10). The size estimates made by an attacker on a small business are likely to have low absolute error but large relative error. Larger businesses often reveal approximate information about themselves on their own web pages or other public reports. For example, Disneyland Resort shared in 2025 that they employ $\approx 36,000$ employees in the Orange County, CA [37]. This information for larger establishments tends to have lower relative error but higher absolute error than information observable about smaller establishments. Assuming Disneyland rounded the true employment to the nearest thousand, an attacker may make an estimate in the range [35,500,36,500], which would have at most an absolute error of 1,000 and at most a relative error of 3%. Other information gathering activities by attackers can include surveys:

Example 5.1. Suppose an establishment $\mathcal{E}$ has x employees and a statistician (e.g., an attacker or labor economist) wants to estimate its employment count. Suppose nearly all employees of $\mathcal{E}$ live within a region whose total population, N, is known with high accuracy (e.g., from Census data). For some $q \in [0,1]$, the statistician may take a random sample (with replacement) of $n = Nq$ people in the region and record the proportion p that work for $\mathcal{E}$. The quantity np can be modeled as the sum of n Bernoulli(x/N) random variables. So, an unbiased estimate of the number of employees of $\mathcal{E}$ is $\hat{x} = np(N/n)$. The variance of this estimate is $\mathrm{Var}(\hat{x}) = n(x/N)(1-x/N)N^2/n^2 = x\frac{N-x}{n} \leq \frac{1}{q}x$. If the sample is without replacement, then np has a Hypergeometric(N,x,n) distribution. Using the unbiased estimate $\hat{x} = np(N/n)$, we get $Var(\hat{x}) = x\frac{N-x}{Nq}\frac{N-Nq}{N-1} \leq x\frac{N(1-q)}{q(N-1)}$. In both cases, the standard deviation is $O(\sqrt{x})$ and so confidence intervals would have lengths $O(\sqrt{x})$. Thus, the attacker's estimate has absolute error $O(\sqrt{x})$ and relative error $O(1/\sqrt{x})$. Again we see the trend of **low absolute error and high relative error for small establishments, and high absolute error but low relative error for large establishments.**

In our security model, an attacker uses this information to create ballpark estimates of an establishment's size. The goal of the data curator is to release approximate query answers that prevent an attacker from significantly sharpening their inference about establishments. That is, in order to get more precise inferences, the attacker must turn to more expensive data gathering activities rather than attacking the published data.

In our framework, a data curator specifies a neighbor function ψ which defines what "ballpark estimate" means. The neighbor function can be selected by the data curator based on how accurate public information appears to be (in this paper, we argue that $\psi = \sqrt{}$ is a sensible choice). Afterwards, the privacy parameter μ determines how hard it is to sharpen inference within this ballpark estimate.

5.2 The Neighbor Function

Metric DP [10] uses a distance metric to help specify which plausible values should be nearly indistinguishable from one another. However, it does not specify how to create application-dependent metrics. Our approach, μ-Gaussian Establishment DP, creates the metric indirectly. First, we introduce a *neighbor function* that can be used to specify intended uncertainty intervals whose absolute size increases with an establishment's size, but whose relative size decreases. Then we define the metric in terms of the neighbor function. This process allows us to use techniques from Gaussian DP [15] to measure the difficulty of distinguishing between plausible values in the same uncertainty interval instead of the default ϵ-DP [17] used by Metric DP or ρ-zCDP [6] used in prior work on business data [21, 43]. As we explain in Section 5.3, this choice results in more accurate and more interpretable quantification of how strong the plausible deniability guarantees are.

Definition 5.2 (Neighbor function, ψ-close)**.** A real-valued function with nonnegative domain $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is called a *neighbor function* if it is (1) strictly increasing, (2) continuous, (3) concave, and (4) the function $g(x) = \psi(\exp(x))$ is convex. Let $\gamma_1, \ldots, \gamma_{k_c}$ be nonnegative real numbers, which we call *distance parameters*. An establishment record $(a_1^{(1)}, \ldots, a_{k_p}^{(1)}, c_1^{(1)}, \ldots, c_{k_c}^{(1)})$ is ψ-close to an establishment record $(a_1^{(2)}, \ldots, a_{k_p}^{(2)}, c_1^{(2)}, \ldots, c_{k_c}^{(2)})$ with respect to those distance parameters when:

- $a_i^{(1)} = a_i^{(2)}$ for $i = 1, \ldots, k_p$ (i.e., the public attributes match), and
- $|\psi(c_j^{(1)}) - \psi(c_j^{(2)})| \leq \gamma_j$ for $j = 1, \ldots, k_c$ (i.e., the confidential attributes are close enough in a transformed space defined by ψ)

For example, two potential values x and y for the employment count at establishment $\mathcal{E}_1$ are considered close with respect to ψ and a distance parameter γ if $|\psi(x) - \psi(y)| \leq \gamma$. This is the same as requiring $\psi(y) \in [\max(\psi(0), \psi(x) - \gamma),\ \psi(x) + \gamma]$. Two candidate records for the same establishment are considered close by Definition 5.2 if all of their corresponding confidential attributes are close. Hence, our metric is the L_∞ metric after the transformation ψ has been applied.

Definition 5.3 (Uncertainty Interval $\mathcal{I}_{\psi,\gamma}$)**.** The *uncertainty interval* around y induced by ψ and γ is defined as $\mathcal{I}_{\psi,\gamma}(y) = [L,\ U]$ where $L = \psi^{-1}(\max(\psi(0), \psi(x) - \gamma))$ and $U = \psi^{-1}(\psi(x) + \gamma)$. Denote the length of the uncertainty interval as $|\mathcal{I}_{\psi,\gamma}| = U - L$.

If $|\psi(x) - \psi(y)| \leq \gamma$ then $x \in \mathcal{I}_{\psi,\gamma}(y)$ and $y \in \mathcal{I}_{\psi,\gamma}(x)$ (i.e., x and y are inside each other's intervals). So, if two records r_1 and r_2 are ψ-close then they are in each other's uncertainty intervals for all confidential attributes. We discuss specific examples of ψ in Section 5.4. The conditions on ψ in Definition 5.2 ensure that uncertainty intervals behave as intended (increasing absolute error but decreasing relative error as the attribute value increases):

Theorem 5.4. Let ψ be a function satisfying the conditions of Definition 5.2. Then for any x, y, t such that $0 \leq x < y$ and $t \geq 0$

(1) ψ^{-1} exists and is increasing.
(2) We have $|\psi(y+t) - \psi(x+t)| \leq |\psi(y) - \psi(x)|$.

(3) The length of the uncertainty interval around y is at least as big as the interval around x: $|\mathcal{I}_{\psi,t}(y)| \geq |\mathcal{I}_{\psi,t}(x)|$.
(4) The length of the uncertainty interval around x relative to x is at least as big as the length of the uncertainty interval around y relative to y, $\frac{|\mathcal{I}_{\psi,t}(y)|}{y} \leq \frac{|\mathcal{I}_{\psi,t}(x)|}{x}$

Remark 5.5. Definition 5.2 uses one neighbor function ψ and then a separate distance parameter γ_i for each confidential attribute C_i. In the most general case, one can have a separate ψ_i and γ_i for each C_i. We chose the simpler version to increase readability. However, the more general version is needed for a few results in Section 5.5 about interactions between privacy definitions.

Using ψ-closeness, we define neighboring datasets as follows:

Definition 5.6 (ψ-neighbors)**.** Let ψ be a function satisfying the requirements of Definition 5.2. Let $\gamma_1, \ldots, \gamma_{k_c}$ be nonnegative real numbers. Two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are ψ-neighbors if $\mathcal{D}_1$ can be obtained from $\mathcal{D}_2$ by replacing one record r with another record r' that is ψ-close to r (with distance parameters $\gamma_1, \ldots, \gamma_{k_c}$).

5.2.1 Extensions when Confidential Attributes can be Negative. The nonnegativity constraint on confidential attributes is used to simplify the discussion in this paper. However, this assumption can be eliminated in several ways, either by modifying the neighbor function or transforming the data:

- Given a neighbor function ψ defined on nonnegative numbers, one can define an extension ψ' over the real numbers as $\psi'(x) = sign(x)\psi(|x|)$. This provides the semantics that values near 0 get uncertainty intervals with high relative errors and values far from 0 get uncertainty intervals with high absolute errors.
- Sometimes an attribute can be naturally expressed as the difference of two nonnegative attributes; e.g., *profit* can be replaced by *revenue* and *cost*, since profit is the difference of the two.
- Otherwise, a confidential attribute C that is sometimes negative can be replaced with $C' = \max(0, C)$ and $C'' = |\min(0, C)|$. If C is profit, C' can be interpreted as the gain and C'' can be interpreted as the loss. Both C' and C'' are nonnegative and $C = C' - C''$.

5.3 μ-Gaussian Establishment DP

Our approach, μ-Gaussian Establishment DP, adapts Gaussian DP [15] to the neighbor function ψ in order to measure how difficult it should be to distinguish between records that are ψ-close.

Definition 5.7 (μ-gedp)**.** Let $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a function satisfying the requirements of Definition 5.2 and let $\gamma_1, \ldots, \gamma_{k_c}$ be nonnegative real numbers. A mechanism $\mathcal{M}$ satisfies μ-Gaussian Establishment DP (μ-gedp) if for all pairs of ψ-neighbors $\mathcal{D}_1, \mathcal{D}_2$ (with distance parameters $\gamma_1, \ldots, \gamma_{k_c}$) and all sets S, the following is satisfied: $\Phi^{-1}(P(\mathcal{M}(\mathcal{D}_2) \in S)) \leq \Phi^{-1}(P(\mathcal{M}(\mathcal{D}_1) \in S)) + \mu$.

Theorem 5.8 (Adaptive Composition)**.** Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a sequence of mechanisms such that $\mathcal{M}_{i+1}$ can access the output of $\mathcal{M}_1, \ldots, \mathcal{M}_i$. Suppose each $\mathcal{M}_i$ satisfies μ_i-gedp with the same neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$. Then the mechanism that releases all of their outputs satisfies $\sqrt{\sum_{i=1}^{K} \mu_i^2}$-gedp with the same neighbor function and distance parameters.

Thus gedp and Gaussian DP share the same accounting method (privacy parameter and the form of the probabilistic inequality). However, they differ in the choice of neighbors $\mathcal{D}_1, \mathcal{D}_2$. Gaussian DP is concerned with equal membership inference guarantees for all entities, which ruins the utility of establishment data (see experiments in Section 8.2). Meanwhile, gedp protects against attribute inferences being more precise than the uncertainty interval. We discuss specific choices of ψ in Section 5.4, then analyze the interactions between different neighboring functions in Section 5.5.

The semantics of the protections come from the Gaussian DP privacy accounting method and are based on statistical hypothesis testing. Suppose an attacker is trying to distinguish between two candidate records (r_1 and r_2) for an establishment, and both candidate records are within each other's uncertainty intervals. The null hypothesis is that r_1 is true and the alternative hypothesis is that r_2 is true. Based on the output ω of a mechanism satisfying μ-Gaussian Establishment DP, the attacker guesses either r_1 or r_2. The *significance level* α of the hypothesis test is the probability the attacker incorrectly guesses r_2 when the true record is r_1. The *power* is the probability the attacker correctly guesses r_2 when the input is r_2 and is bounded by $\Phi(\Phi^{-1}(\alpha) + \mu)$. Given a desired significance level, more privacy means the power should be lower, so lower values of μ guarantee more privacy.

Note that the Gaussian DP accounting method can be easily swapped out for ϵ-DP [17], approximate-DP [16], zCDP [6], Renyi DP [32], and f-DP [15]. We chose Gaussian DP because it has tighter privacy accounting than zCDP and Renyi DP for mechanisms that inject Gaussian noise [15] and has tighter composition properties than ϵ-DP, approximate-DP, and f-DP [15, 33]. As an illustration, we pose the following question. Suppose an additive noise mechanism adds noise to q counting queries. Under the 3 different privacy accounting schemes of ϵ-DP (used by metric differential privacy [10]), ρ-zCDP (used by prior work on establishment data [21, 43]), and Gaussian DP, what per-query variance is needed to achieve a desired power for a given significance level? For counting queries over person-level data, Figure 1 plots the ratio of per-query variance of ϵ-DP vs. Gaussian DP (left) and ρ-zCDP vs. Gaussian DP (right). (Calculation details can be found in Appendix C). The Figure shows that Gaussian DP always requires less variance than ρ-zCDP to achieve the same semantics. If only 1 query is ever issued, the ϵ-DP is preferable to Gaussian DP. However, if more than 1 query needs to be answered, Gaussian DP requires less per-query variance.

5.4 Semantics and Choices of ψ

Suppose establishment $\mathcal{E}$ has x employees. Intuitively, μ-gedp ensures that x is nearly indistinguishable from any y in the uncertainty interval $\mathcal{I}_{\psi,\gamma}(x)$ around x (Definition 5.3). Note that x is also in the uncertainty interval around y. That is, $y \in \mathcal{I}_{\psi,\gamma}(x) \Leftrightarrow x \in \mathcal{I}_{\psi,\gamma}(y)$ which is mathematically equivalent to $|\psi(x) - \psi(y)| \leq \gamma$. The level of indistinguishability between such x and y is controlled by the μ parameter. Suppose an attacker wishes to determine whether the employment count of $\mathcal{E}$ is x or y. For any attack based on the output of a μ-gedp mechanism $\mathcal{M}$, if the false positive rate (over the randomness in $\mathcal{M}$) is α, then the true positive rate is $\leq \Phi(\mu + \Phi^{-1}(\alpha))$ and if the method's false negative rate is α then the true negative rate is $\leq \Phi(\mu + \Phi^{-1}(\alpha))$. Thus, distinguishing between any x and y

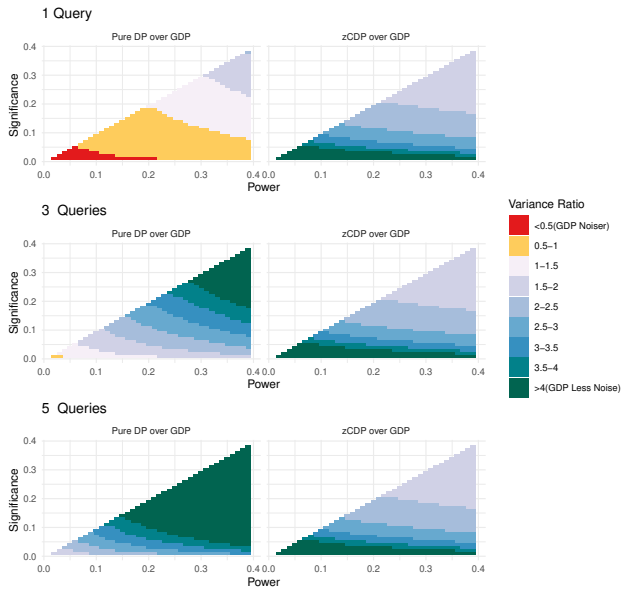

Figure 1: Query variance needed to achieve a given significance level (y-axis) and power (x-axis) when issuing 1 (top), 3 (middle), and 5 (bottom) queries. Left column: ϵ-DP vs. Gaussian DP. Right column: ρ-zCDP vs. Gaussian DP.

in the same uncertainty interval is at least as hard as distinguishing between the Gaussians $N(0,1)$ and $N(\mu,1)$ based on a data point sampled from one of those two Gaussians. Note that this is not a separate guarantee for each attribute. It is a simultaneous guarantee; any two records whose attributes belong to the uncertainty intervals of each other have this indistinguishability guarantee.

Remark 5.9. The group privacy properties of Gaussian DP, which are inherited by gedp, also provide indistinguishability guarantees for wider intervals. If $|\psi(x) - \psi(y)| \leq m\gamma$ then distinguishing between x and y is at least as hard as distinguishing between the Gaussians $N(0,1)$ and $N(m\mu,1)$ based on a point sampled from one of the Gaussians. Stated differently, if $\mathcal{M}$ satisfies μ-gedp with respect to ψ and distance parameters $\gamma_1,\ldots,\gamma_{k_c}$, then $\mathcal{M}$ satisfies $(m\mu)$-gedp with respect to ψ and distance parameters $m\gamma_1,\ldots,m\gamma_{k_c}$.

This group privacy result provides a connection to the membership inference: an establishment's value x is indistinguishable from 0 with privacy parameter $\lceil(\psi(x)-\psi(0))/\gamma\rceil$. It also provides a connection to metric differential privacy, using a discrete metric between records: $d(r_1,r_2) = \max_{i=1,\ldots,k_c}\left\lceil\frac{|\psi(r_1[C_i])-\psi(r_2[C_i])|}{\gamma_i}\right\rceil$ and the privacy parameter decreases with this distance.

Our algorithms work with any ψ compatible with Definition 5.2. However, we believe there are two important potential choices for practical applications:[2] $\psi = \log$ and $\psi = \sqrt{}$. The choice of $\psi = \log$ ensures that the length of the uncertainty interval $\mathcal{I}_{\psi,\gamma}(x)$ is proportional to x. This is consistent with historical tradition of attempting to limit attacker inference about establishments up to within a pre-specified relative error (e.g., 1% or 10%). Notably, this is the motivation for the p% rule used in cell suppression [13] and the purpose of multiplicative EZS noise [20]. The choice of $\psi = \sqrt{}$ is an alternative we advocate for instead. It makes the length of an uncertainty interval around x equal to $O(\sqrt{x})$, which matches the error an attacker would get by alternative means, such as sampling as in Example 5.1.

We next argue that reasonable protections for establishment data should adhere to the following principles (for which $\sqrt{}$ is more suitable than log). The reason is that they match the type of information an attacker can obtain from alternate sources, as discussed in Section 5.1.

Principle 1: *Relative* error protections should be *decreasing* with establishment size.

Principle 2: *Absolute* error protections should be *increasing* with establishment size.

Why are these principles reasonable? As an example, consider $\psi = \log$ with distance parameter $\gamma = 0.1$. The resulting uncertainty interval around any confidential x is approximately $[0.9x,\ 1.10x]$ – meaning all establishments, large and small, are to be protected with the same relative error of 10% . For a small ice-cream shop with 3 employees, the statement that the number of employees is between 2.7 and 3.3 offers no protections at all (under-protection). What about a large establishment? In 2025, Disneyland Resort reported they had 36,000 employees [37]. It is presumably rounded to the nearest thousand and so they may be comfortable with an uncertainty interval somewhere between ±500 (rounding to the nearest thousand) and ±50 (rounding to the nearest hundred). This represents a desired relative error/uncertainty somewhere between 0.3% and 2.8%. Thus, protecting inference about the true employment for Disneyland to 10% is unnecessary and loses utility. Lowering the distance parameter γ will further erode protections for the smaller establishments, while raising it would unnecessarily lose utility for the larger establishments. Thus a fixed relative error seems to be unsuitable — smaller establishments need higher relative error than larger ones. In contrast to $\psi = \log$, the choice of $\psi = \sqrt{}$ with distance parameter, say, 0.5 naturally increases relative error for smaller establishments while decreasing it for the larger establishments (see Table 2 for comparison of $\psi = \sqrt{}$ vs. $\psi = \log$).

	Neighbor function	
x	$\sqrt{}$	log
3	[1.5 – 5.0]	[2.7 – 3.3]
36	[30.2 – 42.2]	[32.6 – 39.8]
360	[341.3 – 379.2]	[325.7 – 397.9]
36,000	[35,810.5 – 36,190.0]	[32,574.1 – 39.786.2]

Table 2: Uncertainty interval $\mathcal{I}_{\psi,\gamma}(x)$ for $\psi(x) = \sqrt{x}$ (with $\gamma = 0.5$) and $\psi(x) = \log(x)$ (with $\gamma = 0.1$).

5.5 Deeper Semantics Analysis

We next consider the following more advanced confidentiality semantics questions. ⟨1⟩ What protections are received by individuals (units smaller than an establishment) and firms (groups of establishments with the same owners)? ⟨2⟩ Can (person-based) differential privacy protections be expressed in the μ-gedp framework? ⟨3⟩ How

[2] More generally, $\psi(x) = \log(x+a)$ or $\sqrt{x+a}$ for some constant $a \geq 0$.

can one combine the protections and semantics of two different neighbor functions? ⟨4⟩ What happens when different data publishers use different neighbor functions when publishing statistics about the same establishments (heterogeneous composition)?

5.5.1 Confidentiality for Firms. A firm is a business made up of one or more establishments, and its confidentiality is equivalent to the group privacy semantics of gedp. Consider a firm that consists of ℓ establishments. We can model the ability of an attacker to detect ℓ total changes to the data from those establishments as follows. Given ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$, suppose $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\ell$ are a sequence of pairwise-neighboring datasets; i.e., for any i, $\mathcal{D}_{i+i}$ is obtainable from $\mathcal{D}_i$ by replacing 1 record with a ψ-close record (Definition 5.2). Let $\mathcal{M}$ be a mechanism that satisfies μ-gedp with respect to $\psi, \gamma_1, \ldots, \gamma_{k_c}$. Then for any set S, the group privacy properties from Remark 5.9 imply that $\Phi^{-1}(P(\mathcal{M}(\mathcal{D}_1) \in S)) \leq (\ell - 1)\mu + \Phi^{-1}(P(\mathcal{M}(\mathcal{D}_\ell) \in S))$. Equivalently, distinguishing between $\mathcal{D}_1$ and $\mathcal{D}_\ell$ based on the output of $\mathcal{M}$ is equivalent to distinguishing between $N(0,1)$ and $N((\ell-1)\mu, 1)$ based on a sampled point from one of those two distributions.

5.5.2 Confidentiality for People. When considering the direct and indirect (group privacy) properties of μ-gedp, protection for individuals is at least as strong as for establishments having one employee. Employees in larger establishments even benefit from blending in a crowd – the protections are obtained by taking the uncertainty intervals around the *establishment* attributes, and re-centering them around the *person's* attributes. For example, let $\psi = \sqrt{\ }$ and $\gamma = 100$ for the wages attributes. Suppose person X earns \$20,000 in a quarter and is known to work at establishment $\mathcal{E}$.

- If the total quarterly wages for $\mathcal{E}$ is $\$20,000$ (i.e., X is the only worker), the uncertainty interval for wages is $\approx$ [\$1,715.7 − \$58,284.3] and so distinguishing between X's true salary vs. any other salary in that range is at least as hard as distinguishing between $N(0,1)$ vs. $N(\mu,1)$.
- If the total quarterly wages for $\mathcal{E}$ is $\$1,000,000$, the uncertainty interval for $\mathcal{E}$'s wages is $\approx$ $[\$810,000 - \$1,210,000]$. Thus any decrease in wages by $\$190,000$ or increase by up to $\$210,000$ are nearly indistinguishable. So, changes to X's quarterly wages by up to $\$190,000$ are nearly indistinguishable (hence the presence of other workers in $\mathcal{E}$ help to better hide the wages of X).

More generally, and formally, suppose the confidential numerical attributes of individual X are $C_1^{(p)}, \ldots, C_{k_c}^{(p)}$ and for establishment $\mathcal{E}$, the aggregate values are $C_1^{(e)}, \ldots, C_{k_c}^{(e)}$. Let $\mathcal{M}$ be a μ-gedp mechanism with ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$. Define the intervals $I_1, \ldots, I_{k_c}$ as the uncertainty intervals for $\mathcal{E}$ that are re-centered around the values for X; i.e., $I_i = \mathcal{I}_{\psi,\gamma_i}(C_i^{(e)}) - C_i^{(e)} + C_i^{(p)}$. Then distinguishing between the true values $C_1^{(p)}, \ldots, C_{k_c}^{(p)}$ vs. any other choice $\widehat{C}_1^{(p)}, \ldots, \widehat{C}_{k_c}^{(p)}$ from those intervals is at least as hard as distinguishing between $N(0,1)$ vs. $N(\mu,1)$.

Group privacy (Remark 5.9) lets us consider what protections are provided within wider intervals. For instance, given an integer $k \geq 1$, if we define the wider intervals $I_1^{(k)}, \ldots, I_{k_c}^{(k)}$ as $I_i^{(k)} = \mathcal{I}_{\psi,k\gamma_i}(C_i^{(e)}) - C_i^{(e)} + C_i^{(p)}$ then distinguishing between the true values for X and any other choices from those intervals is at least as hard as $N(0,1)$ vs. $N(k\mu,1)$.

5.5.3 Encoding (person-based) Gaussian DP in gedp. We next consider how to directly encode DP protections in this framework; in Section 5.5.4, we consider how to combine protections from different ψ choices, such as combining DP protections for people with $\psi = \sqrt{\ }$ protections for establishments. Suppose the contributions of any person to an establishment's confidential attributes C_i is bounded by a constant d_i (e.g., person X can only add at most 1 to an employment count, or salary is clipped by some constant d before being aggregated). Then set neighbor function $\psi^{(DP)}(x) = x$ and distance parameters $\gamma_1^{(DP)}, \ldots, \gamma_{k_c}^{(DP)}$ to $d_1, \ldots, d_{k_c}$, respectively. This makes two datasets neighbors if they differ on the contributions of one person to an establishment's values.

5.5.4 Combining Neighbor Functions. We next consider how to combine the protections provided by some neighbor function $\psi^{(A)}$ having distance parameters $\gamma_1^{(A)}, \ldots, \gamma_{k_c}^{(A)}$ with another neighbor function $\psi^{(B)}$ having distance parameters $\gamma_1^{(B)}, \ldots, \gamma_{k_c}^{(B)}$. This requires using the more general form of μ-gedp, as discussed in Remark 5.5, in which each confidential attribute C_i has its own neighbor function ψ_i^* (instead of all confidential attributes sharing the same neighbor function). The goal is to obtain a new neighbor function ψ_i^* and distance parameter γ_i^* for each attribute C_i so that for every x and i, the resulting uncertainty interval $\mathcal{I}_{\psi_i^*,\gamma_i^*}(x)$ contains the uncertainty intervals $\mathcal{I}_{\psi^{(A)},\gamma_i^{(A)}}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma_i^{(B)}}(x)$. This non-obvious result is provided by the following theorem.

Theorem 5.10. Let $\psi^{(A)}$ and $\psi^{(B)}$ be neighbor functions satisfying the conditions of Definition 5.2. Let $\gamma^{(A)}$ and $\gamma^{(B)}$ be positive numbers. Then $\psi^{(A)}$ and $\psi^{(B)}$ are continuously differentiable almost everywhere. Set $\gamma^* = 1$ and define ψ^* as follows:

$$\frac{d\psi^*(x)}{dx} = \min\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x)}{dx}, \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x)}{dx}\right)$$

$$\psi^*(x) = \int_0^x \frac{d\psi^*(x)}{dx}\, dx$$

Then ψ^* satisfies the conditions of Definition 5.2 to be a neighbor function. Furthermore, the uncertainty intervals defined by $\psi^{(A)}, \gamma^{(A)}$ and $\psi^{(B)}, \gamma^{(B)}$ are contained in the uncertainty interval defined by $\psi^*, \gamma*$. That is for any x, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \subseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma^{(B)}}(x) \subseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

As an application of Theorem 5.10, we show how to combine the setting $\psi = \sqrt{\ }$ having distance parameters $\gamma_1, \ldots, \gamma_{k_c}$ with the differential privacy setting (from Section 5.5.3) of neighbor function $\psi^{(DP)}(x) = x$ and distance parameters $d_1, \ldots, d_{k_c}$. Applying the theorem for each attribute, we get $\gamma_i^* = 1$ and

$$\psi_i^*(x) = \begin{cases} \frac{x}{d_i} & \text{if } x \leq \frac{d_i^2}{4\gamma_i^2} \\ \frac{\sqrt{x}}{\gamma_i} - \frac{d_i}{4\gamma_i^2} & \text{if } x \geq \frac{d_i^2}{4\gamma_i^2} \end{cases}$$

for $i = 1, \ldots, k_c$. Thus, two records r_1 and r_2 are said to be "close" if they have the same value for their public attributes and

$$\left|\psi_i^*(r_1[C_i]) - \psi_i^*(r_2[C_i])\right| \leq \gamma_i^* \text{ for } i = 1, \ldots, k_c$$

and two datasets would be neighbors if one can be obtained from the other by replacing one record with another record that is close. The mechanisms we propose in Section 6 work directly with this generalization of closeness. However, to keep the amount of notation manageable, the discussion in the rest of the paper uses the same neighbor function for each confidential attribute.

5.5.5 Heterogeneous Composition. Suppose two different data publishers A and B release statistics about the same underlying data, but with completely different settings. One uses $\mu^{(A)}$-gedp with $\psi^{(A)}$ and $\gamma_1^{(A)}, \ldots, \gamma_{k_c}^{(A)}$ and the other uses $\mu^{(B)}$-gedp with $\psi^{(B)}$ and $\gamma_1^{(B)}, \ldots, \gamma_{k_c}^{(B)}$. We call this *heterogeneous* composition. The goal is to find new neighbor function ψ_i^* and distance parameter γ_i^* for each attribute, such that the resulting uncertainty intervals are contained inside both $\mathcal{I}_{\psi^{(A)},\gamma_i^{(A)}}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma_i^{(B)}}(x)$ for each i and x. This means mechanism $\mathcal{M}^{(A)}$ also satisfies $\mu^{(A)}$-gedp with respect to the (per-attribute) neighbor functions $\psi_1^*, \ldots, \psi_{k_c}^*$ and distance parameters $\gamma_1^*, \ldots, \gamma_{k_c}^*$ (i.e., the general version of gedp discussed in Remark 5.5 and Section 5.5.4) and $\mathcal{M}^{(B)}$ satisfies $\mu^{(B)}$-gedp for those neighbor functions and distance parameters as well. Then this reduces to the standard composition and the overall confidentiality parameter becomes $\sqrt{(\mu^{(A)})^2 + (\mu^{(B)})^2}$. The following theorem shows to create the appropriate ψ_i^* and γ_i^* for each attribute.

Theorem 5.11. Let $\psi^{(A)}$ and $\psi^{(B)}$ be neighbor functions satisfying the conditions of Definition 5.2. Let $\gamma^{(A)}$ and $\gamma^{(B)}$ be positive numbers. Then $\psi^{(A)}$ and $\psi^{(B)}$ are continuously differentiable almost everywhere. Set $\gamma^* = 1$ and define ψ^* as follows:

$$\frac{d\psi^*(x)}{dx} = \max\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x)}{dx}, \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x)}{dx}\right)$$

$$\psi^*(x) = \int_0^x \frac{d\psi^*(x)}{dx}\, dx$$

Then ψ^* satisfies the conditions of Definition 5.2 to be a neighbor function. Furthermore, the uncertainty intervals defined by $\psi^{(A)}, \gamma^{(A)}$ and $\psi^{(B)}, \gamma^{(B)}$ contain the uncertainty interval defined by ψ^*, γ^*. That is for any x, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \supseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma^{(B)}}(x) \supseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

6 Mechanisms for gedp

Mechanisms for Gaussian Establishment DP can be designed using the appropriate concept of sensitivity. However, the requirement that absolute error protections increase with establishment size means that the mechanisms have data-dependent variance. We first present the sensitivity-based framework for gedp. We then propose two mechanisms: (1) the ψ-mechanism, which has data-dependent variance that cannot be released exactly (to avoid compromising the confidentiality guarantees), and (2) the pnc-mechanism, which piggybacks on the ψ-mechanism to answer additional queries with data-dependent variances that are safe to release.

Definition 6.1 (ψ-neighbor sensitivity)**.** Given a vector-valued query function q, the ψ-neighbor sensitivity of q denoted as $\Delta_\psi(q)$ is defined as: $\Delta_\psi(q) = \sup_{\mathcal{D}_1,\mathcal{D}_2} ||q(\mathcal{D}_1) - q(\mathcal{D}_2)||_2$, where the supremum is taken over all pairs of ψ-neighboring datasets with associated distance parameters $\gamma_1, \ldots, \gamma_{k_c}$.

The resulting Gaussian-style mechanism is similar to the one used by GDP [15], but with ψ-neighbor sensitivity in place of L_2 sensitivity:

Definition 6.2 (Estab-Gaussian mechanism)**.** Let μ be a positive real-valued privacy parameter and let q be a vector-valued function. Given a neighbor function ψ with associated distance parameters $\gamma_1, \ldots, \gamma_{k_c}$ and an input dataset $\mathcal{D}$, the Estab-Gaussian mechanism for gedp returns $q(\mathcal{D}) + N\left(0, \frac{{\Delta_\psi}^2(q)}{\mu^2} I\right)$.

Theorem 6.3. Given a privacy budget $\mu > 0$, neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$, the Estab-Gaussian mechanism (Definition 6.2) satisfies μ-gedp.

6.1 The ψ-mechanism.

Group-by sum queries are some of the most important types of queries for establishment data. We represent them as $gsum_{\langle \mathcal{G}, C_i\rangle}$, where $\mathcal{G}$ is a function that creates groups from the public attributes of an establishment and C_i is the confidential attribute to sum over. For example, consider the query "total wages for each combination of county and 2-digit NAICS prefix." Here $\mathcal{G}$ would return a tuple consisting of the county and 2-digit NAICS prefix of an establishment, and C_i would be the wages attribute.

Definition 6.4. Let $gsum_{\langle \mathcal{G}, C_i\rangle}$ be a groupby-sum query that aggregates over confidential attribute C_i having distance parameter γ_i. Let $\mu > 0$ be a privacy parameter. The ψ-mechanism $\mathcal{M}_\psi$ is defined as $\mathcal{M}_\psi(\mathcal{D}) = \psi(gsum_{\langle \mathcal{G}, C_i\rangle}(\mathcal{D})) + N(\mathbf{0}, \frac{\gamma_i^2}{\mu^2}\mathbf{I})$ – it applies ψ to every entry in the group-by query result, then adds independent Gaussian noise, with standard deviation γ_i/μ, to each entry.

Theorem 6.5. Let ψ be a neighbor function satisfying Definition 5.2. Let $\gamma_1, \ldots, \gamma_{k_c}$ be the associated distance parameters. Then the ψ-mechanism $\mathcal{M}_\psi$ satisfies μ-gedp with neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$.

The personalized DP framework for establishment data by Finley et al. [21] also propose mechanisms that first transform the data and then add Gaussian noise. Every ψ-mechanism from our framework is a valid transformation mechanism in the framework of [21] (i.e., a concave, strictly increasing function with domain from $[a, \infty)$ with $a \in \mathbb{R}$). However, only a subset of their transformation mechanisms correspond to a ψ-mechanism in our framework. The reason is that not every transformation in their framework satisfies Definition 5.2, specifically Condition 4. That condition prevents small establishments from having *lower* relative error protections than larger establishments, as the following example shows:

Example 6.6. Consider the neighbor function ψ and its inverse:

$$\psi(x) = \begin{cases} 10x & \text{if } x \le 10 \\ \frac{x}{100} + 99.9 & \text{if } x \ge 10 \end{cases} \quad \psi^{-1}(y) = \begin{cases} \frac{y}{10} & \text{if } y \le 100 \\ 100(y - 99.9) & \text{if } y \ge 100 \end{cases}$$

This function satisfies the requirements of [21] because it is concave and strictly increasing. It does not satisfy our requirements of a neighbor function because $\psi(\exp(x))$ is not convex. We next show that releasing $\psi(x) + N(0, 1)$ would result in counterintuitive

protections where small establishments are unprotected while large establishments get much higher relative error protections.

Let x_1 and x_2 be the employment counts of two establishments. Suppose $\psi(x_1) + N(0,1) = 51$. What can we infer about x_1? Using the properties of the Gaussian distribution, it means that with probability > 0.999, $\psi(x_1) \in 51 \pm 3.3$ and so with overwhelming probability $x_1 \in 5.1 \pm 0.33$. The relative length of this 0.999-confidence band is $2\frac{0.33}{5.1} < 0.13$. So x_1 is almost certainly 5. The corresponding establishment received very little relative and absolute error protections. What if $\psi(x_2) + N(0,1) = 106.3$? In this case, $\psi(x_2) \in 106.3 \pm 3.3$ with probability > 0.999. Using the inverse ψ^{-1}, we see that $x_2 \in [640 \pm 330]$ with overwhelming probability. This confidence band has relative size $2\frac{330}{640} > 1.03$ which is over 7 times larger than for the small establishment x_1. We can tell that x_2 is large, but its relative and absolute errors are much higher.

Additionally, the above example shows that if we get a noisy answer ω_g corresponding to a group g in a group-by query, then if $[-z, z]$ is a q% confidence interval for the standard Gaussian, then $[\psi^{-1}(\omega_g - z\gamma_i/\mu), \quad \psi^{-1}(\omega_g + z\gamma_i/\mu)]$ is a q% confidence interval for the true aggregation of C_i in that group g. The quantity $\psi^{-1}(\omega_g)$ is the maximum likelihood estimate of the true value, but is biased. The following theorem provides (1) unbiased estimators[3] when $\psi = \sqrt{}$ or $\psi = \log$, (2) the distribution of the unbiased estimators and (3) the true variance. Since the variance is a function of the true value, it cannot be released but may be estimated from the noisy answer. The following theorem shows that the variance can be accurately estimated when $\psi = \sqrt{}$, but for $\psi = \log$, the variance estimates will never converge.

Theorem 6.7. Given a group-by sum query $gsum_{\langle \mathcal{G}, C_i \rangle}$, let x be the true sum within a given group g, and let $\omega_g = \psi(x) + N(0, \gamma_i^2/\mu^2)$ be the noisy answer provided by $\mathcal{M}_\psi$. Then

- If $\psi = \sqrt{}$, then ω_g^2 has the distribution of a non-central χ^2 random variable with 1 degree of freedom and non-centrality parameter $x\mu^2/\gamma^2$, multiplied by γ_i^2/μ^2. Then $\widetilde{x} \equiv \omega_g^2 - (\gamma_i^2/\mu^2)$ is an unbiased estimate of x. The variance of $\widetilde{x}$ is $v \equiv 2(\gamma_i/\mu)^2\left(2x + (\gamma_i/\mu)^2\right)$. Replacing x with $\widetilde{x}$ in this formula results in a releasable estimate $\widetilde{v}$ of the variance, and $\widetilde{v}/v \to 1$ in probability as $x \to \infty$.
- If $\psi = \log$, then e^{ω_g} has the log-normal distribution with location parameter $\log(x)$ and scale γ_i^2/μ^2. Then $\widetilde{x} \equiv \exp(\omega_g - \gamma_i^2/(2\mu^2))$ is an unbiased estimate of x. The variance of $\widetilde{x}$ is $v \equiv x^2(-1 + \exp(\gamma_i^2/\mu^2))$. Replacing x with $\widetilde{x}$ in this formula gives a releasable estimate $\widetilde{v}$, but $\widetilde{v}/v$ does not converge in probability as $x \to \infty$.

Unbiased (noisy) query answers and estimates of their variances are important for the postprocessing step of converting the noisy answers into confidentiality-preserving microdata (see Section 7). Such microdata are useful to statistical agencies, who are often requested to release additional special tabulations. Having confidentiality-preserving microdata on hand means that the agencies can tabulate query answers from the microdata. It is an advantage over the prevailing cell suppression approach, for which it is difficult to analyze interactions with new data products.

From Theorem 6.7, we see that use of $\psi = \log$, which is consistent with historical tradition of constant relative error protections, is poorly suited for the microdata creation process as it does not allow good estimates of noisy query variances – the variance estimates do not converge no matter how large an establishment is. The requirement to estimate the variance can have serious consequences on the quality of the confidentiality-preserving microdata (as we explain in Section 7). Hence we next propose the pnc-mechanism whose data-dependent variance is safe to release.

[3]The results of [21, 50] can be used to obtain unbiased estimators and their true variances in a more general setting.

6.2 Probably-No-Clipping (pnc) Mechanism

To answer group-by sum queries $gsum_{\langle \mathcal{G}, C_i \rangle}$ with data-dependent Gaussian noise but variances that are safe to release, we propose the probably-no-clipping mechanism (pnc-mechanism). The prerequisite is a simultaneous high-probability upper bound $u_{j,i}$ on $r_j[C_i]$ (the value of confidential attribute C_i for the record belonging to establishment $\mathcal{E}_j$) for all i, j. **Privacy is not affected by failure of the upper bound.** Given such an upper bound, we replace the true values $r_j[C_i]$ with $\min(r_j[C_i], u_{j,i})$ when aggregating records in a group, and add Gaussian noise to satisfy μ-gedp. We obtain the $u_{j,i}$ by running a ψ-mechanism to answer the identity query and applying Theorem 6.8.

Theorem 6.8. Given establishment records $r_1, \ldots, r_n$, a neighboring function ψ with distance parameters $\gamma_1, \ldots, \gamma_{k_c}$, and a privacy loss budget allocations $\mu_1, \ldots, \mu_{k_c}$, define the noisy values $\omega_{j,i}$ as $\omega_{j,i} = \psi(r_j[C_i]) + N(0, \gamma_i^2/\mu_i^2)$. Let $\zeta \in (0,1)$ be a confidence parameter. First, define the *probably-no-clipping parameter* $\widehat{\tau} = \Phi^{-1}((1-\zeta)^{1/(k_c n)})$. Next, define upper bounds $u_{j,i}$ as $u_{j,i} = \psi^{-1}(\omega_{j,i} + \gamma_i\widehat{\tau}/\mu_i)$. Then with probability $1 - \zeta$ the following inequalities simultaneously hold: $r_j[C_i] \le u_{j,i}$ for all j and i.

The upper bounds $u_{j,i}$ from Theorem 6.8 are *public*, since they result from postprocessing the answers to a ψ-mechanism. We use them to define the pnc-mechanism for group-by sum queries.

Definition 6.9 (Probably no clipping mechanism)**.** Let $gsum_{\langle \mathcal{G}, C_i \rangle}$ be a group-by sum-query, $g_1, \ldots, g_\ell$ be the corresponding disjoint groupings of establishments, and the $u_{j,i}$ be the upper bounds from Theorem 6.8. The pnc-mechanism for $gsum_{\langle \mathcal{G}, C_i \rangle}$ with privacy budget $\mu > 0$ is: (1) For each group g_ℓ compute the maximum establishment contribution: $u_\ell^* = \max_{\mathcal{E}_j \in g_\ell} u_{j,i}$. (2) Compute the per-group sensitivity for each group g_ℓ: $\Delta_\ell = u_\ell^* - \psi^{-1}(\max(0, \psi(u_\ell^*) - \gamma_i))$. (3) Produce the per-group truncated summation for each group g_ℓ: $T_{g_\ell} = \sum_{\mathcal{E}_j \in g_\ell} \min(r[C_i], u_\ell^*)$. (4) Release the aggregate values. For each group g_ℓ release $T_{g_\ell} + N(0, (\Delta_\ell/\mu)^2)$ and the variance $(\Delta_\ell/\mu)^2$.

Theorem 6.10. The pnc-mechanism from Definition 6.9 satisfies μ-gedp with neighboring function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$.

The complete workflow is: (1) Run the ψ-mechanism with privacy budgets $\mu_1^{(a)}, \ldots, \mu_{k_c}^{(a)}$ to answer the identity queries for confidential attributes $C_1, \ldots, C_{k_c}$. (2) Compute the $u_{j,i}$ using Theorem 6.8. (3) Use the pnc-mechanism to answer group-by queries $gsum_{\langle \mathcal{G}_1, C_{i_1} \rangle}, \ldots, gsum_{\langle \mathcal{G}_m, C_{i_m} \rangle}$ with budgets $\mu_1^{(b)}, \ldots, \mu_m^{(b)}$. The total privacy budget is:

$$\mu^{(total)} = \sqrt{(\mu_1^{(a)})^2 + \cdots + (\mu_{k_c}^{(a)})^2 + (\mu_1^{(b)})^2 + \cdots + (\mu_m^{(b)})^2}.$$

Note: in our experiments, we set the clipping parameter to be 0.01. This means that with probability 0.99, no clipping occurs at all when answering a batch of queries.

7 Postprocessing into Microdata

Postprocessing a set of confidentiality-preserving query answers into confidentiality-preserving establishment-level data $\widehat{\mathcal{D}}$ (also known as microdata) is an important task in practice [2, 29] as it allows agencies to answer additional queries from $\widehat{\mathcal{D}}$ without spending additional privacy budget. A common approach is inverse-variance-weighted least squares [2, 24, 29]. Starting with a collection of queries, $q_1, \ldots, q_k$ and their respective noisy answers, $a_1, \ldots, a_k$ (e.g., produced via the ψ-mechanism with postprocessing via Theorem 6.7 or pnc-mechanism), let $v_1, \ldots, v_k$ be the corresponding public estimates of the noisy answer variances.[4] In the case of the pnc-mechanism, the true variances are publicly known; for the ψ-mechanism with $\psi = \log$ or $\psi = \sqrt{}$ (followed by postprocessing), the v_i are approximations (e.g., using Theorem 6.7). To create $\widehat{\mathcal{D}}$, one solves the following optimization problem:

$$\widehat{\mathcal{D}} \leftarrow \arg\min_{\widehat{\mathcal{D}}} \sum_{i=1}^{k} \frac{1}{v_i} (q_i(\widehat{\mathcal{D}}) - a_i)^2. \tag{1}$$

Non-negativity or other domain-specific constraints can also be added, although their effects on data quality can have unintended consequences [1] that require study for each application.

8 Experiments

We evaluate our proposed methods on real and synthetic data. The real dataset is a historical QCEW dataset which we used via a cooperative agreement with the BLS. The state, year, and quarter cannot be disclosed for legal and policy reasons. We generated synthetic data by combining publicly available data sources [7, 8, 18] with statistical techniques like imputation; the details of this process can be found in Appendix H. The synthetic data, the code to generate it, and the code to evaluate it can be found online [36]. In both datasets, the public attributes of an establishment are its NAICS code and county, while the 4 private attributes are the employment counts for the first, second and third months in the quarter, and the total wages paid during the quarter. The neighbor function for Gaussian Establishment DP is set to $\psi = \sqrt{}$. We compare the accuracy of (1) privacy-preserving microdata generated from queries entirely answered by the $\sqrt{}$-mechanism (i.e., ψ-mechanism with $\psi = \sqrt{}$) combined with Theorem 6.7, followed by conversion into microdata to (2) privacy-preserving microdata generated from queries answered by the pnc-mechanism (except for the identity query, which must be answered by the $\sqrt{}$-mechanism, as explained in Section 6.2).

8.1 Evaluations on Confidential QCEW Data

We next describe the settings for the experiment on QCEW data for which we have permission to share results, which only cover employment data (not wages). The **measurement queries** are the group-by queries for which the mechanisms produced noisy answers for the 3 monthly employment counts using the following groupings: (Q1) Identity – no grouping, (Q2) State total – grouping all establishments in the (redacted) state covered by the data, (Q3) 5-digit NAICS prefix, (Q4) County, (Q5) County by 5-digit NAICS prefix. For each of the 3 months, the respective query budgets were: (Q1) 0.7, (Q2) 0.2, (Q3) 0.6, (Q4) 0.6, (Q5) 0.7. This results in an overall privacy budget (square root of the sum of squares) $\mu = \sqrt{0.7^2 * 3 + 0.2^2 * 3 + 0.6^2 * 3 + 0.6^2 * 3 + 0.7^2 * 3} \approx 2.28$. The

[4] The $v_1, \ldots, v_k$ depend solely on the mechanisms and privacy parameters, and not on statistical models of the data

Level	$\sqrt{}$-mechanism 1st	Median	Mean	3rd	pnc-mechanism 1st	Median	Mean	3rd
State Level Employment								
Total	-6,198	-6,198	-6,198	-6,198	-118	-118	-118	-118
Supersector	-1,083	-293	-563	-233	-100	-48	-11	24
NAICS-5	-32	-9	-9	8	-14	-1	-0	12
County Level Employment								
Total	-420	-252	-4,604	-143	-64	-11	-4,331	27
Supersector	-69	-18	-419	16	-43	-4	-394	30
NAICS-5	-7	-1	-10	4	-6	-1	-9	5

Table 3: Quartiles of *signed* differences between group-by query answers computed from confidentiality-preserving microdata vs. ground truth. Note super-sector is essentially 2-digit NAICS prefix.

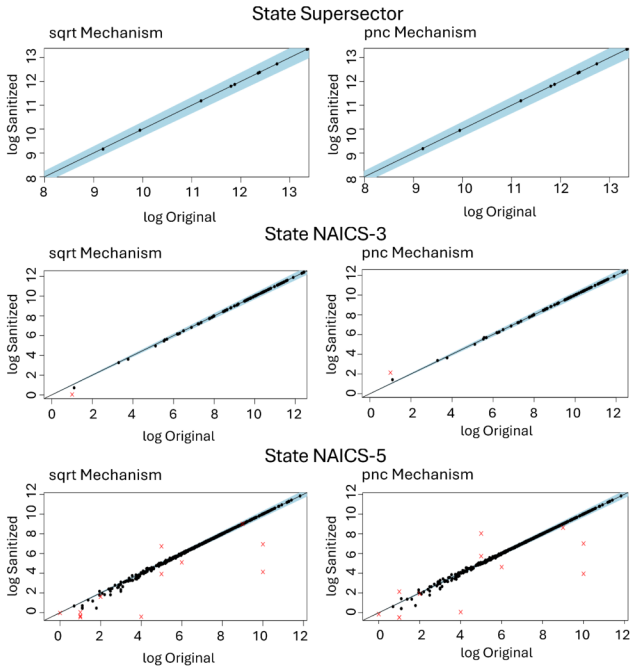

Figure 2: (Employment) Scatter plot of confidentiality-preserving microdata vs. ground truth for the groups in group-by queries for the workflow based on the $\sqrt{}$-mechanism (first column) and pnc-mechanism (second column). The blue band shows ±3% of the original value. Points outside this band are colored red. Only the smallest groups (where error must be large to protect confidentiality) have values that are further than 3% from the ground truth.

distance parameter for employment was $\gamma_{emp} = 0.5$. The pnc-mechanism confidence parameter was $\zeta = 0.01$ (i.e., with probability 0.99, no group-by query was changed by clipping). The noisy measurement query answers are postprocessed into confidentiality-preserving microdata. The **evaluation queries** are the group-by queries for which we evaluated accuracy – they can be different from the measurement queries.

Table 3 shows the *signed* difference between employment group-by-queries computed from the confidentiality-preserving microdata and the ground truth (i.e., we don't take the absolute value). Negative numbers indicate a negative difference (i.e., downward bias). In addition to state total, the groupings used for evaluation are supersector (2-digit NAICS prefix) and 5-digit NAICS prefix. The table summarizes the differences by computing the quartiles across groups in the group-by-queries (for state total, all establishments are grouped together into 1 group, so all quartiles are the same). We see that using the pnc-mechanism produces greater accuracy and less bias. This serves as additional empirical validation for Section 7, as it shows the pnc-mechanism reduces bias in the microdata. Some bias is still present because usage of the pnc-mechanism requires that the identity query be answered by the ψ-mechanism (with all other group-by queries answered by the pnc-mechanism).

Figure 2 shows a scatter-plot of true vs. noisy answers for individual groups within employment group-by-queries (supersector, 3-digit NAICS prefix, 5-digit NAICS prefix), with a 3% band around the true value. Points outside the band are colored red. The results show both methods are accurate. All but the smallest groups have answers within 3% of the true value. The groups with small employment counts, especially if they contain few establishments, would raise the most confidentiality concerns, and therefore should exhibit large errors. This is reflected by more red points in fine-grained groupings (e.g., group by 5-digit NAICS prefix) compared to coarser groupings (e.g., group by 3-digit NAICS prefix).

8.2 Evaluations on Synthetic Data

Since experiments on confidential QCEW data are limited by policy, we synthesized a dataset with the same schema using public information [7, 8, 18] and statistical imputation (see the Appendix H for details). We use synthetic New Jersey (Syn-NJ) and Synthetic Rhode Island (Syn-RI) data, which serve as examples of fairly large and fairly small states, for further experimentation. This allows us to add results for wages as well. Due to space constraints, only a subset of our results is presented.

Synthetic New Jersey (Syn-NJ) includes 208,868 establishments and 21 counties. The number of establishments in a county ranges from 999 to 28,689 with mean 9,894, median 9,385 and standard deviation 6,968. The number of establishments for different NAICS codes present in Syn-NJ ranges from 1 to 8,172 with mean 364 and median 79 – so this distribution is highly skewed. The largest establishment's employment is 7,629 and the largest wages is 3,136,080. Synthetic Rhode Island has only 5 counties and 25,420 establishments. The number of establishments in a county ranges from 1,102 to 14,414 with mean 5,084, median 3,354, and standard deviation 5,340. The number of establishments for different NAICS codes present in Syn-RI ranges from 1 to 610, with mean 13, median 4 and standard deviation 34. The largest Syn-RI establishment's employment is 4,050 and the largest wages is 510,039. As a baseline, we can compare against the Gaussian mechanism of GDP at the establishment level (i.e., providing equal defense against membership inference for all establishments). We use the values of the largest establishment in each state as sensitivity bounds used by the GDP Gaussian mechanism (giving it a slight non-private advantage).

For the privacy settings, we set the privacy budget allocation for wages to be the same as the QCEW allocations for employment in Section 8.1 and then rescaled everything so that the total privacy budget was nearly the same as in Section 8.1. Specifically, each month of employment, and also the quarterly wages had the following privacy budgets for the group-by query groupings: (Q1) Identity – no grouping: 0.611, (Q2) State total – grouping all establishments in the same: 0.179, (Q3) 5-digit NAICS prefix: 0.525, (Q4) County: 0.525, (Q5) County by 5-digit NAICS prefix: 0.611. We used the distance parameter $\gamma_w = 50$. The overall privacy budget was $\mu = \sqrt{0.611^2 * 4 + 0.179^2 * 4 + 0.525^2 * 4 + 0.525^2 * 4 + 0.611^2 * 4} \approx 2.31$.

For each workflow (pnc-mechanism vs. $\sqrt{\ }$-mechanism vs. GDP mechanism), we compute the absolute error of query answers and then repeat the experiment for 34 repetitions in order to assess variability of the errors.

Figures 3 and 4 compare the absolute errors for employment and wages on a variety of group-by queries on Synthetic New Jersey and Rhode Island. The figures summarize the errors using a box plot, which shows the median, upper and lower quartiles (the top and bottom of the box) and the spread of points outside the quartiles (the whiskers and dots outside the boxes). Figure 3 summarizes the absolute error for each group within a query is averaged across the replicates. The group-by queries cover (1) grouping by county to get county-level employment and wage totals, (2) grouping by 3-digit NAICS prefix – this is an important query that is not a measurement query, and (3) grouping by 5-digit NAICS prefix – this is an important query and is expected to be sparse (i.e., many of the groups have very few establishments). Figure 4 summarizes the absolute error from each replicate of the trivial grouping in which all establishments are combined together to get state-level employment and wages totals.

Both of our proposed mechanisms perform better than the GDP mechanism (left plot of Figure 3). The scale of the noise from the GDP mechanism renders the data fairly useless and thus is omitted from the rest of our discussion. For employment counts (the top rows of the Figures 3 and 4), the pnc-mechanism performs substantially better than $\sqrt{\ }$-mechanism. This is because the amount of noise injected into a group by $\sqrt{\ }$-mechanism depends on the total within that group. However, the amount of noise added to a group by pnc-mechanism depends on the value of the largest establishment in the group. Specifically, *pnc − mechanism* uses the noisy identity query to estimate simultaneous high-probability upper bounds on each establishment and then clips each establishment's wages and employment to ensure that the upper bounds are not exceeded. It adds up the clipped values in a group,[5] and then adds noise based on the largest establishment upper bound in the group. If no establishment dominates the group (i.e., is much larger than

[5]Since the upper bound is high probability, it means that all of the clipped values are probably identical to the true values

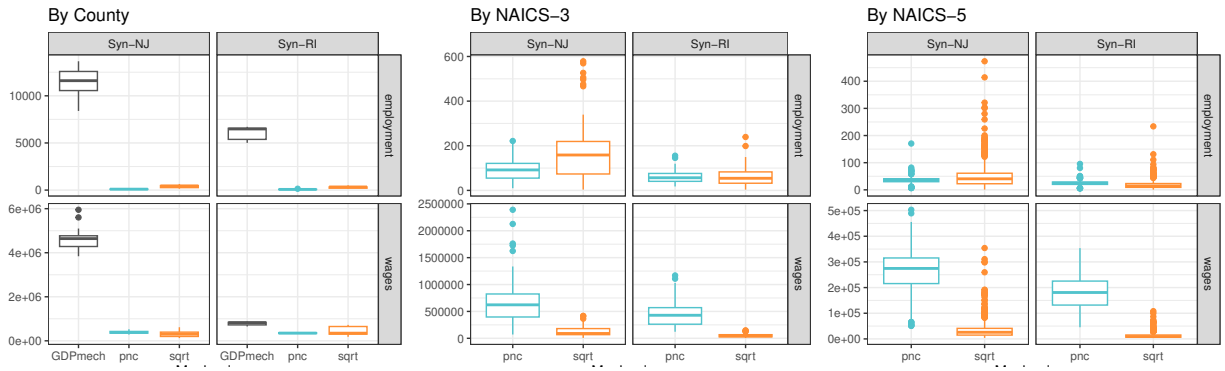

Figure 3: Absolute error averaged across 34 replications for employment group-by queries (top) and wages (bottom) for Synthetic New Jersey and Synthetic Rhode Island data.

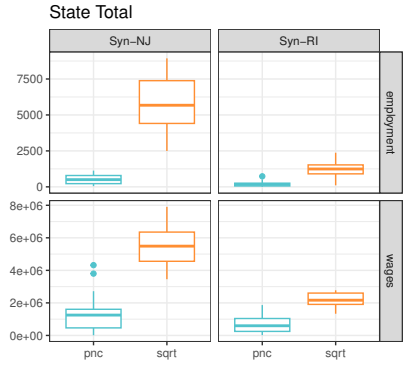

Figure 4: Absolute error of trivial grouping at state-level for employment (top) and wages (bottom) for Synthetic New Jersey and Synthetic Rhode Island data across 34 replicates.

the rest), then the largest of the establishment upper bounds estimated by pnc-mechanism is significantly smaller than the total within a group, leading to less noise compared to $\sqrt{\ }$-mechanism.

Thus, for high-level aggregations like state total (Figure 4), there are so many establishments in a single group that no establishment can dominate the employment or wages. Hence pnc-mechanism significantly outperforms $\sqrt{\ }$-mechanism on state totals. However, as the number of establishments in a group decreases, a single establishment is more likely to dominate the total within its group, giving the pnc-mechanism an advantage over the $\sqrt{\ }$-mechanism.

For medium-level aggregations, like County (first plot of Figure 3), there is an increasing chance that some groups in the group-by queries contain a single establishment that can dominate the rest. Hence, the advantage of pnc-mechanism over $\sqrt{\ }$-mechanism starts to decrease. It still performs better on employment and also Syn-RI wages, but the $\sqrt{\ }$-mechanism is slightly better for Syn-NJ wages.

The comparison for the NAICS-3 (second plot of Figure 3) and NAICS-5 (last plot of Figure 3) is particularly interesting because NAICS-5 is a sparse group-by query (this should benefit the $\sqrt{\ }$-mechanism) containing many groups, and NAICS-3 is a non-measurement query (all the information about it contained in the confidentiality-preserving microdata mostly comes from the noisy answers to the group-by 5-digit NAICS prefix query). Surprisingly, we continue to see that pnc-mechanism outperforms $\sqrt{\ }$-mechanism on employment counts for NAICS-5 and NAICS-3, although the difference between the performance of the two mechanisms has become very small. However, for total wages grouped by NAICS-3 or NAICS-5, we see that $\sqrt{\ }$-mechanism clearly outperforms pnc-mechanism. This suggests that wages data are more highly skewed than employment count and an establishment is more likely to dominate the total wages in its industry than the total employment.

The exact relationship between a group (in a group-by query) and which mechanism works best for it depends on the skewness in a group, the distance parameters, and privacy parameters. An interesting direction for future work is to use the confidentiality-preserving answers to the identity query to help estimate skewness in all groups of all group-by queries and (along with the public privacy parameters) to determine whether the pnc-mechanism or $\sqrt{\ }$-mechanism should be used for a particular query.

9 Conclusions and Future Work

In this paper, we introduced Gaussian Establishment DP, a privacy definition for establishment data, designed to be interpretable to mathematical statisticians in statistical agencies. We proposed two mechanisms, the ψ-mechanism with non-releasable data-dependent variance, and the pnc-mechanism with *releasable* data-dependent variance.

Future work includes mechanisms that answer queries over multiple confidential attributes. Another interesting direction is how to optimize a workload – adaptively selecting measurement queries and privacy budgets in order to maximize accuracy. It is likely to be an iterative process, where prior query answers are used to estimate the expected gain in accuracy that can be obtained from candidate future measurement queries.

Acknowledgments

This work was supported by a NSF BAA grant (NCSES-BAA 49100421C0022 to The Pennsylvania State University). We thank David Talan for help with data access and Spencer Jobe for help with running our experiments on internal BLS data.

References

[1] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, and Pavel Zhuravlev. 2021. An uncertainty principle is a price of privacy-preserving microdata. *Advances in neural information processing systems* 34 (2021), 11883–11895.

[2] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. 2022. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review* 2 (2022).

[3] Nabil R Adam and John C Worthmann. 1989. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)* 21, 4 (1989), 515–556.

[4] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. 2020. Hypothesis Testing Interpretations and Renyi Differential Privacy. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 2496–2506. https://proceedings.mlr.press/v108/balle20a.html

[5] Scott A Brave, Charles Gascon, William Kluender, and Thomas Walstrum. 2021. Predicting benchmarked US state employment data in real time. *International Journal of Forecasting* 37, 3 (2021), 1261–1275.

[6] Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*. Springer, 635–658.

[7] U.S. Census Bureau. (retrieved May 14, 2024). County Business Patterns 2016. https://www.census.gov/data/datasets/2016/econ/cbp/2016-cbp.html.

[8] U.S. Census Bureau. (retrieved May 14, 2024). Quarterly Workforce Indicators 2016. https://ledextract.ces.census.gov/qwi/all.

[9] George Casella and Roger L. Berger. 2002. *Statistical Infence* (2 ed.). Duxbury.

[10] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *13th International Symposium (PETS)*.

[11] Steve Cohen and Bogong T Li. 2006. A Comparison of Data Utility between Publishing Cell Estimates as Fixed Intervals or Estimates based upon a Noise Model versus Traditional Cell Suppression on Tabular Employment Data. https://www.bls.gov/osmr/research-papers/2006/pdf/st060100.pdf.

[12] Lawrence H Cox. 1980. Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.* 75, 370 (1980), 377–385.

[13] Lawrence H. Cox. 1981. Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference* 5, 2 (1981), 153–164.

[14] Michael Dalton, Elizabeth Weber Handwerker, and Mark A Loewenstein. 2020. Employment changes by employer size using the current employment statistics survey microdata during the COVID-19 pandemic. *Covid Economics* 46 (Sept. 2020), 123–136.

[15] Jinshuo Dong, Aaron Roth, and Weijie Su. 2021. Gaussian Differential Privacy. *Journal of the Royal Statistical Society* (2021).

[16] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*. Springer, 486–503.

[17] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[18] Fabian Eckert, Teresa C Fort, Peter K Schott, and Natalie J Yang. 2020. County Business Patterns Database. https://doi.org/10.3886/E117464V1

[19] Fabian Eckert, Teresa C Fort, Peter K Schott, and Natalie J Yang. 2020. *Imputing Missing Values in the US Census Bureau's County Business Patterns*. Working Paper 26632. National Bureau of Economic Research. https://doi.org/10.3386/w26632

[20] Timothy Evans, Laura Zayatz, and John Slanta. 1996. Using noise for disclosure limitation of establishment tabular data. In *Proceedings of the Annual Research Conference, US Bureau of the Census, Washington, DC*, Vol. 20233. 65–86.

[21] Brian Finley, Anthony M Caruso, Justin C Doty, Ashwin Machanavajjhala, Mikaela R Meyer, David Pujol, William Sexton, and Zachary Terner. 2024. Slowly Scaling Per-Record Differential Privacy. *arXiv preprint arXiv:2409.18118* (2024).

[22] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. 2008. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 265–273.

[23] Samuel Haney, Ashwin Machanavajjhala, John M Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. Utility cost of formal privacy for releasing national employer-employee statistics. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1339–1354.

[24] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. 2009. Boosting the accuracy of differentially-private histograms through consistency. *arXiv preprint arXiv:0904.0942* (2009).

[25] Scott H. Holan, Daniell Toth, Marco A. R. Ferreira, and Alan F. Karr. 2010. Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality. *J. Amer. Statist. Assoc.* 105, 490 (2010), 564–577. https://doi.org/10.1198/jasa.2009.ap08629

[26] Zach Jorgensen, Ting Yu, and Graham Cormode. 2015. Conservative or liberal? Personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*. IEEE, 1023–1034.

[27] Satkartar K Kinney, Jerome P Reiter, Arnold P Reznek, Javier Miranda, Ron S Jarmin, and John M Abowd. 2011. Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International statistical review* 79, 3 (2011), 362–384.

[28] H. O. Lancaster. 1969. *The Chi-Squared Distribution*. Hutchinston Educational.

[29] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. 2015. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal* 24 (2015), 757–781.

[30] Jonathan Meer and Jeremy West. 2016. Effects of the minimum wage on employment dynamics. *Journal of Human Resources* 51, 2 (2016), 500–522.

[31] Microsoft. 2025. Facts about Miscrosoft. https://news.microsoft.com/facts-about-microsoft/.

[32] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 263–275.

[33] Jack Murtagh and Salil Vadhan. 2015. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*. Springer, 157–175.

[34] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 75–84.

[35] Michelle Pistner, Aleksandra Slavković, and Lars Vilhuber. 2018. Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In *International Conference on Privacy in Statistical Databases*. Springer, 92–108.

[36] Prottay Protivash, John Durrell, Kaitlyn Webb, and Daniel Kifer. 2026. bls_privacy_public. https://github.com/cmla-psu/bls_privacy_public.

[37] Disneyland Resorts. (retrieved June 20, 2025). Frequently asked questions about working at the Disneyland Resort. https://disneyexperiences.com/disneyland/our-cast/frequently-asked-questions-about-working-at-the-disneyland-resort/.

[38] Ralph Tyrell Rockafellar. 2015. Convex analysis:(pms-28). (2015).

[39] Akbar Sadeghi and Kevin Cooksey. 2021. Business employment dynamics by wage class. *Monthly Lab. Rev.* 144 (2021), 1.

[40] Akbar Sadeghi, David M. Talan, and Richard L. Clayton. 2016. Establishment, firm, or enterprise: does the unit of analysis matter? *Monthly Labor Review* (2016). https://doi.org/10.21916/mlr.2016.51

[41] Lester M Salamon and S Wojciech Sokolowski. 2005. Nonprofit organizations: New insights from QCEW data. *Monthly Lab. Rev.* 128 (2005), 19.

[42] Ian M Schmutte and Lars Vilhuber. 2020. Balancing privacy and data usability: An overview of disclosure avoidance methods. *Handbook on Using Administrative Data for Research and Evidence-Based Policy* (2020).

[43] Jeremy Seeman, William Sexton, David Pujol, and Ashwin Machanavajjhala. 2024. Privately Answering Queries on Skewed Data via Per-Record Differential Privacy. *Proc. VLDB Endow.* 17, 11 (Aug. 2024), 3138–3150. https://doi.org/10.14778/3681954.3681989

[44] Adam Smith and Abhradeep Thakurta. 2022. Fully Adaptive Composition for Gaussian Differential Privacy. https://doi.org/10.48550/ARXIV.2210.17520

[45] Daniell Toth. 2014. Data smearing: an approach to disclosure limitation for tabular data. *Journal of official statistics* 30, 4 (2014), 839–857.

[46] Tran Tran, Matthew Reimherr, and Aleksandra Slavković. 2023. Differentially Private Synthetic Heavy-tailed Data. arXiv:2309.02416 [stat.AP] https://arxiv.org/abs/2309.02416

[47] U.S. Bureau of Labor Statistics. (retrieved August 21, 2024). Handbook of Methods: Quarterly Census of Employment and Wages. https://www.bls.gov/opub/hom/cew/.

[48] U.S. Bureau of Labor Statistics. (retrieved August 21, 2024). QCEW Public-Use Data Files. https://www.bls.gov/cew/downloadable-data-files.htm.

[49] U.S. Census Bureau. (retrieved August 21, 2024). North American Industry Classification System. https://www.census.gov/naics/.

[50] Yasutoshi Washio, Haruki Morimoto, and Nobuyuki Ikeda. 1956. Unbiased estimation based on sufficient statistics. *Bulletin of Mathematical Statistics* 6 (1956), 69–93.

[51] Larry Wasserman and Shuheng Zhou. 2010. A Statistical Framework for Differential Privacy. *J. Amer. Statist. Assoc.* 105, 489 (2010), 375–389. http://www.jstor.org/stable/29747034

[52] Y.M. Yang, A. Mushtaq, S. Pramanik, F. Scheuren, D. Hiles, M. Buso, and S. Butani. 2011. Evaluation of Three Disclosure Limitation Models for the QCEW Program. BLS Research Papers https://www.bls.gov/osmr/research-papers/2011/pdf/st110120.pdf.

[53] Y. Michael Yang, Ali Mushtaq, Santanu Pramanik, Fritz Scheuren, David Hiles, Michael Buso, and Shail Butani. 2011. An Evaluation of BLS Noise Research

for the Quarterly Census of Employment and Wages. BLS Research Papers https://www.bls.gov/osmr/research-papers/2011/pdf/st110120.pdf.

A Analyzing attacker error due to sampling.

The following example demonstrates why $\psi = \sqrt{}$ makes statistical sense and how it relates to estimating establishment count.

Example A.1. Suppose an establishment $\mathcal{E}$ has x employees and a statistician (e.g., an attacker or labor economist) wants to estimate its employment count. Suppose nearly all employees of $\mathcal{E}$ live within a region whose total population, N, is known with high accuracy (e.g., from Census data). For some $q \in [0, 1]$, the statistician may take a random sample of $n = Nq$ people in the region and record the proportion p that work for $\mathcal{E}$. The quantity np can be modeled as the sum of n Bernoulli(x/N) random variables. So, an unbiased estimate of the number of employees of $\mathcal{E}$ is $\hat{x} = np(N/n)$. The variance of this estimate is $\text{Var}(\hat{x}) = n(x/N)(1 - x/N)N^2/n^2 = x\frac{N-x}{n} \leq \frac{1}{q}x$. The standard deviation is therefore $O(\sqrt{x})$ and confidence intervals would therefore have lengths $O(\sqrt{x})$.

B Discussion of Cell Suppression Weaknesses

The BLS is obligated to use statistical disclosure limitation methods to protect sensitive values (employment and wages) of individual establishments. Internal policy and federal law, specifically the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002, require confidentiality protections to all data acquired and maintained by the BLS under a pledge of confidentiality.

The current disclosure avoidance method for the QCEW is cell suppression. According to the cell suppression methodology [3], sensitive cells are initially identified by either the p% rule, the qp-rule or the nk-rule [42]. The p% rule, for example, considers the question of whether the second largest establishment in a cell can use the cell total to learn the largest establishment's confidential value up to p% **relative error** (a cell is flagged as sensitive if the cell total minus the contributions of the two largest establishments is less than p% of the largest establishment). These cells are suppressed in a step called *primary suppression*. The *secondary suppression* phase then identifies combinations of cells from which the suppressed cells can be deduced via linear algebra. This phase removes additional cells so that such combinations no longer exist.

Primary and secondary suppression procedures result in the BLS suppressing the majority of the cells before they are released [47]. Over 60% of more than 3.6 million cells in the QCEW tables are suppressed [11, 45, 52]. Apart from concerns about loss of data, secondary suppression can be complex to implement and often requires solving computationally expensive linear programs [12].

Cell suppression also has two major confidentiality vulnerabilities. First, it is brittle against external knowledge – it is possible for, say Pennsylvania, to release additional data about its establishments that, when combined with cell-suppressed QCEW data, allows attackers to reverse engineer other cells (e.g., compromising confidentiality for businesses in New York). Second, cell suppression is only designed to protect against *linear attackers* – those who only use addition, subtraction, and scalar multiplication to attack data. It is not designed to protect against more sophisticated statistical attacks [25], which is why the p parameter from the p% rule is also kept confidential. It is therefore not possible to realistically determine what confidential values might become vulnerable to disclosure when new types of data tabulations are published. This prevents the BLS from publishing new statistics and analyses from the QCEW microdata because of the unknown ways that new data products could impact the disclosure protections of existing products.

Table 4: Establishment microdata and three marginals computed from it. Primary suppression must remove the cells for County X and Sector 52 (red), as they are directly revealing of the wages $\mathbb{B}$. Secondary suppression is needed to prevent $\mathbb{B}$ from being reconstructed by the remaining cells.

Establishment	County	Sector	Wages
Bob's Building	X	52	$\mathbb{B}$
Carlos's Construction	Y	23	$\mathbb{C}$
Green Hills Farming	Y	11	$\mathbb{G}$
Isaac's Foundations	Z	23	$\mathbb{F}$
Alice's Apples	Z	11	$\mathbb{A}$

Total Wages

$\mathbb{B}$+$\mathbb{C}$+$\mathbb{G}$+$\mathbb{F}$+$\mathbb{A}$

By County

X	Y	Z
$\mathbb{B}$	$\mathbb{C}$+$\mathbb{G}$	$\mathbb{F}$+$\mathbb{A}$

By Sector

52	23	11
$\mathbb{B}$	$\mathbb{C}$+$\mathbb{F}$	$\mathbb{A}$+$\mathbb{G}$

Table 4 shows a simplified microdata set and 3 marginals computed from it: total wages, wages by county, and wages by sector (2-digit NAICS code) prefix. The values in cells for County X and also for Sector 52 (displayed in red) are directly disclosive because they only contain one establishment, and hence would undergo primary suppression.

The secondary suppression phase then identifies which non-suppressed cells can be used to deduce the cells suppressed in the first phase, or can isolate additional establishments. For example, the total wages minus the combined wages of counties Y and Z, or total wages minus the combined wages of sectors 23 and 11 recover the wages $\mathbb{B}$. Meanwhile, the wages in counties Y and Z along with the wages in sectors 23 and 11 form a linear system of 4 equations and 4 unknowns and hence is uniquely solvable. A subset of these problematic cells needs to be suppressed with the goal that the remaining cells cannot be used to reconstruct sensitive values. Thus, for example, one could additionally suppress the total wages table, along with total wages of County Y. All that remains to be published are the wages of County Z, Sector

23, and Sector 11. In our example 50% of the cells end up being suppressed. This is not an anomaly. Primary and secondary suppression procedures results in the BLS suppressing the majority of the cells before they are released [47].

C Deriving Variance from Power and Significance

We can frame differential privacy definitions through power and significance level similar to the GDP framework in Section 3. Let y be the output of a privacy-preserving mechanism $\mathcal{M}$. Consider the hypothesis test:

$$H_0 : y \text{ from } \mathcal{M}(\mathcal{D}_1) \qquad H_1 : y \text{ from } \mathcal{M}(\mathcal{D}_2) \tag{2}$$

for any pair of neighboring $\mathcal{D}_1$ and $\mathcal{D}_2$.

Fixing a significance and power level, we calculate the smallest variance for the privacy noise added by a standard mechanism to satisfy pure DP, Gaussian DP and zCDP.

C.1 Pure DP

If a mechanism,$\mathcal{M}$, satisfies ϵ-DP, also known as pure differential privacy, by definition

$$P(\mathcal{M}(\mathcal{D}_1) \in S) \leq e^{\epsilon} P(\mathcal{M}(\mathcal{D}_2) \in S)$$

for any pair of neighboring datasets, $\mathcal{D}_1, \mathcal{D}_2$ and $S \in \text{Range}(\mathcal{M})$ [17].

Since this result must hold for any pair of neighboring datasets and any S in the range of $\mathcal{M}$. Let y be the output of $\mathcal{M}(D_1)$ and let S be the rejection region. By definition of significance, $P(\mathcal{M}(\mathcal{D}_1) \in S) = \alpha$, where α is the significance level (or Type I error). Thus $\alpha \leq e^{\epsilon} P(\mathcal{M}(\mathcal{D}_2) \in S)$. However, $P(\mathcal{M}(\mathcal{D}_2) \in S) \leq 1 - \beta$, by definition of power. Since the definition must hold for any S in the range of $\mathcal{M}$, a similar result holds for the complement of S[6].

$$1 - \alpha \leq e^{\epsilon} \beta \tag{3}$$

$$\alpha \leq e^{\epsilon} (1 - \beta) \tag{4}$$

If we fix significance level α and power $(1 - \beta)$, we find an lower bound for ϵ.

From (3) we have

$$\frac{1-\alpha}{\beta} \leq e^{\epsilon} \implies \log\left(\frac{1-\alpha}{\beta}\right) \leq \epsilon$$

From (4), we have

$$\alpha \geq e^{-\epsilon}(1 - \beta) \implies \epsilon \geq \log\left(\frac{1-\beta}{\alpha}\right) \tag{5}$$

Thus $\epsilon \geq \max\left(\log\left(\frac{1-\alpha}{\beta}\right), \log\left(\frac{1-\beta}{\alpha}\right)\right)$

Let m be the number of queries. By composition properties of pure DP, each query should have a $\epsilon_i = \frac{\max\left(\log\left(\frac{1-\alpha}{\beta}\right), \log\left(\frac{1-\beta}{\alpha}\right)\right)}{m}$ to obtain the fixed significance and power levels.

If $\mathcal{M}$ is the commonly used Laplace mechanism [17], then the noise added to each query result has variance $2\left(\frac{m}{\epsilon_i}\right)^2$.

C.2 GDP

Once again, let S be a rejection region and $y \leftarrow \mathcal{M}(\mathcal{D}_1)$. A mechanism $\mathcal{M}$ satisfies μ-GDP if

$$\Phi^{-1}(P(\mathcal{M}(\mathcal{D}_1) \in S)) \leq \mu + \Phi^{-1}(P(\mathcal{M}(\mathcal{D}_2) \in S))$$

Thus

$$\Phi^{-1}(\alpha) \leq \mu + \Phi^{-1}(1 - \beta)$$

This means $\mu = \Phi^{-1}(\alpha) - \Phi^{-1}(1 - \beta)$ is the smallest privacy parameter that satisfies the given significance and power levels. For m queries, by composition properties each query must satisfy $\mu_i = \mu/\sqrt{m}$. The Gaussian mechanism satisfies μ_i-GDP if the variance of the infused noise is $1/(\mu_i^2)$.

[6] This result is also derived by Wasserman & Zhou [51].

C.3 zCDP

We omit the original definition of ρ-zCDP, but include the hypothesis testing interpretation of it [4]. Once again, let S be a rejection region and $y \leftarrow \mathcal{M}(\mathcal{D}_1)$. A mechanism $\mathcal{M}$ satisfies ρ-zCDP if

$$P(\mathcal{M}(\mathcal{D}_1) \in S)^{\eta} P(\mathcal{M}(\mathcal{D}_2) \in S)^{1-\eta} + P(\mathcal{M}(\mathcal{D}_1) \notin S)^{\eta} + P(\mathcal{M}(\mathcal{D}_2) \notin S)^{1-\eta} \leq e^{\rho\eta(\eta-1)}$$

for any $\eta > 1$ [4]. This is equivalent to

$$\alpha^{\eta}(1-\beta)^{1-\eta} + (1-\alpha)^{\eta}\beta^{1-\eta} \leq e^{\rho\eta(\eta-1)} \tag{6}$$

Thus

$$\rho \geq \max_{\eta>1} \left(\frac{\log\left(\alpha^{\eta}(1-\beta)^{1-\eta} + (1-\alpha)^{\eta}\beta^{1-\eta}\right)}{\eta(\eta-1)} \right)$$

This is solved numerically to find the smallest ρ for a fixed power and significance level. Denote this solution as ρ^*. By composition properties, for m queries, each query should satisfy $\rho_i = \rho^*/m$.

A typical mechanism that satisfies ρ_i-zCDP is the Gaussian mechanism with variance $1/2\rho_i$ [6].

D Proofs from Section 5

Theorem 5.4. Let ψ be a function satisfying the conditions of Definition 5.2. Then for any x, y, t such that $0 \leq x < y$ and $t \geq 0$

(1) ψ^{-1} exists and is increasing.
(2) We have $|\psi(y+t) - \psi(x+t)| \leq |\psi(y) - \psi(x)|$.
(3) The length of the uncertainty interval around y is at least as big as the interval around x: $|\mathcal{I}_{\psi,t}(y)| \geq |\mathcal{I}_{\psi,t}(x)|$.
(4) The length of the uncertainty interval around x relative to x is at least as big as the length of the uncertainty interval around y relative to y, $\frac{|\mathcal{I}_{\psi,t}(y)|}{y} \leq \frac{|\mathcal{I}_{\psi,t}(x)|}{x}$

Proof of Theorem 5.4. To prove Item 1, we note that since ψ is increasing, the inverse must exist and clearly it is increasing as well (since $x < y \Leftrightarrow \psi(x) < \psi(y)$).

To prove Item 2, we take advantage of the concavity of ψ. Given numbers $a < b < c$, we note that $b = a\left(\frac{c-b}{c-a}\right) + c\left(\frac{b-a}{c-a}\right) = a\left(\frac{c-b}{c-a}\right) + c\left(1 - \frac{c-b}{c-a}\right)$

$$\begin{aligned}
&\frac{\psi(c)-\psi(b)}{c-b} - \frac{\psi(b)-\psi(a)}{b-a} \\
&= \frac{\psi(c) - \psi\left(a\left(\frac{c-b}{c-a}\right) + c\left(1-\frac{c-b}{c-a}\right)\right)}{c-b} - \frac{\psi\left(a\left(\frac{c-b}{c-a}\right) + c\left(1-\frac{c-b}{c-a}\right)\right) - \psi(a)}{b-a} \\
&\leq \frac{\psi(c) - \psi(a)\left(\frac{c-b}{c-a}\right) - \psi(c)\left(1-\frac{c-b}{c-a}\right)}{c-b} - \frac{\psi(a)\left(\frac{c-b}{c-a}\right) + \psi(c)\left(1-\frac{c-b}{c-a}\right) - \psi(a)}{b-a} \\
&\text{by concavity of } \psi \\
&= \frac{(\psi(c)-\psi(a))\left(\frac{c-b}{c-a}\right)}{c-b} - \frac{(\psi(c)-\psi(a))\left(1-\frac{c-b}{c-a}\right)}{b-a} \\
&= (\psi(c)-\psi(a))\left(\frac{1}{c-a}\right) - \frac{(\psi(c)-\psi(a))\left(\frac{b-a}{c-a}\right)}{b-a} \\
&= (\psi(c)-\psi(a))\left(\frac{1}{c-a}\right) - (\psi(c)-\psi(a))\left(\frac{1}{c-a}\right) = 0
\end{aligned}$$

and thus $\frac{\psi(c)-\psi(b)}{c-b} \leq \frac{\psi(b)-\psi(a)}{b-a}$ when $c > b > a$. Noting that item 2 is trivial when $t = 0$, we only need to consider $t > 0$. The first case we consider is when $y = x + t$, so that the intervals $[x, y]$ and $[x + t, y + t]$ are adjacent. In this case, we directly apply the derived relation to get:

$$0 \leq \frac{\psi(y+t)-\psi(x+t)}{y-x} \leq \frac{\psi(x+t)-\psi(x)}{x+t-x} = \frac{\psi(y)-\psi(x)}{y-x}$$

from which we get $|\psi(y+t) - \psi(x+t)| \leq |\psi(y) - \psi(x)|$. The second case we consider is when $[x + t, y + t]$ and $[x, y]$ are disjoint (so that $x < y < x + t < y + t$). Applying the derived relation twice, we get:

$$0 \leq \frac{\psi(y+t)-\psi(x+t)}{y-x} \leq \frac{\psi(x+t)-\psi(y)}{x+t-y} \leq \frac{\psi(y)-\psi(x)}{y-x}$$

from which we get $|\psi(y+t)-\psi(x+t)| \le |\psi(y)-\psi(x)|$. The remaining case to consider is when $[x+t, y+t]$ and $[x, y]$ overlap at more than just an endpoint (so that $x < x+t < y < y+t$). In this case,

$$\begin{aligned}
&\Big(\psi(y+t)-\psi(x+t)\Big) - \Big(\psi(y)-\psi(x)\Big) \\
&= \Big(\psi(y+t)-\psi(y)+\psi(y)-\psi(x+t)\Big) - \Big(\psi(y)-\psi(x+t)+\psi(x+t)-\psi(x)\Big) \\
&= \Big(\psi(y+t)-\psi(y)\Big) - \Big(\psi(x+t)-\psi(x)\Big)
\end{aligned}$$

apply the derived relation twice (as we did in the case of the disjoint intervals), we get

$$0 \le \frac{\psi(y+t)-\psi(y)}{t} \le \frac{\psi(x+t)-\psi(x)}{t}$$

and this allows us to conclude that $|\psi(y+t)-\psi(x+t)| \le |\psi(y)-\psi(x)|$.

To prove Item 3, we see that the case $t = 0$ is trivial and so we can assume $t > 0$. Next, we note that since ψ is increasing and concave, then ψ^{-1} is increasing and convex. To see why, concavity implies $\psi(ax+(1-a)y) \ge a\psi(x)+(1-a)\psi(y)$ and applying ψ^{-1} to both sides, we get $ax+(1-a)y \ge \psi^{-1}(a\psi(x)+(1-a)\psi(y))$. Replacing $\psi(x)$ with x' and $\psi(y)$ with y', we have $a\psi^{-1}(x')+(1-a)\psi^{-1}(y') \ge \psi^{-1}(ax'+(1-a)y')$.

Next, following the proof of Item 2, but replacing concavity with convexity, it follows that given numbers a, b, c, d with a being the smallest and d being the largest, we have $\frac{\psi^{-1}(d)-\psi^{-1}(c)}{d-c} \ge \frac{\psi^{-1}(b)-\psi^{-1}(a)}{b-a}$ (i.e., convexity not only implies that derivatives are non-decreasing, but same with secants). We consider several cases based on the relationship between $\psi(x)-t$, $\psi(0)$, and $\psi(y)-t$. The case is trivial if $\psi(x)-t < \psi(y)-t \le \psi(0)$. Next we consider the case where $\psi(0) \le \psi(x)-t \le \psi(y)-t$. Substituting $d = \psi(y)+t$ and $c = \psi(y)-t$ and $b = \psi(x)+t$ and $a = \psi(x)-t$, then noting that $d-c = b-a = 2t > 0$, we get $\psi^{-1}(\psi(y)+t)-\psi^{-1}(\psi(y)-t) \ge \psi^{-1}(\psi(x)+t)-\psi^{-1}(\psi(x)-t) \ge 0$. Finally, if $\psi(x)-t \le \psi(0) \le \psi(y)-t$, we substitute $a = \psi(0)$ and $c = \psi(y)-t$ and get $\frac{\psi^{-1}(\psi(y)+t)-\psi^{-1}(\psi(y)-t)}{2t} \ge \frac{\psi^{-1}(\psi(x)+t)}{\psi(x)-\psi(0)+t}$. Since $\psi(x)-\psi(0)+t \le \psi(x)-(\psi(x)-t)+t = 2t$, we get $\psi^{-1}(\psi(y)+t)-\psi^{-1}(\psi(y)-t) \ge \psi^{-1}(\psi(x)+t)-\psi^{-1}(\psi(x)-t) \ge 0$.

To prove Item 4, we note that if $g(a) = \psi(\exp(a))$ is convex then $h(b) = g^{-1}(b) = \log(\psi^{-1}(b))$ is concave, and both g and h are increasing functions. Thus we know that for numbers a, b, c, d with a being the smallest and d being the largest, $\frac{h(d)-h(c)}{d-c} \le \frac{h(b)-h(a)}{b-a}$. This means that

$$\frac{1}{d-c}\log\frac{\psi^{-1}(d)}{\psi^{-1}(c)} \le \frac{1}{b-a}\log\frac{\psi^{-1}(b)}{\psi^{-1}(a)}$$

Substituting $d = \psi(y)+t$ and $c = \psi(y)$ and $b = \psi(x)+t$ and $a = \psi(x)$, then noting that $d-c = b-a = t > 0$, we get

$$\frac{\psi^{-1}(\psi(y)+t)}{y} \le \frac{\psi^{-1}(\psi(x)+t)}{x} \tag{7}$$

Thus the result holds when $\psi(x)-t < \psi(y)-t \le \psi(0)$. When $\psi(0) < \psi(x)-t < \psi(y)-t$, we substitute $d = \psi(y)$ and $c = \psi(y)-t$ and $b = \psi(x)$ and $a = \psi(x)-t$, then noting that $d-c = b-a = t > 0$, we get $\frac{y}{\psi^{-1}(\psi(y)-t)} \le \frac{x}{\psi^{-1}(\psi(x)-t)}x$. Since $\psi(x)-t > \psi(0)$ implies $\psi^{-1}(\psi(y)-t) > \psi^{-1}(\psi(x)-t) > 0$, we find

$$\frac{\psi^{-1}(\psi(y)-t)}{y} \ge \frac{\psi^{-1}(\psi(x)-t)}{x} \tag{8}$$

Thus,

$$\begin{aligned}
&\frac{\psi^{-1}(\psi(y)+t)}{y} - \frac{\psi^{-1}(\psi(y)-t)}{y} \\
&\le \frac{\psi^{-1}(\psi(x)+t)}{x} - \frac{\psi^{-1}(\psi(y)-t)}{y} && \text{by (7)} \\
&\le \frac{\psi^{-1}(\psi(x)+t)}{x} - \frac{\psi^{-1}(\psi(x)-t)}{x} && \text{by (8).}
\end{aligned}$$

The result holds when $\psi(0) < \psi(x)-t < \psi(y)-t$. Finally, when $\psi(x)-t \le \psi(0) \le \psi(y)-t$, we know $\psi^{-1}(\psi(y)-t) \ge 0$. Thus,

$$\begin{aligned}
&\frac{\psi^{-1}(\psi(y)+t)}{y} - \frac{\psi^{-1}(\psi(y)-t)}{y} \\
&\le \frac{\psi^{-1}(\psi(x)+t)}{x} - \frac{\psi^{-1}(\psi(y)-t)}{y} && \text{by (7)} \\
&\le \frac{\psi^{-1}(\psi(x)+t)}{x}
\end{aligned}$$

Therefore $\frac{\psi^{-1}(\psi(y)+t)}{y} - \frac{\psi^{-1}(\max(\psi(0),\psi(y)-t)}{y} \le \frac{\psi^{-1}(\psi(x)+t)}{y} - \frac{\psi^{-1}(\max(\psi(0),\psi(x)-t)}{x}$ for all cases of $0 \le x < y$ and $t \ge 0$. □

It is important to note that a piecewise linear function ψ_{pl} with 2 or more pieces *cannot* be a neighbor function, even if it is concave and increasing. The reason is that $\psi_{pl}(\exp(x))$ is not convex and hence will not allow relative error for large establishments to decrease.

Theorem D.1. Let ψ_{pl} be a non-decreasing concave piece-wise linear function with 2 or more pieces. Then ψ_{pl} is not a neighbor function.

Proof of Theorem D.1. A concave function defined on an interval is continuous except possibly at the endpoints. Thus, it a concave piecewise linear function has at least 2 pieces, there exists a point x_0 where two pieces meet. Thus in a small neighborhood to the left of x_0, the function $\psi_{pl}(x) = a_1x + b_1$ for some a_1, b_1 and in a small neighborhood to the right of x_0, we have $\psi_{pl}(x) = a_2x + b_2$ for some a_2, b_2. Importantly, by continuity, $a_1x_0 + b_1 = a_2x_0 + b_2$. Since $\psi_{pl}(x) = a_1x + b_1$ is concave, $a_1 > a_2$.

Now, the function $\psi_{pl}(\exp(x))$ is equal to $e^{a_1x+b_1}$ in a small neighborhood to the left of x_0, and has derivative $a_1e^{a_1x+b_1}$. Meanwhile, $\psi_{pl}(\exp(x))$ is equal to $e^{a_2x+b_2}$ in a small neighborhood to the right and has derivative $a_2e^{a_2x+b_2}$. Thus, left derivative at x_0 is $a_1/a_2 > 1$ times bigger than the right derivative at x_0. This means that as we approach and pass x_0, the derivative of $\psi_{pl}(\exp(x))$ decreases and hence it cannot be convex.

Thus ψ_{pl} is not a neighbor function. □

Theorem 5.8 (Adaptive Composition)**.** Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a sequence of mechanisms such that $\mathcal{M}_{i+1}$ can access the output of $\mathcal{M}_1, \ldots, \mathcal{M}_i$. Suppose each $\mathcal{M}_i$ satisfies μ_i-gedp with the same neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$. Then the mechanism that releases all of their outputs satisfies $\sqrt{\sum_{i=1}^{K} \mu_i^2}$-gedp with the same neighbor function and distance parameters.

Proof of Theorem 5.8. This is proven in the same way as adaptive composition in Gaussian Differential Privacy [15]. □

Theorem 5.10. Let $\psi^{(A)}$ and $\psi^{(B)}$ be neighbor functions satisfying the conditions of Definition 5.2. Let $\gamma^{(A)}$ and $\gamma^{(B)}$ be positive numbers. Then $\psi^{(A)}$ and $\psi^{(B)}$ are continuously differentiable almost everywhere. Set $\gamma^* = 1$ and define ψ^* as follows:

$$\frac{d\psi^*(x)}{dx} = \min\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x)}{dx},\ \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x)}{dx}\right)$$

$$\psi^*(x) = \int_0^x \frac{d\psi^*(x)}{dx}\,dx$$

Then ψ^* satisfies the conditions of Definition 5.2 to be a neighbor function. Furthermore, the uncertainty intervals defined by $\psi^{(A)}, \gamma^{(A)}$ and $\psi^{(B)}, \gamma^{(B)}$ are contained in the uncertainty interval defined by $\psi^*, \gamma*$. That is for any x, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \subseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma^{(B)}}(x) \subseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

Proof of Theorem 5.10. Continuous differentiability almost everywhere is a result of concavity of the functions [38] (Theorem 25.5). Thus, the set of points at which $\frac{d\psi}{dx}$ can not be defined has measure 0 and the set contains at most countable many points.

To show that ψ^* is increasing, we note that $\psi^{(A)}$ and $\psi^{(B)}$ are increasing, so their derivative is positive almost everywhere, hence the derivative of ψ^* is positive almost everywhere and hence ψ^* is increasing.

Next, ψ^* is continuous because it is defined as an integral.

Next, we show that ψ^* is concave. Since $\psi^{(A)}$ and $\psi^{(B)}$ are concave, their derivatives are non-increasing. The same is true for ψ^* since for $x_1 < x_2$, we have

$$\begin{aligned}\frac{d\psi^*(x_1)}{dx} &= \min\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x_1)}{dx},\ \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x_1)}{dx}\right)\\ &\geq \min\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x_2)}{dx},\ \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x_2)}{dx}\right)\\ &= \frac{d\psi^*(x_2)}{dx}\end{aligned}$$

and hence ψ^* is also concave.

Next, we show that $\psi^*(\exp(x))$ is convex as a function of x. We note that if $y = \exp(x)$ then $\frac{d\psi^{(A)}(\exp(x))}{dx} = e^x\frac{d\psi^{(A)}(y)}{dy} = y\frac{d\psi^{(A)}(y)}{dy}$. Hence the fact that $\psi^{(A)}(\exp(x))$ and $\psi^{(B)}(\exp(x))$ are convex means that $y\frac{d\psi^{(A)}(y)}{dy}$ and $y\frac{d\psi^{(B)}(y)}{dy}$ are non-decreasing functions of y when $y > 0$. Thus for $0 < y_1 < y_2$,

$$\begin{aligned}y_1\frac{d\psi^*(y_1)}{dy} &= \min\left(y_1\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(y_1)}{dy},\ y_1\frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(y_1)}{dy}\right)\\ &\leq \min\left(y_2\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(y_2)}{dy},\ y_2\frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(y_2)}{dy}\right)\end{aligned}$$

$$= y_2 \frac{d\psi^*(y_2)}{dy}$$

hence $y\frac{d\psi^*(y)}{dy}$ is also non-decreasing when $y > 0$ and it follows that $\psi^*(\exp(x))$ is convex.

To prove containment of uncertainty intervals, we first show that for $x_1 < x_2$, both $\frac{\psi^*(x_2)-\psi^*(x_1)}{\gamma^*} \leq \frac{\psi^{(A)}(x_2)-\psi^{(A)}(x_1)}{\gamma^{(A)}}$ and $\frac{\psi^*(x_2)-\psi^*(x_1)}{\gamma^*} \leq \frac{\psi^{(B)}(x_2)-\psi^{(B)}(x_1)}{\gamma^{(B)}}$. To see why, note that

$$\begin{aligned}
\frac{\psi^*(x_2) - \psi^*(x_1)}{\gamma^*} &= \psi^*(x_2) - \psi^*(x_1) \\
&= \int_{x_1}^{x_2} \frac{d\psi^*(x)}{dx}\, dx \\
&\leq \frac{1}{\gamma^{(A)}} \int_{x_1}^{x_2} \frac{d\psi^{(A)}(x)}{dx} \\
&= \frac{\psi^{(A)}(x_2) - \psi^{(A)}(x_1)}{\gamma^{(A)}}
\end{aligned}$$

and same with $\psi^{(A)}$ replaced by $\psi^{(B)}$.

Now pick a point x and let $z^{(+A)} = (\psi^{(A)})^{-1}\left(\psi^{(A)}(x) + \gamma^{(A)}\right)$ and also let $z^{(+*)} = (\psi^*)^{-1}(\psi^*(x) + \gamma^*)$. Then:

$$\begin{aligned}
1 &= \frac{\psi^{(A)}(x) + \gamma^{(A)} - \psi^{(A)}(x)}{\gamma^{(A)}} \\
&= \frac{\psi^{(A)}(z^{(+A)}) - \psi^{(A)}(x)}{\gamma^{(A)}} \\
&\geq \frac{\psi^*(z^{(+A)}) - \psi^*(x)}{\gamma^*} \\
&= \psi^*(z^{(+A)}) - \psi^*(x)
\end{aligned}$$

but, $1 = \psi^*(z^{(+*)}) - \psi^*(x)$ and since ψ^* is increasing, this means $z^{(+*)} \geq z^{(+A)}$. Hence the right endpoint of the uncertainty interval determined by the combination $\psi^{(A)}, \gamma^{(A)}$ is less than or equal to the right endpoint of the uncertainty interval determined by the combination ψ^*, γ^*.

The left endpoint is trickier because it has a more complex expression. Pick a point x. Define $\gamma_L^{(A)} = \min(\gamma^{(A)}, \quad \psi^{(A)}(x) - \psi^{(A)}(0))$. This means the uncertainty interval's left endpoint is

$$\begin{aligned}
z^{(-A)} &= (\psi^{(A)})^{-1}\left(\max(\psi^{(A)}(0), \quad \psi^{(A)}(x) - \gamma^{(A)})\right) \\
&= (\psi^{(A)})^{-1}\left(\psi^{(A)}(x) - \gamma_L^{(A)}\right)
\end{aligned}$$

Similarly, define

$$\begin{aligned}
\gamma_L^* &= \min(\gamma^*, \quad \psi^*(x) - \psi^*(0)) \\
z^{(-*)} &= (\psi^*)^{-1}(\max(\psi^*(0), \quad \psi^*(x) - \gamma^*)) \\
&= (\psi^*)^{-1}\left(\psi^*(x) - \gamma_L^*\right)
\end{aligned}$$

Now, if $z^{(-*)} = 0$ then the left endpoint of the uncertainty interval associated with ψ^* and γ^* is the smallest possible and hence $\leq$ the left endpoint associated with $\psi^{(A)}$ and $\gamma^{(A)}$. Hence, **for the rest of the proof, we just need to consider the case where $\mathbf{z^{(-*)} = (\psi^*)^{-1}(\psi^*(x) - \gamma^*)}$.**

Then we have:

$$\begin{aligned}
1 &= \frac{\psi^{(A)}(x) - (\psi^{(A)}(x) - \gamma_L^{(A)})}{\gamma_L^{(A)}} \\
&= \frac{\psi^{(A)}(x) - \psi^{(A)}(z^{(-A)})}{\gamma_L^{(A)}} \\
&= \frac{1}{\gamma_L^{(A)}} \int_{r=z^{(-A)}}^{x} \frac{d\psi^{(A)}(r)}{dr}\, dr \\
&\geq \frac{1}{\gamma_L^{(A)}} \int_{r=z^{(-A)}}^{x} \gamma^{(A)} \frac{d\psi^*(r)}{dr}\, dr \quad \text{(by construction of } \psi^*)
\end{aligned}$$

$$= \frac{\gamma^{(A)}}{\gamma_L^{(A)}} \left(\psi^*(x) - \psi^*(z^{(-A)})\right)$$
$$\geq \psi^*(x) - \psi^*(z^{(-A)}) \quad (\text{since } \gamma^{(A)} \geq \gamma_L^{(A)})$$

But, since we are dealing with the case that $z^{(-*)} = (\psi^*)^{-1}(\psi^*(x) - \gamma^*)$, we have:

$$1 = \psi^*(x) - (\psi^*(x) - \gamma^*)$$
$$= \psi^*(x) - \psi^*(z^{(-*)})$$

So this means $\psi^*(x) - \psi^*(z^{(-*)}) \geq \psi^*(x) - \psi^*(z^{(-A)})$ and so $\psi^*(z^{(-A)}) \geq \psi^*(z^{(-*)})$. Since ψ^* is increasing, this means $z^{(-A)} \geq z^{(-*)}$ and so the left endpoint associated with $\psi^{(A)}$ and $\gamma^{(A)}$ is $\geq$ the left endpoint associated with ψ^* and γ^*. Hence, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \subseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

A similar argument holds for $\psi^{(B)}$.

□

Example D.2. We show how to combine the neighbor function $\psi^{(DP)}(x) = x$ with distance parameter d_1 with neighbor function $\psi^{(E)}(x) = \sqrt{x}$ with distance parameter γ_1. First, we set $\gamma^* = 1$. Then we note that

$$\frac{1}{d_1}\frac{d\psi^{(DP)}(x)}{dx} = \frac{1}{d_1}$$
$$\frac{1}{\gamma_1}\frac{d\psi^{(E)}(x)}{dx} = \frac{1}{2\gamma_1\sqrt{x}}$$
$$\frac{d\psi^*(x)}{dx} = \begin{cases} \frac{1}{d_1} & \text{if } x \leq \frac{d_1^2}{4\gamma_1^2} \\ \frac{1}{2\gamma_1\sqrt{x}} & \text{if } x \geq \frac{d_1^2}{4\gamma_1^2} \end{cases}$$
$$\psi^*(x) = \int_0^x \frac{d\psi^*(t)}{dt}\, dt$$
$$= \begin{cases} \frac{x}{d_1} & \text{if } x \leq \frac{d_1^2}{4\gamma_1^2} \\ \frac{\sqrt{x}}{\gamma_1} - \frac{d_1}{4\gamma_1^2} & \text{if } x \geq \frac{d_1^2}{4\gamma_1^2} \end{cases}$$

Theorem 5.11. Let $\psi^{(A)}$ and $\psi^{(B)}$ be neighbor functions satisfying the conditions of Definition 5.2. Let $\gamma^{(A)}$ and $\gamma^{(B)}$ be positive numbers. Then $\psi^{(A)}$ and $\psi^{(B)}$ are continuously differentiable almost everywhere. Set $\gamma^* = 1$ and define ψ^* as follows:

$$\frac{d\psi^*(x)}{dx} = \max\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x)}{dx}, \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x)}{dx}\right)$$
$$\psi^*(x) = \int_0^x \frac{d\psi^*(x)}{dx}\, dx$$

Then ψ^* satisfies the conditions of Definition 5.2 to be a neighbor function. Furthermore, the uncertainty intervals defined by $\psi^{(A)}, \gamma^{(A)}$ and $\psi^{(B)}, \gamma^{(B)}$ contain the uncertainty interval defined by ψ^*, γ^*. That is for any x, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \supseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$ and $\mathcal{I}_{\psi^{(B)},\gamma^{(B)}}(x) \supseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

Proof of Theorem 5.11. Continuous differentiability almost everywhere is a result of concavity of the functions [38] (Theorem 25.5). Thus, the set of points at which $\frac{d\psi}{dx}$ can not be defined has measure 0 and the set contains at most countably many points.

To show that ψ^* is increasing, we note that $\psi^{(A)}$ and $\psi^{(B)}$ are increasing, so their derivative is positive almost everywhere, hence the derivative of ψ^* is positive almost everywhere and hence ψ^* is increasing.

Next, ψ^* is continuous because it is defined as an integral.

Next, we show that ψ^* is concave. Since $\psi^{(A)}$ and $\psi^{(B)}$ are concave, their derivatives are non-increasing. The same is true for ψ^* since for $x_1 < x_2$, we have

$$\frac{d\psi^*(x_1)}{dx} = \max\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x_1)}{dx}, \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x_1)}{dx}\right)$$
$$\geq \max\left(\frac{1}{\gamma^{(A)}}\frac{d\psi^{(A)}(x_2)}{dx}, \frac{1}{\gamma^{(B)}}\frac{d\psi^{(B)}(x_2)}{dx}\right)$$
$$= \frac{d\psi^*(x_2)}{dx}$$

and hence ψ^* is also concave.

Next, we show that $\psi^*(\exp(x))$ is convex as a function of x. We note that if $y = \exp(x)$ then $\frac{d\psi^{(A)}(\exp(x))}{dx} = e^x \frac{d\psi^{(A)}(y)}{dy} = y\frac{d\psi^{(A)}(y)}{dy}$. Hence the fact that $\psi^{(A)}(\exp(x))$ and $\psi^{(B)}(\exp(x))$ are convex means that $y\frac{d\psi^{(A)}(y)}{dy}$ and $y\frac{d\psi^{(B)}(y)}{dy}$ are non-decreasing functions of y when $y > 0$. Thus for $0 < y_1 < y_2$,

$$\begin{aligned} y_1 \frac{d\psi^*(y_1)}{dy} &= \max\left(y_1 \frac{1}{\gamma^{(A)}} \frac{d\psi^{(A)}(y_1)}{dy},\ y_1 \frac{1}{\gamma^{(B)}} \frac{d\psi^{(B)}(y_1)}{dy} \right) \\ &\le \max\left(y_2 \frac{1}{\gamma^{(A)}} \frac{d\psi^{(A)}(y_2)}{dy},\ y_2 \frac{1}{\gamma^{(B)}} \frac{d\psi^{(B)}(y_2)}{dy} \right) \\ &= y_2 \frac{d\psi^*(y_2)}{dy} \end{aligned}$$

hence $y\frac{d\psi^*(y)}{dy}$ is also non-decreasing when $y > 0$ and it follows that $\psi^*(\exp(x))$ is convex.

To prove containment of uncertainty intervals, we first show that for $x_1 < x_2$, both $\frac{\psi^*(x_2)-\psi^*(x_1)}{\gamma^*} \ge \frac{\psi^{(A)}(x_2)-\psi^{(A)}(x_1)}{\gamma^{(A)}}$ and $\frac{\psi^*(x_2)-\psi^*(x_1)}{\gamma^*} \ge \frac{\psi^{(B)}(x_2)-\psi^{(B)}(x_1)}{\gamma^{(B)}}$. To see why, note that

$$\begin{aligned} \frac{\psi^*(x_2) - \psi^*(x_1)}{\gamma^*} &= \psi^*(x_2) - \psi^*(x_1) \\ &= \int_{x_1}^{x_2} \frac{d\psi^*(x)}{dx}\, dx \\ &\ge \frac{1}{\gamma^{(A)}} \int_{x_1}^{x_2} \frac{d\psi^{(A)}(x)}{dx} \\ &= \frac{\psi^{(A)}(x_2) - \psi^{(A)}(x_1)}{\gamma^{(A)}} \end{aligned}$$

and same with $\psi^{(A)}$ replaced by $\psi^{(B)}$.

Now pick a point x and let $z^{(+A)} = (\psi^{(A)})^{-1}\left(\psi^{(A)}(x) + \gamma^{(A)}\right)$ and also let $z^{(+*)} = (\psi^*)^{-1}(\psi^*(x) + \gamma^*)$. Then:

$$\begin{aligned} 1 &= \frac{\psi^{(A)}(x) + \gamma^{(A)} - \psi^{(A)}(x)}{\gamma^{(A)}} \\ &= \frac{\psi^{(A)}(z^{(+A)}) - \psi^{(A)}(x)}{\gamma^{(A)}} \\ &\le \frac{\psi^*(z^{(+A)}) - \psi^*(x)}{\gamma^*} \\ &= \psi^*(z^{(+A)}) - \psi^*(x) \end{aligned}$$

but, $1 = \psi^*(z^{(+*)}) - \psi^*(x)$ and since ψ^* is increasing, this means $z^{(+*)} \le z^{(+A)}$. Hence the right endpoint of the uncertainty interval determined by the combination $\psi^{(A)}, \gamma^{(A)}$ is greater than or equal to the right endpoint of the uncertainty interval determined by the combination ψ^*, γ^*.

The left endpoint is trickier because it has a more complex expression. Pick a point x. Define $\gamma_L^{(A)} = \min(\gamma^{(A)},\quad \psi^{(A)}(x) - \psi^{(A)}(0))$. This means the uncertainty interval's left endpoint is

$$\begin{aligned} z^{(-A)} &= (\psi^{(A)})^{-1}\left(\max(\psi^{(A)}(0),\quad \psi^{(A)}(x) - \gamma^{(A)})\right) \\ &= (\psi^{(A)})^{-1}\left(\psi^{(A)}(x) - \gamma_L^{(A)}\right) \end{aligned}$$

Similarly, define

$$\begin{aligned} \gamma_L^* &= \min(\gamma^*,\quad \psi^*(x) - \psi^*(0)) \\ z^{(-*)} &= (\psi^*)^{-1}\left(\max(\psi^*(0),\quad \psi^*(x) - \gamma^*)\right) \\ &= (\psi^*)^{-1}\left(\psi^*(x) - \gamma_L^*\right) \end{aligned}$$

Now, if $z^{(-A)} = 0$ then the left endpoint of the uncertainty interval associated with $\psi^{(A)}$ and $\gamma^{(A)}$ is the smallest possible and hence $\le$ the left endpoint associated with ψ^* and γ^*. Hence, **for the rest of the proof, we just need to consider the case where** $\mathbf{z^{(-A)} = (\psi^{(A)})^{-1}(\psi^{(A)}(x) - \gamma^{(A)})}$.

Then we have:

$$\begin{aligned}
1 &= \frac{\psi^{(A)}(x) - (\psi^{(A)}(x) - \gamma^{(A)})}{\gamma^{(A)}} \\
&= \frac{\psi^{(A)}(x) - \psi^{(A)}(z^{(-A)})}{\gamma^{(A)}} \\
&= \frac{1}{\gamma^{(A)}} \int_{r=z^{(-A)}}^{x} \frac{d\psi^{(A)}(r)}{dr}\, dr \\
&\leq \int_{r=z^{(-A)}}^{x} \frac{d\psi^{*}(r)}{dr}\, dr \quad \text{(by construction of } \psi^*) \\
&= \left(\psi^*(x) - \psi^*(z^{(-A)})\right) \\
&\leq \frac{\gamma^*}{\gamma_L^*} \left(\psi^*(x) - \psi^*(z^{(-A)})\right) \quad \text{(since } \gamma^* \geq \gamma_L^*) \\
&= \frac{1}{\gamma_L^*} \left(\psi^*(x) - \psi^*(z^{(-A)})\right) \quad \text{(since } \gamma^* = 1)
\end{aligned}$$

But, we also have

$$\begin{aligned}
1 &= \frac{\psi^*(x) - (\psi^*(x) - \gamma_L^*)}{\gamma_L^*} \\
&= \frac{\psi^*(x) - \psi^*(z^{(-*)})}{\gamma_L^*}
\end{aligned}$$

So this means $\psi^*(x) - \psi^*(z^{(-*)}) \leq \psi^*(x) - \psi^*(z^{(-A)})$ and so $\psi^*(z^{(-A)}) \leq \psi^*(z^{(-*)})$. Since ψ^* is increasing, this means $z^{(-A)} \leq z^{(-*)}$ and so the left endpoint associated with $\psi^{(A)}$ and $\gamma^{(A)}$ is $\leq$ the left endpoint associated with ψ^* and γ^*. Hence, we have $\mathcal{I}_{\psi^{(A)},\gamma^{(A)}}(x) \supseteq \mathcal{I}_{\psi^*,\gamma^*}(x)$.

A similar argument holds for $\psi^{(B)}$. □

E Proofs from Section 6

Several results in Section 6 rely on existing properties of Gaussian Differential Privacy [15] and adapt definitions associated with the GDP framework. The ψ-neighbor sensitivity (Definition 6.1) adapts L_2-sensitivity (Definition E.1) to our proposed framework.

Definition E.1 (L_2 sensitivity)**.** The L_2 sensitivity of a vector-valued function q is defined as $\Delta_2(q) \equiv \sup_{\mathcal{D}_1 \sim \mathcal{D}_2} ||q(\mathcal{D}_1) - q(\mathcal{D}_2)||_2$, where the supremum is taken over all pairs of datasets that differ on the value of one record.

The ψ-neighbor sensitivity restricts the pairs of dataset to any which are ψ-neighbors (Definition 5.6) where as L_2-sensitivity considers all pairs of datasets differing on one record. Then the Estab-Gaussian mechanism in Definition 6.2 adapts the Gaussian Mechanism (Definition E.2) using the ψ-neighbor sensitivity (Definition 6.1).

Definition E.2 (Gaussian Mechanism [15])**.** Let q be a query represented by a vector-valued function. Given a privacy parameter μ and input dataset $\mathcal{D}$, the Gaussian Mechanism returns $q(\mathcal{D}) + N\left(0, \frac{{\Delta_2}^2(q)}{\mu^2}\mathbf{I}\right)$ and satisfies μ-GDP.

Theorem 6.3. Given a privacy budget $\mu > 0$, neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$, the Estab-Gaussian mechanism (Definition 6.2) satisfies μ-gedp.

Proof of Theorem 6.3. These results follow directly from the proofs of the Gaussian mechanism of Gaussian Differential Privacy. □

Theorem 6.5. Let ψ be a neighbor function satisfying Definition 5.2. Let $\gamma_1, \ldots, \gamma_{k_c}$ be the associated distance parameters. Then the ψ-mechanism $\mathcal{M}_\psi$ satisfies μ-gedp with neighbor function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$.

Proof of Theorem 6.5. First, note that each group in a group-by sum query $gsum_{\langle \mathcal{G}, C_i \rangle}$ is formed from the public attributes. Hence modifying one establishment only changes one of the groups. Let x be the value of C_i of an establishment in such a group before modification, let y be the value of C_i for that establishment after modification, and let t be the sum of the C_i values in the rest of the establishments in that group. By definition of ψ-closeness, $|\psi(x) - \psi(y)| \leq \gamma_i$. Theorem 5.4, item 2, guarantees that the ψ-neighbor sensitivity (Definition 6.1) of $\psi \circ gsum_{\langle \mathcal{G}, C_i \rangle}$ is $\leq \gamma_i$ since $|\psi(x + t) - \psi(y + t)| \leq |\psi(x) - \psi(y)|$. Hence, by Theorem 6.3, $\mathcal{M}_\psi$ satisfies μ-gedp. □

Theorem 6.7. Given a group-by sum query $gsum_{\langle \mathcal{G}, C_i \rangle}$, let x be the true sum within a given group g, and let $\omega_g = \psi(x) + N(0, \gamma_i^2/\mu^2)$ be the noisy answer provided by $\mathcal{M}_\psi$. Then

- If $\psi = \sqrt{\ }$, then ω_g^2 has the distribution of a non-central χ^2 random variable with 1 degree of freedom and non-centrality parameter $x\mu^2/\gamma^2$, multiplied by γ_i^2/μ^2. Then $\widetilde{x} \equiv \omega_g^2 - (\gamma_i^2/\mu^2)$ is an unbiased estimate of x. The variance of $\widetilde{x}$ is $v \equiv 2(\gamma_i/\mu)^2\left(2x + (\gamma_i/\mu)^2\right)$. Replacing x with $\widetilde{x}$ in this formula results in a releasable estimate $\widetilde{v}$ of the variance, and $\widetilde{v}/v \to 1$ in probability as $x \to \infty$.
- If $\psi = \log$, then e^{ω_g} has the log-normal distribution with location parameter $\log(x)$ and scale γ_i^2/μ^2. Then $\widetilde{x} \equiv \exp(\omega_g - \gamma_i^2/(2\mu^2))$ is an unbiased estimate of x. The variance of $\widetilde{x}$ is $v \equiv x^2(-1 + \exp(\gamma_i^2/\mu^2))$. Replacing x with $\widetilde{x}$ in this formula gives a releasable estimate $\widetilde{v}$, but $\widetilde{v}/v$ does not converge in probability as $x \to \infty$.

Proof of Theorem 6.7. Let x be the true sum within a given group g and $\widetilde{x} \equiv \omega_g^2 - (\gamma_i^2/\mu^2)$ where $\omega_g = \sqrt{x} + N(0, \gamma_i^2/\mu^2)$.

Then ω_g^2 is equal in distribution to $\left(\frac{\gamma_i}{\mu}\right)^2 \chi_1^2(\eta)$ where $\eta = \left(\frac{\sqrt{x_i}\mu}{\gamma_i}\right)^2$ and $\chi_1^2(\eta)$ is a non-central chi-squared random variable with 1 degree of freedom and non-centrality parameter η (since the distribution of $\chi_1^2(\eta)$ is, by definition, the distribution of the square of a $N(\sqrt{\eta}, 1)$ random variable). Since $\mathrm{E}(\chi_1^2(\eta)) = 1 + \eta$ and $\mathrm{Var}(\chi_1^2(\eta)) = 2(1 + 2\eta)$ [28], then $\widetilde{x}$ is an unbiased estimator for x. Additionally, $v = \mathrm{Var}\left(\widetilde{x}\right) = 2\left(\frac{\gamma_i}{\mu}\right)^2\left(2x + \left(\frac{\gamma_i}{\mu}\right)^2\right)$. A plug-in estimate for the variance is $\widetilde{v} = 2\left(\frac{\gamma_i}{\mu}\right)^2\left(2\widetilde{x} + \left(\frac{\gamma_i}{\mu}\right)^2\right)$. Thus,

$$\frac{\widetilde{v}}{v} - 1 = \frac{\widetilde{v} - v}{v} = \frac{2\frac{\gamma_i^2}{\mu^2}\left(\widetilde{x} - x\right)}{\frac{\gamma_i^2}{\mu^2}\left(2x + (\gamma_i/\mu)^2\right)}$$

To show that $\widetilde{v}/v \to 1$ in probability as $x \to \infty$, we must show $\lim_{x\to\infty} P\left(\left|\frac{2(\widetilde{x}-x)}{2x+(\gamma_i/\mu)^2}\right| > \epsilon_{lim}\right) = 0$ for any $\epsilon_{lim} > 0$.

We can express this probability in terms of x and a standard normal random variable Z. For ease of notation let $a_i^2 = \frac{\gamma_i^2}{\mu^2}$. Since $x > -0$, the resulting probability

$$P\left(\left|\frac{2}{2x + a_i^2}\left(a_i^2\left(Z + \sqrt{x}/a_i\right)^2 - x - a_i^2\right)\right| > \epsilon_{lim}\right)$$

can be expanded to be $1 - p_+(x) + p_-(x)$ where

$$p_+(x) = P\left(\left(Z + \sqrt{x}/a_i\right)^2 \le \frac{2x(1 + \epsilon_{lim}) + (2 + \epsilon_{lim})a_i^2}{2a_i^2}\right)$$

and

$$p_-(x) = P\left(\left(Z + \sqrt{x}/a_i\right)^2 < \frac{2x(1 - \epsilon_{lim}) + (2 - \epsilon_{lim})a_i^2}{2a_i^2}\right)$$

. Then $p_+(x)$ can be expressed with the standard normal cumulative distribution function Φ:

$$p_+(x) = \Phi\left(\frac{\sqrt{2(1 + \epsilon_{lim})x + (2 + \epsilon_{lim})a_i^2} - \sqrt{2x}}{a_i\sqrt{2}}\right) - \Phi\left(\frac{-\sqrt{2(1 + \epsilon_{lim})x + (2 + \epsilon_{lim})a_i^2} - \sqrt{2x}}{a_i\sqrt{2}}\right).$$

Since $\epsilon_{lim} > 0$, we know $\lim_{x\to\infty} p_+(x) = 1$. Regarding $p_-(x)$, if $\epsilon_{lim} > 1$, then $\lim_{x\to\infty} \frac{(1-\epsilon_{lim})}{a_i^2}x + 1 - \frac{\epsilon_{lim}}{2} < 0$ and $\lim_{x\to\infty} p_-(x) = 0$. If $0 < \epsilon_{lim} \le 1$, then

$$p_-(x) = \Phi\left(\frac{\sqrt{2(1 - \epsilon_{lim})x + (2 - \epsilon_{lim})a_i^2} - \sqrt{2x}}{\sqrt{2a_i^2}}\right) - \Phi\left(\frac{-\sqrt{2(1 - \epsilon_{lim})x + (2 - \epsilon_{lim})a_i^2} - \sqrt{2x}}{\sqrt{2a_i^2}}\right).$$

We know $\pm\sqrt{2(1 - \epsilon_{lim})x + (2 - \epsilon_{lim})a_i^2} - \sqrt{2x} \to -\infty$ as $x \to \infty$ for $0 < \epsilon_{lim} \le 1$. Thus $\lim_{x\to\infty} p_-(x) = 0$ for any $\epsilon_{lim} > 0$. Therefore, $\widetilde{v}/v$ converges in probability to 1 as $x \to \infty$.

For $\psi = \log$ with x restricted to positive values, $\omega_g = \log(x) + N(0, \gamma_i^2/\mu^2)$. Then e^{ω_g} is a log-normal random variable with location parameter $\eta = \log(x)$ and scale parameter $\sigma^2 = \frac{\gamma_i^2}{\mu^2}$. Using known formulas for expectation and variance of log-normal random variables [9], we know $\widetilde{x} \equiv \exp\left(\omega_g - \gamma_i^2/(2\mu^2)\right)$ is unbiased for x. The variance of $\widetilde{x}$ is: $v = \mathrm{Var}\left(\exp\left(\omega_g - \gamma_i^2/(2\mu^2)\right)\right) = x^2\left(e^{\gamma_i^2/\mu^2} - 1\right)$. The

plug-in estimate is $\widetilde{v} = \widetilde{x}^2\left(e^{\gamma_i^2/\mu^2} - 1\right)$. We consider $\widetilde{v}/v = \widetilde{x}^2/x^2$. Since $\widetilde{x} \equiv x\exp\left(\frac{\gamma_i}{\mu}Z - \frac{\gamma_i^2}{2\mu^2}\right)$ where $Z \sim N(0,1)$, $P\left(\left|\frac{\widetilde{v}}{v} - 1\right| > \epsilon_{lim}\right) = 1 - P\left(\exp\left(\frac{2\gamma_i}{\mu}Z - \frac{\gamma_i^2}{\mu^2}\right) \le 1 + \epsilon_{lim}\right) + P\left(\exp\left(\frac{2\gamma_i}{\mu}Z - \frac{\gamma_i^2}{\mu^2}\right) < 1 - \epsilon_{lim}\right)$. This probability does not depend on the value of x and a selection of any $\epsilon_{lim} > 1$ makes $P\left(\left|\frac{\widetilde{v}}{v} - 1\right| \ge \epsilon_{lim}\right) > 0$. Thus $\widetilde{v}/v$ does not converge in probability.

□

Figure 5 compares the two estimators for group-by sum queries given in Theorem 6.7 for the choices of $\psi = \sqrt{\ }$ and $\psi = \log$ using the same settings examples as in Section 5.4. As with the privacy settings (where log under-protects small establishments and overprotects large ones), the choice of $\psi = \log$ has undesirable utility properties: it has larger errors for large aggregates. But the large aggregates (such as total employment in the United States) are so politically sensitive that they require extremely low relative errors in practice. This is yet another reason to prefer $\psi = \sqrt{\ }$.

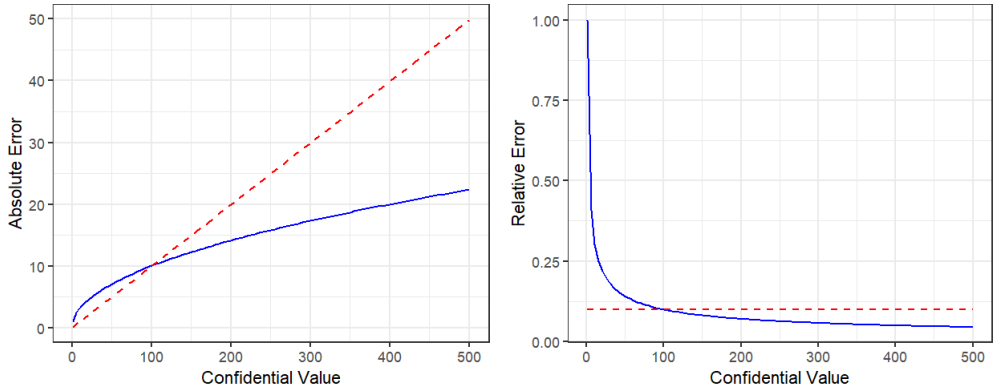

Figure 5: Comparison of the absolute and relative error of unbiased privacy preserving estimators of group-by sum queries, as given by Theorem 6.7. The x axis is the true answer. Solid blue line: $\psi = \sqrt{\ }$ with $\gamma = 0.5$ and $\mu = 1$. Dashed red line: $\psi = \log$ with $\gamma = 0.1$.

Theorem 6.8. Given establishment records $r_1, \ldots, r_n$, a neighboring function ψ with distance parameters $\gamma_1, \ldots, \gamma_{k_c}$, and a privacy loss budget allocations $\mu_1, \ldots, \mu_{k_c}$, define the noisy values $\omega_{j,i}$ as $\omega_{j,i} = \psi(r_j[C_i]) + N(0, \gamma_i^2/\mu_i^2)$. Let $\zeta \in (0,1)$ be a confidence parameter. First, define the *probably-no-clipping parameter* $\widehat{\tau} = \Phi^{-1}((1-\zeta)^{1/(k_c n)})$. Next, define upper bounds $u_{j,i}$ as $u_{j,i} = \psi^{-1}(\omega_{j,i} + \gamma_i\widehat{\tau}/\mu_i)$. Then with probability $1 - \zeta$ the following inequalities simultaneously hold: $r_j[C_i] \le u_{j,i}$ for all j and i.

Proof of Theorem 6.8.

$$\begin{aligned}
&P\Big(r_j[C_i] \le u_{j,i} \text{ for } j = 1, \ldots, n \text{ and } i = 1, \ldots, k_c\Big)\\
&= P\Big(\psi(r_j[C_i]) \le \omega_{j,i} + \gamma_i\widehat{\tau}/\mu_i \text{ for } j = 1, \ldots, n \text{ and } i = 1, \ldots, k_c\Big)\\
&= P\Big(-N(0, \sigma_i^2/\mu_i^2) \le \sigma_i\widehat{\tau}/\mu_i \text{ for } j = 1, \ldots, n \text{ and } i = 1, \ldots, k_c\Big)\\
&= P(N(0,1) \le \widehat{\tau} \text{ for } nk_c \text{ independent Gaussians})\\
&= (\Phi(\tau))^{nk_c} = (\Phi(\Phi^{-1}((1-\zeta)^{\frac{1}{nk_c}}))^{nk_c} = 1 - \zeta
\end{aligned}$$

□

Theorem 6.10. The pnc-mechanism from Definition 6.9 satisfies μ-gedp with neighboring function ψ and distance parameters $\gamma_1, \ldots, \gamma_{k_c}$.

Proof of Theorem 6.10. Clearly, the privacy cost of different groups compose in parallel since they affect different establishments, and the choice of which establishment goes into which group is based on the public attributes. Thus we just need to analyze one group, compute the sensitivity of the statistic for that group, and then apply Theorem 6.3 for the Estab-Gaussian mechanism.

Given a group g_ℓ, the statistic being computed is $\sum_{\mathcal{E}_j \in g_\ell} \min(r[C_i], u_\ell^*)$. Clearly the sensitivity is the same as the solution to

$$\begin{aligned}
&\max_{x \ge 0, y \ge 0} |x - y|\\
&\text{s.t. } |\psi(x) - \psi(y)| \le \gamma_i
\end{aligned}$$

$$\text{and } x \le u_\ell^*$$
$$\text{and } y \le u_{\ell^*}$$

Without loss of generality, we can set $x > y$. Noting that ψ is a neighbor function and hence increasing, we need to make y as small as possible given x. Thus we can set y so that $\psi(y) = \psi(x) - \gamma_i$, meaning that $y = \psi^{-1}(\max(0, \psi(x) - \gamma_i))$. Thus the sensitivity equals:

$$\max_{x \in [0, u_\ell^*]} |x - \psi^{-1}(\max(0, \psi(x) - \gamma_i))|$$

Next, since ψ is concave and increasing, then ψ^{-1} is convex and increasing. This means for any $a \ge \gamma_i$, $\psi^{-1}(a) - \psi^{-1}(a - \gamma_i)$ is an increasing function of a. Setting $a = \psi(x)$ proves the result. □

F Approximation Errors in Estimating Query Answer Variances can Impact on Postprocessed Data Quality

It is important to analyze how approximation errors in $v_1, \ldots, v_k$ affect the quality of $\widehat{D}$. The pnc-mechanism and ψ-mechanism are too complex to analyze in closed form, so instead we next analyze simple mathematical models that are tractable. They show that if the v_i are approximations to the true variances, the error of queries computed from $\widehat{D}$ can increase. If the approximations are data-dependent, then this introduces bias as well. Numerical simulations in Appendix G and Section 8 support these findings and hence support the use of pnc-mechanism as much as possible when bias is a concern.

Case 1: known variances. Suppose there is just one establishment $\mathcal{E}$ with X employees and we have n independent unbiased noisy estimates $a_1, \ldots, a_n$ of the employment count (i.e., $E[a_i] = X$). Each a_i has known variance σ^2. Applying Equation 1, we get the intuitive result that the overall guess of the employment count X should be $(a_1 + \cdots + a_n)/n$ and the variance of this guess is σ^2/n.

Case 2: estimated, but data-independent, variances. Next, we examine how error increases when the true variances of the a_i (i.e., σ^2) are unknown, but noisy variance estimates $v_1, \ldots, v_n$ are available. Suppose the variance estimates v_i have independent inverse Gamma distributions.[7] If each v_i has distribution $\text{IG}(2 + \tau, \sigma^2(1 + \tau))$, then $E[v_i] = \sigma^2$, and $Var(v_i) = \sigma^4/\tau$ and τ controls how accurate the v_i are. Solving Equation 1, the guess for employment count X is $\frac{\sum_{i=1}^n a_i/v_i}{\sum_{j=1}^n 1/v_j}$. It is unbiased (the expected value equals X) and the squared error of the guess is:

$$E\left[\left(X - \frac{\sum_i a_i/v_i}{\sum_j 1/v_j}\right)^2\right] = \sum_i E\left[\frac{(X - a_i)^2/v_i^2}{(\sum_j 1/v_j)^2}\right] = \frac{\sigma^2}{n}\frac{n(3+\tau)}{1+n(2+\tau)},$$

where the last equality follows from the following facts: (1) If v_i has the $\text{IG}(\alpha, \beta)$ distribution, then $1/v_i$ has the $\text{Gamma}(\alpha, \beta)$ distribution. (2) If $1/v_1, \ldots, 1/v_n$ are independent samples from a $\text{Gamma}(\alpha, \beta)$ distribution then the vector $\frac{(1/v_1, \ldots, 1/v_n)}{\sum_j 1/v_j}$ has the $\text{Dirichlet}(\alpha, \ldots, \alpha)$ distribution and its i^{th} component has mean α/n, variance $(\alpha/n)(1 - \alpha/n)/(1 + \alpha/n)$, and second moment $\frac{1}{n}\frac{1+\alpha}{1+n\alpha}$. Comparing to Case 1, we see that **uncertainty about the variance of the a_i increases the squared error by a factor of** $\frac{n(3+\tau)}{1+n(2+\tau)}$. For example, with $n = 2$ and $\tau = 1$, this factor is ≈ 1.14, representing a 14% increased error simply for not knowing exactly how accurate are our noisy estimates of the employment at establishment $\mathcal{E}$.

Case 3: estimated and data-dependent variances. To make the variances of the a_i both uncertain and data-dependent (like with the ψ-mechanism), suppose $a_1, \ldots, a_n$ are samples from an Inverse Gamma $\text{IG}(2 + X/c, X + X^2/c)$ distribution, where $c > 0$ is some constant. Then $E[a_i] = X$ and $Var(a_i) = cX$. Thus the variance estimate is $v_i = a_i/c$. This relation between the mean and estimated variance is similar to when $\psi = \sqrt{\ }$ (see Theorem 6.7). Solving Equation 1, the guess for employment count X is $\frac{\sum_{i=1}^n a_i/v_i}{\sum_{j=1}^n 1/v_j} = \frac{n}{\sum_j 1/a_j}$. We use the following facts: (1) if v_i has distribution $IG(\alpha, \beta)$ then $1/v_i$ has the $\text{Gamma}(\alpha, \beta)$ distribution, and (2) the sum of n independent $\text{Gamma}(\alpha, \beta)$ random variables is a $\text{Gamma}(n\alpha, \beta)$ distribution. Then the expected value of the guess is $\frac{n(2+X/c)-n}{n(2+X/c)-1}$, which is slightly *less* than X; i.e., the guess has a downward bias.

The above analysis suggests that the pnc-mechanism (whose true variance is publicly releasable) is preferable to the ψ-mechanism (whose variance is both data dependent and must be approximated) when bias in the confidentiality-preserving microdata $\widehat{D}$ is a concern. Further numerical results can be found in Appendix G and Section 8.

Useful properties about distributions. These are the facts needed about the distributions used in the analytical analysis of inverse weighted least squares.

(1) The Inverse Gamma distribution $IG(\alpha, \beta)$ with shape parameter α and scale parameter β has mean $\frac{\beta}{\alpha - 1}$ (for $\alpha > 1$) and variance $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (when $\alpha > 2$). Note, the variance is the mean squared divided by $(\alpha - 2)$.
(2) The Gamma distribution $\text{Gamma}(\alpha, \beta)$ with shape parameter α and rate parameter β has mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$.
(3) If v follows the $\text{IG}(\alpha, \beta)$ distribution, then $1/v$ follows the $\text{Gamma}(\alpha, \beta)$ distribution with shape parameter α and rate parameter β.

[7] The inverse Gamma distribution $\text{IG}(\alpha, \beta)$ is a distribution over nonnegative numbers. It has mean $\beta/(\alpha - 1)$ and variance $\beta^2/(\ (\alpha - 1)^2(\alpha - 2)\ )$.

(4) If $1/v_1, \ldots, 1/v_n$ are independent and each $1/v_i$ follows the Gamma(α_i, β) distribution (with the same β but possibly different α_i) then $\sum_{i=1}^n 1/v_i$ follows the Gamma$(\sum_{i=1}^n \alpha_i, \beta)$ distribution.

(5) If $1/v_1, \ldots, 1/v_n$ are independent and each $1/v_i$ follows the Gamma(α_i, β) distribution then the joint distribution of the n quantities $\frac{1/v_1}{\sum_{j=1}^n 1/v_j}, \ldots, \frac{1/v_n}{\sum_{j=1}^n 1/v_j}$ is the Dirichlet$(\alpha_1, \ldots, \alpha_n)$ distribution.

(6) If the joint distribution of $w_1, \ldots, w_n$ is Dirichlet$(\alpha_1, \ldots, \alpha_n)$, then $E[w_i] = \frac{\alpha_i}{\sum_{j=1}^n \alpha_j}$ and $Var(w_i) = \frac{\frac{\alpha_i}{\sum_{j=1}^n \alpha_j}\left(1 - \frac{\alpha_i}{\sum_{j=1}^n \alpha_j}\right)}{1+\sum_{i=j}^n \alpha_j}$ and

$$\begin{aligned}
E[w_i^2] &= Var(w_i) + E[w_i]^2 \\
&= \frac{\alpha_i^2}{(\sum_{j=1}^n \alpha_j)^2} + \frac{\frac{\alpha_i}{\sum_{j=1}^n \alpha_j}\left(1 - \frac{\alpha_i}{\sum_{j=1}^n \alpha_j}\right)}{1 + \sum_{i=j}^n \alpha_j} \\
&= \frac{\alpha_i^2 + \alpha_i^2 \sum_{j=1}^n \alpha_j + \alpha_i(-\alpha_i + \sum_{j=1}^n \alpha_j)}{(\sum_{j=1}^n \alpha_j)^2(1 + \sum_{i=j}^n \alpha_j)} \\
&= \frac{\alpha_i^2 \sum_{j=1}^n \alpha_j + \alpha_i \sum_{j=1}^n \alpha_j}{(\sum_{j=1}^n \alpha_j)^2(1 + \sum_{i=j}^n \alpha_j)} \\
&= \frac{\alpha_i^2 + \alpha_i}{(\sum_{j=1}^n \alpha_j)(1 + \sum_{i=j}^n \alpha_j)} \\
&= \frac{\alpha_i(1 + \alpha_i)}{(\sum_{j=1}^n \alpha_j)(1 + \sum_{i=j}^n \alpha_j)}
\end{aligned}$$

Analysis of Case 2: Recall $a_i, \ldots, a_n$ are random variables where $E[a_i] = X$ and $Var(a_i) = \sigma^2$. Recall that each v_i has distribution IG$(2 + \tau, \sigma^2(1 + \tau))$. Let $\widehat{X} = \frac{\sum_{i=1}^n a_i/v_i}{\sum_{j=1}^n 1/v_j}$. Then

$$\begin{aligned}
E[\widehat{X}] &= E\left[\frac{\sum_{i=1}^n a_i/v_i}{\sum_{j=1}^n 1/v_j}\right] = \sum_{i=1}^n E\left[\frac{a_i/v_i}{\sum_{j=1}^n 1/v_j}\right] \\
&= \sum_{i=1}^n E\left[\frac{X/v_i}{\sum_{j=1}^n 1/v_j}\right] = XE\left[\frac{\sum_{i=1}^n 1/v_i}{\sum_{j=1}^n 1/v_j}\right] = X \\
E\left[\left(X - \frac{\sum_i a_i/v_i}{\sum_j 1/v_j}\right)^2\right] &= E\left[\left(\frac{\sum_i (X - a_i)/v_i}{\sum_j 1/v_j}\right)^2\right] \\
&= \sum_i E\left[\frac{(X - a_i)^2/v_i^2}{(\sum_j 1/v_j)^2}\right] \\
&\quad \text{(because the cross terms have expected value} \\
&\quad \text{0 when expanding the square)} \\
&= n\sigma^2 E\left[\frac{1/v_1^2}{(\sum_j 1/v_j)^2}\right] \\
&\quad \text{(because the } v_i \text{ have the same distribution)} \\
&\quad (v_1 \text{ has IG}(2 + \tau, \sigma^2(1 + \tau)) \text{ distribution)} \\
&\quad (1/v_1 \text{ has Gamma}(2 + \tau, \sigma^2(1 + \tau)) \text{ distribution)} \\
&\quad \left(\frac{v_1}{\sum_j 1/v_j}, \ldots, \frac{v_n}{\sum_j 1/v_j} \text{ has Dirichlet}(2 + \tau, \ldots, 2 + \tau) \text{ distribution}\right) \\
&= n\sigma^2 \frac{2 + \tau}{n(2 + \tau)} \frac{1 + 2 + \tau}{1 + n(2 + \tau)} \\
&= \sigma^2 \frac{1 + 2 + \tau}{1 + n(2 + \tau)} \\
&= \frac{\sigma^2}{n} \frac{n + n(2 + \tau)}{1 + n(2 + \tau)}
\end{aligned}$$

Analysis of Case 3:

Let $a_1, \ldots, a_n$ be samples from an Inverse Gamma $\text{IG}(2 + X/c, X + X^2/c)$ distribution, where $c > 0$ is some constant. Then

$$\begin{aligned}
E\left[\frac{n}{\sum_j 1/a_j}\right] &= nE[1/\text{Gamma}(n(2 + X/c), X + X^2/c)] \\
&= nE\left[\text{IG}(n(2 + X/c), X + X^2/c)\right] \\
&= n\frac{X + X^2/c}{n(2 + X/c) - 1} \\
&= X\frac{n(2 + X/c) - n}{n(2 + X/c) - 1} \equiv \theta \\
E\left[\left(X - \frac{n}{\sum_j 1/a_j}\right)^2\right] &= E\left[\left(X - \theta + \theta - \frac{n}{\sum_j 1/a_j}\right)^2\right] \\
&= (X - \theta)^2 + E\left[\left(\theta - \frac{n}{\sum_j 1/a_j}\right)^2\right] \\
&\text{(cross terms have 0 expected value)} \\
&= (X - \theta)^2 + Var(\text{IG}(n(2 + X/c), X + X^2/c))
\end{aligned}$$

G An Empirical Exploration of Bias Caused by Estimated Variances

For a simple, but illuminating scenario, consider a dataset with 100 counties and 2 establishments per county. To simplify interpretation of the outcome, set the ground truth to be 10 employees in each establishment. The queries of interest are the **ID**: the identity query (how many employees are in each establishment), **County**: the number of employees in each county, and **Total:** the total employment. In the first setting, we add noise to each query using ψ-mechanism with $\psi = \sqrt{}$ ($\gamma = 0.5$ and $\mu = 1$); in the second setting, $\psi = \log$ ($\gamma = 0.1$ and $\mu = 1$). In both cases, Theorem 6.7 is used to recover unbiased query answers. We consider 3 postprocessing strategies:

- **Est**: Postprocessing with Equation 1 using estimated variances.
- **Act**: (Non-privacy-preserving) ablation experiment using the true variances in Equation 1.
- **Hybrid**: (Non-privacy-preserving) ablation experiment that uses Equation1 with estimated variances for the **ID** query the true variances for the rest. By comparing to **Act**, we will see the effect of having exact variances for non-identity queries. The full recommended setup, using ψ-mechanism (with estimated variances) for the identity query and pnc-mechanism (whose exact variances are safe to release) will be studied in Section 8.

Table 5 shows that the impact of uncertain variances is significant – the total employment query, which often is the most important, has the most significant accuracy degradation (**Est** vs. **Act** columns). The effects are ameliorated when the true variance is unknown only for the identity query (**Hybrid** column). Hence, one should avoid mechanisms with non-releasable variance whenever possible.

	Ablation Setting, Reporting Mean Squared Errors					
	$\psi = \sqrt{}$			$\psi = \log$		
Query	*Est*	*Act*	*Hybrid*	*Est*	*Act*	*Hybrid*
ID	7.5	7.6	7.5	18.0	23.7	19.2
County	10.5	10.0	10.1	40.0	37.9	35.0
Total	624.4	335.2	407.4	22, 215.3	1, 882.8	4, 842.6

Table 5: Postprocessing ablation study for evaluating the impact of using estimated variances with Equation 1. When $\psi = \sqrt{}$, then $\gamma = 0.5, \mu = 1$. When $\psi = \log$, then $\gamma = 0.1, \mu = 1$. Metric: mean-squared errors averaged over $1,000,000$ trials.

H Generation of Synthetic Evaluation data

We aim to create a synthetic establishment-level dataset that mirrors the structure of the QCEW data from publicly accessible data sources in order to run experiments to access our proposed privacy mechanisms and postprocessing procedures. The QCEW structure includes 18 variables, as described in Table 6.

We use three main sources of publicly accessible data for this task: the Quarterly Workforce Indicators (QWI) [8], the County Business Patterns (CBP) dataset [7], and an employment count imputed CBP created by Eckert et al.[18, 19]. The variables of these datasets are

summarized in Table 7. We focus on privately-owned establishment data from the 50 states for quarter 1 of 2016 since this matches the most recent imputed CBP dataset [18].

There are four main steps to create the synthetic data. First, we combine the data sources by county and industry code. Next, we extract and impute monthly employment counts and quarterly wages for each county and NAICS-4 combination. Then, we get month 1 and 3 employment counts, quarterly wages, and number of establishments for each NAICS-6 by county combination. Finally, we generate establishment-level data that matches the QCEW structure.

Variable	Description
year	year of data; all entries = 2016
qtr	quarter of data; all entries = 1
state	SOC numeric state/territory value from 1 to 56
cnty	numeric county value, numbered within each state
naics	6-digit industry code using North American Industry Classification System (NAICS)
naics5, naics4, naics3	5, 4, and 3 digit industry code respectively
sector	2-digit NAICS sector code
supersector	BLS sector code groups NAICS sectors
own	ownership code (our data is all code-5 = “privately owned")
m1emp, m2emp, m3emp	monthly employment counts
wage	total quarterly wages
primary_key	unique identifier for each establishment
can_agg	(metadata variable) all entries ="Y"
rectype	(metadata variable) all entries ="C"

Table 6: Variables for the establishment-level synthetic microdata.

H.1 Public Access Data

The U.S. Census offers many publicly accessible datasets that contain information similar to the QCEW data, namely the Quarterly Workforce Indicators (QWI) [8] and the County Business Patterns (CBP) [7]. The CBP dataset has values aggregated at the County by NAICS-6 level up to the County by Sector level, while county by NAICS-4 is the most precise aggregation level of the QWI dataset. Additionally, both datasets have many missing and suppressed values (Table 8). The CBP dataset includes values for number of establishments (`estnum`), a mid-March (month 3 of quarter 1) employment count (`emp`) and quarterly total payroll values (`qp1`) (Table 7). The CBP dataset is not ideal for creating the monthly employment counts for the synthetic QCEW data, since it only has one of the three monthly employment counts that the QCEW data has. The QWI dataset has several employment variables discussed in Table 7: the employment at the beginning and end of the quarter (`Emp` and `EmpEnd` respectively) and the number of workers stably employed through the whole quarter (`EmpS`). The CBP data can help fill suppressed QWI month 3 employment values, especially since Eckert et al. have imputed the 2016 suppressed CBP employment values up to the county by NAICS-6 level and released the imputed dataset [18]. The quarterly payroll value of the CBP data is a good base for the total quarterly data. However, 50% of `qp1` are suppressed at the county by NAICS-4 level and the imputed CBP dataset does not include wages (Table 4. The QWI dataset has the average monthly earnings of workers employed at the beginning of the quarter (`EarnBeg`) and only 2% of the county by NAICS-4 cells are suppressed. Using `EarnBeg` and the monthly employment counts we can fill in suppressed values at the county by NAICS-4 level. For denoting a variable from here on, we refer to it by its data source, public attribute grouping, and then variable name. For example, `QWI.cnty.EmpEnd` is the employment at the end of the quarter at the county level from the QWI dataset.

H.2 Synthetic County by NAICS-4 Data

Since the the NAICS-4 by county level is the most specific level of public attribute groups in the QWI data, we start by creating a dataset of monthly employment counts and quarterly wages at this level. We describe the process in detail and in Algorithms 1 to 4. Algorithm 1 gets the monthly employment values and uses Algorithm 2 to get the month 2 employment count. Algorithm 2 is used in the final stage of the data generation as well to get the establishment level month 2 employment counts. Algorithm 3 is for the wage values and uses Algorithm 4 to bound the wages above using hierarchical industry values.

We use a combination of `QWI.cntyN4.Emp`, `QWI.cntyN4.EmpEnd`, and `imputeCBP.cntyN4.emp` to generate the monthly employment counts for the county by NAICS-4 code cells (Algorithm 1). We denote these monthly employment counts as `cntyN4.m1emp`, `cntyN4.m2emp`, and `cntyN4.m3emp` to avoid confusion with the monthly employment counts at the establishment level or at the county by NAICS-6 level.

When `QWI.cntyn4.Emp` values are not suppressed, we set `cntyn4.m1emp` to be equal to them. However, when they are suppressed we use a regression model to impute the values.

First, we fit the model on the rows where `QWI.cntyN4.Emp` and `QWI.cntyN4.EmpEnd` are available. The following linear model is used:

$$\texttt{QWI.cntyN4.Emp}_{c,k} = \beta_0 + \beta_1 \texttt{QWI.cntyN4.EmpEnd}_{c,k} + \beta_2 \texttt{CBP.cntyN4.estnum}_{c,k} + \beta_3 \texttt{cntyN4.state}_{c,k} + \beta_4 \texttt{cntyN4.sector}_{c,k} + \epsilon_{c,k} \quad (9)$$

All Datasets (never suppressed)	
Public Attributes	state, county, sector, NAICS3, NAICS4
County Business Patterns (CBP) [7]	
Public Attribute Levels	County by Sector (cntySect), County by NAICS-3 (cntyN3) to County by NAICS-6 (cntyN6)
Variables	estnum: the number of establishments qp1: quarterly total payroll values
Imputed CBP (imputeCBP) [18]	
Public Attribute Levels	County by Sector (cntySect), County by NAICS-3 (cntyN3) to County by NAICS-6 (cntyN6)
Variable	emp: the imputed month 3 employment count
Quarterly Workforce Indicators (QWI) [8]	
Public Attribute Levels	County (cnty), County by NAICS4 (cntyN4)
Variables	Emp: Beginning of quarter employment EmpEnd: End of quarter employment EmpS: number of employees who work the whole quarter EarnBeg: average month 1 earnings of workers

Table 7: The three public access datasets we use have a variety of variables describing the number of establishments, the number of employees, and the quarterly wages aggregated at various county and industry levels. Within the text, we refer to a variable by its data source, public attribute grouping, and then variable name. For example, `CBP.cntyN4.qp1` is the quarterly total payroll aggregated by county and NAICS4 from the CBP dataset.

	Geographic Level				
	cntySect	cntyN3	cntyN4	cntyN5	cntyN6
CBP.estnum	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
CBP.qp1	10,813 (18%)	67,190 (36%)	210,047 (50%)	393,634 (58%)	480,987 (61%)
imputeCBP.emp	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

	Geographic Level	
	cnty	cntyN4
QWI.Emp	1 (0%)	187,814 (43%)
QWI.EmpEnd	1 (0%)	189,021 (43%)
QWI.EmpS	1 (0%)	184,082 (42%)
QWI.EarnBeg	0 (0%)	8,100 (2%)

Table 8: The number suppress values at each public attribute group level for each variable. The percent of rows at a given level that are suppressed in shown in parentheses.

where $\epsilon_{c,k} \overset{iid}{\sim} N(0, \sigma^2)$ for all (c, k) such that $\texttt{QWI.cntyN4.Emp}_{c,k}$ and $\texttt{QWI.cntyN4.EmpEnd}_{c,k}$ are not missing for the combination of county code c and NAICS-4 code k. This model was selected to include both geographic and industry code features as well as the number of establishments in cell. For the geographic predictor variable, we cannot use county as a predictor since there are counties that suppress all QWI.cntyN4.Emp values. State is used as the geographic predictor instead. Sector is used for similar reasons. Once the model is fitted, we check that the residuals for normality and equal variance, and identify one outlier. The model is refitted without the outlier. If QWI.cntyN4.EmpEnd is suppressed, we use the impute.cntyN4.emp as the predictor. Then we simulate the missing cntyN4.m1emp values using the refitted model. The QWI.cntyN4.EmpEnd filled with the imputeCBP.cntyN4.emp values are used as month 3 employment counts.

Then we generate month 2 employment counts using Algorithm 2 and noise parameter $\eta = 1.69$. This algorithm sets cntyN4.m2emp to zero if month 1 and 3 employment counts are zero. Otherwise the change from the midpoint of month 1 and 3 is randomly generated from a normal distribution with mean 0 and variance proportional to the ratio of the absolute difference and the average of the monthly employment values. More specifically, the variance is $\sigma^2 = \eta \frac{|\texttt{cntyN4.m3emp} - \texttt{cntyN4.m1emp}|}{\texttt{cntyN4.m1emp} + \texttt{cntyN4.m3emp}}$. Thus there is more variation from the midpoint when the monthly employment counts are vastly different. When the change between month 1 and month 3 is large relative to the monthly employment counts, the variation from the midpoint will be larger than if the same change is small relative to the monthly employment counts. If industries with a large number of employees are likely to have more steady changes in employment than small but fast-growing industries, then this algorithm mimics that behavior. The Month 2 Employment Algorithm is also used later when generating the establishment level month 2 employment counts.

Now that all three monthly employment counts have been determined, we check and adjust these values. First, if a stable employment count QWI.cntyN4.EmpS is available, we force this value to be a lower bound on each monthly employment count. We then use in county total QWI.cnty.Emp values to adjust any model-simulated m1emp values. Denote the value for county c as $\texttt{QWI.cnty.Emp}_c$. We are unable to adjust the values in Kalawao County, Hawaii (FIPS code 15005), since the county total is suppressed. For each other county, we sum the QWI.cntyN4.Emp values in the county and sum the m1emp values that came from the fitted model for that county. For county c, denote the sum of the QWI.cntyN4.Emp data as $\Sigma_{c,QWI}$Emp and the sum from the fitted model as $\Sigma_{c,model}$m1emp. If $\Sigma_{c,model}\texttt{m1emp} \neq 0$, then each

Algorithm 1: *(Employment for County by NAICS-4)* We use QWI.cntyN4.Emp and QWI.cntyN4.EmpEnd as m1emp and m3emp when available. Otherwise, we use impute.CBP.cntyN4.emp as m3emp and predict m1emp from a model fitted on a combination of QWI.cntyN4 and CBP.cntyN4 variables. Then m2emp is derived from Algorithm 2. Adjustments to the employment counts are made based on QWI.cnty.Emp and QWI.cntyN4.EmpS when available. For an set of indices such as $\mathcal{I}_*$ let the complement be denoted $\mathcal{I}_*^C$. Let the sum of values over an empty set be equal to 0.

Input: QWI.cntyN4.Emp, QWI.cntyN4.EmpEnd, QWI.cntyN4.EmpS; /* county by NAICS-4 employment values from QWI */
QWI.cnty.Emp ; /* county-level employment values from QWI */
imputeCBP.cntyN4.emp, CBP.cntyN4.estnum ; /* county by NAICS-4 values from CBP and imputed CBP */
$\mathcal{I} = \{(c,k) : (c,k) \in$ QWI.cntyN4 $\cap$ imputeCBP.cntyN4$\}$, ; /* county by NAICS-4 indices in QWI and imputed CBP */
cntyN4.state, cntyN4.sector, ; /* state and sector corresponding to county by NAICS-4 cells */
$\eta > 0$; /* noise parameter for m2emp */
/* Defining sets of indices based on missing values */
Let $\mathcal{I}_{/\text{Emp}} = \{(c,k) \in \mathcal{I} :$ QWI.cntyN4.Emp$_{c,k}$ is missing$\}$;
Let $\mathcal{I}_{/\text{EmpEnd}} = \{(c,k) \in \mathcal{I} :$ QWI.cntyN4.EmpEnd$_{c,k}$ is missing$\}$;
Let $\mathcal{I}_* = \{(c,k) \in \mathcal{I} :$ QWI.cntyN4.Emp$_{c,k}$ and QWI.cntyN4.EmpEnd$_{c,k}$ are not missing$\}$;
/* Fit a model to predict QWI.cntyN4.Emp from QWI.cntyN4.EmpEnd and other predictors. */
Fit model (9) with $(c,k) \in \mathcal{I}_*$;
If necessary, identify outliers based on residual normality plot and fitted verses residual plot. Remove them and refit model (9);
Let $\hat{f}_{\text{cntyN4.Emp}}(\cdot)$ be the function that takes in the predictor variable values and returns the fitted value of the model;
Let $se_{pred}(\hat{y})$ be the function that takes in a fitted value, $\hat{y}$, from the model and returns the standard error of the prediction;
foreach $(c,k) \in \mathcal{I}_*$ **do**
  m3emp$_{c,k}$ $\leftarrow$ QWI.cntyN4.EmpEnd$_{c,k}$; /* use QWI.cntyN4.Emp and QWI.cntyN4.EmpEnd is available */
  m1emp$_{c,k}$ $\leftarrow$ QWI.cntyN4.Emp$_{c,k}$
foreach $(c,k) \in \mathcal{I}_{/\text{EmpEnd}}$ **do**
  m3emp$_{c,k}$ $\leftarrow$ imputeCBP.cntyN4.emp$_{c,k}$; /* use imputeCBP.cntyN4.emp if QWI.cntyN4.EmpEnd is missing */
foreach $(c,k) \in \mathcal{I}_{/\text{Emp}}$ **do**
  $\hat{y}_{c,k} \leftarrow \hat{f}_{\text{cntyN4.Emp}}$(m3emp$_{c,k}$, CBP.cntyN4.estnum, QWI.cntyN4.state, QWI.cntyN4.sector);
  m1emp$_{c,k}$ $\leftarrow \hat{y}_{c,k} + Z_{c,k}$ where $Z_{c,k} \overset{indep.}{\sim} N(0, se^2_{pred}(\hat{y}_{c,k}))$; /* use refitted model if QWI.cntyN4.Emp is missing */
foreach $(c,k) \in \mathcal{I}$ **do**
  m2emp$_{c,k}$ $\leftarrow$ Month 2 Employment$(\eta$, m1emp$_{c,k}$, m3emp$_{c,k})$ where Month 2 Employment is Algorithm 2
  if QWI.cntyN4.EmpS$_{c,k}$ *is not missing* **then**
    mMemp$_{c,k}$ $= \max($mMemp$_{c,k}$, QWI.cntyN4.EmpS$_{c,k})$ for $M = 1, 2, 3$; /* QWI.cntyN4.EmpS lower bound if available */
  else
    mMemp$_{c,k}$ $= \max($mMemp$_{c,k}, 0)$ for $M = 1, 2, 3$; /* lower bound by 0 if QWI.cntyN4.EmpS is missing */
/* Adjust m1emp values from model using the county total value QWI.cnty.Emp if available. */
foreach c' *such that* $(c',k) \in \mathcal{I}_{/\text{Emp}}$ **do**
  if QWI.cnty.Emp$_{c'}$ *is not missing* **then**
    Let $\mathcal{I}_{c',model} = \{k : (c',k) \in \mathcal{I}_{/\text{Emp}}\}$;
    $\Sigma_{c',model}$m1emp $\leftarrow \sum_{k \in \mathcal{I}_{c',model}}$ m1emp$_{c',k}$;
    $\Sigma_{c',QWI}$QWI.cntyN4.Emp $\leftarrow \sum_{k \in \mathcal{I}_{c',QWI}}$ QWI.cntyN4.Emp$_{c',k}$ where $\mathcal{I}_{c',QWI} = \{k : (c',k) \in \mathcal{I}^C_{/\text{Emp}}\}$;
    if $\Sigma_{c',model}$m1emp $\neq 0$ **then**
      m1emp$_{c',k}$ $\leftarrow$ m1emp$_{c',k}$ + (QWI.cnty.Emp$_{c'} - \Sigma_{c',QWI}$QWI.cntyN4.Emp)$\frac{\text{m1emp}_{c',k}}{\Sigma_{c',model}\text{m1emp}}$ for each $k \in \mathcal{I}_{c',model}$
    else
      m1emp$_{c',k}$ $\leftarrow$ m1emp$_{c',k}$ + $\frac{1}{|\mathcal{I}_{c',model}\text{m1emp}|}$(QWI.cnty.Emp$_{c'} - \Sigma_{c',QWI}$QWI.cntyN4.Emp) for each $k \in \mathcal{I}_{c',model}$
Let cntyN4.m1emp be a vector of $\max(0,$ m1emp$_{c,k})$ for $(c,k) \in \mathcal{I}$;
Let cntyN4.m2emp be a vector of $\max(0,$ m2emp$_{c,k})$ for $(c,k) \in \mathcal{I}$;
Let cntyN4.m3emp be a vector of $\max(0,$ m3emp$_{c,k})$ for $(c,k) \in \mathcal{I}$;
Result: cntyN4.m1emp, cntyN4.m2emp, cntyN4.m3emp

model-fitted m1emp is adjusted proportional to its share of the model-fitted sum:

$$\texttt{cntyN4.m1emp}_{c,k} = \texttt{cntyN4.m1emp}_{c,k} + \left(\texttt{QWI.cnty.Emp}_c - \Sigma_{c,QWI}\texttt{QWI.cntyN4.Emp}\right) \frac{\texttt{cntyN4.m1emp}}{\Sigma_{c,model}\texttt{m1emp}}.$$

Otherwise m1emp is adjusted by $(\texttt{QWI.cnty.Emp}_c - \Sigma_{c,QWI}\texttt{QWI.cntyN4.Emp})$ over the number of cells that were simulated from the model.

Algorithm 2: *(Month 2 Employment)* Using a noise parameter η, we make a m2emp value from the m1emp and m3emp values. This algorithm is used within Algorithms 1 and 7

Input: $\eta > 0$, m1emp≥ 0, and m3emp≥ 0

if m1emp = m3emp = 0 **then**
  m2emp = 0
else
  $\sigma^2 = 2\eta \frac{|\texttt{m3emp}-\texttt{m1emp}|}{\texttt{m3emp}+\texttt{m1emp}}$;
  Generate $Z \sim \text{Normal}(0, \sigma^2)$;
  $\texttt{m2emp} = \texttt{m1emp} + \frac{\texttt{m3emp}-\texttt{m1emp}}{2} + Z$;
  if m2emp ≤ 0 **then**
    if m3emp = 0|m1emp = 0 **then**
      m2emp = 0
    else
      m2emp = 1

Result: m2emp

Finally, we can consider the wage values using Algorithm 3. If the CBP.cntyN4.qp1 value is not suppressed, it is used as the wage value. Otherwise, if the average monthly employee earnings at the beginning of the quarter QWI.cntyN4.EarnBeg is available,then we set wage = QWI.cntyN4.EarnBeg(cntyN4.m1emp + cntyN4.m2emp + cntyN4.m3emp) where the monthly employment counts are from Algorithm 1. For the remaining missing NAICS-4 by county cells, we use a regression model. Instead of predicting the wage value outright, we predict the difference between the county by NAICS-4 wage value and the sum of available county by NAICS-6 wage value within the NAICS-4 code.

First, we get the sum of the available NAICS-6 CBP.cntyN6.qp1 values and the count of suppressed NAICS-6 cells corresponding to the NAICS-4 by county cell. For NAICS-4 code k and county c, denote the sum of available NAICS-6 values as $\Sigma_{c,k}^{(6,4)}$, the count of suppressed cells as $M_{c,k}$. If there are no available county by NAICS-6 values for county c and NAICS-4 code k, then $\Sigma_{c,k}^{(6,4)} = 0$. For cells that have a CBP.cntyN4.qp1 value available, we fit a model:

$$\begin{aligned}\sqrt{\texttt{CBP.cntyN4.qp1}_{c,k} - \Sigma_{c,k}^{(6,4)}} = \beta_0 + \beta_1\texttt{cntyN4.state}_{c,k} + \beta_2\texttt{cntyN4.sector}_{c,k} + \beta_3 M_{c,k} + \beta_4\texttt{cntyN4.m3emp}_{c,k} \\ + \beta_5\texttt{cntyN4.m3emp}_{c,k}M_{c,k} + \beta_6\texttt{cntyN4.m3emp}_{c,k}^2 + \beta_7\texttt{cntyN4.m3emp}_{c,k}^3 + \epsilon_{c,k}\end{aligned} \tag{10}$$

where $\epsilon_{c,k} \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$. The predictors cntyN4.state and cntyN4.sector were chosen for similar reasons as the model in (9). The number of suppressed cells is important when predicting the amount missing from a cell higher in the hierarchy of industry codes. The cntyN4.m3emp is chosen over the other monthly employment counts since it was not imputed by our model (9). The square root transformation of the response variable is select trial and error until the residual plots suggested the assumptions on the residuals seemed reasonable. The structure of the equation for the predictors was chosen based on trial and error and comparing the AIC and residual plots across considered models. Once the model is fitted, we find five outliers (FIPS_NAICS-4: 21033_4452, 21107_5619, 36007_2361, 6007_3399, 6109_5231). These are removed and the model is refitted. For the wage cells at the NAICS-4 by county level that are not generate from CBP.cntyN4.qp1 or QWI.cntyN4.EarnBeg, we simulate $\sqrt{\texttt{qp1}_{c,k} - \Sigma_{c,k}^{(6,4)}}$ using the refitted model. Since $\Sigma_{c,k}^{(6,4)}$ is known for all combinations of counties and NAICS-4 codes, we transform the simulated value from the model to generate $\texttt{wage}_{c,k}$ which becomes the wage value for county c and NAICS-4 code k. Then we bound all $\texttt{wage}_{c,k}$ values below by $\Sigma_{c,k}^{(6,4)}$. In Algorithm 4, we find an upper bound for $\texttt{wage}_{c,k}$ using information from the county by NAICS-3, and county by NAICS sector and county levels. Denote the sum of available d-digit NAICS qp1 values in the code k of NAICS r level for county c as $\Sigma_{k,c}^{(d,r)}$. If the NAICS-3 by county $\texttt{CBP.cntyN3.qp1}_{c,j}$ value is not suppressed where j is the NAICS-3 code corresponding to NAICS-4 code k, then the maximum $\texttt{wage}_{c,k}$ is $\texttt{CBP.cntyN3.qp1}_{c,j} - \Sigma_{c,k}^{(4,3)}$. Otherwise, if the NAICS sector (i.e. 2 digit NAICS) by county $\texttt{CBP.cntySect.qp1}_{c,i}$ value is not suppressed where i is the NAICS Sector corresponding to NAICS-4 code k, then the maximum is $\texttt{CBP.cntySect.qp1}_{c,i} - \Sigma_{c,k}^{(3,2)}$. If both the NAICS 3-digit and NAICS sector are suppressed, we use the county total wages minus the sum of available NAICS sector wages as the maximum.

Algorithm 3: *(Wage for County by NAICS-4)* For an set of indices such as $\mathcal{I}_{/\texttt{cntyN4.qp1}}$ let the complement be denoted $\mathcal{I}^C_{/\texttt{cntyN4.qp1}}$.Let any sum of values over an empty set be equal to 0.

Input: QWI.cntyN4.EarnBeg, ; /* county by NAICS-4 average earnings values from QWI */

cntyN4.m1emp,cntyN4.m2emp, cntyN4.m3emp ; /* county by NAICS-4 employment values from Algorithm 1 */

CBP.cntyN6.qp1,CBP.cntyN4.qp1, CBP.cntyN3.qp1,CBP.cntySect.qp1, CBP.cnty.qp1 ; /* quarterly payroll values */

$\mathcal{I} = \{(c,k) : (c,k) \in \texttt{QWI.cntyN4} \cap \texttt{CBP.cntyN4}\}$, ; /* county by NAICS-4 indices in QWI and CBP */

cntyN4.state, cntyN4.sector, ; /* state and sector corresponding to county by NAICS-4 cells */

foreach $(c,k) \in \mathcal{I}$ **do**

  Let $\mathcal{I}^{(6,4)}_{c,k} = \{\ell \in$ NAICS-6 codes within k : $\texttt{CBP.cntyN6.qp1}_{c,\ell}$ is not missing$\}$;

  $\Sigma^{(6,4)}_{c,k} \leftarrow \sum_{\ell \in \mathcal{I}^{(6,4)}_{c,k}} \texttt{CBP.cntyN6.qp1}_{c,\ell}$ $M_{c,k} \leftarrow |\{\ell \in$ NAICS-6 codes within k : $\texttt{CBP.cntyN6.qp1}_{c,\ell}$ is suppressed$\}|$;

Let $\mathcal{I}_{/\texttt{cntyN4.qp1}} = \{(c,k) \in \mathcal{I}$: CBP.cntyN4.qp1 is missing$\}$;

Let $\mathcal{I}_{/\texttt{EarnBeg}} = \{(c,k) \in \mathcal{I}$: QWI.cntyN4.EarnBeg is missing$\}$;

foreach $(c,k) \in \mathcal{I}^C_{/\texttt{cntyN4.qp1}}$ **do**

  $\texttt{wage}_{c,k} \leftarrow \texttt{CBP.cntyN4.qp1}_{c,k}$; /* use CBP.cntyN4.qp1 if available */

Fit model (10) with $(c,k) \in \mathcal{I}^C_{/\texttt{cntyN4.qp1}}$;

If necessary, identify outliers based on residual normality plot and fitted verses residual plot. Remove them and refit model (10);

Let $\hat{f}_{\texttt{cntyN4.qp1}}(\cdot)$ be the function that takes in the predictor variable values and returns the fitted value of the model;

Let $se_{pred}(\hat{y})$ be the function that takes in a fitted value, $\hat{y}$, from the model and returns the standard error of the prediction;

foreach $(c,k) \in \mathcal{I}_{/\texttt{cntyN4.qp1}} \cap \mathcal{I}_{/\texttt{EarnBeg}}$ **do**

  $\hat{y}_{c,k} \leftarrow \hat{f}_{\texttt{cntyN4.qp1}}(\texttt{cntyN4.state}, \texttt{cntyN4.sector}, M_{c,k}, \texttt{cntyN4.m3emp})$;

  $\texttt{wage}_{c,k} \leftarrow (\hat{y}_{c,k} + Z_{c,k})^2 + \Sigma^{(6,4)}_{c,k}$ where $Z_{c,k} \overset{indep.}{\sim} N(0, se^2_{pred}(\hat{y}_{c,k}))$; /* predict wage value with model */

foreach $(c,k) \in \mathcal{I}_{/\texttt{cntyN4.qp1}} \cap \mathcal{I}^C_{/\texttt{EarnBeg}}$ **do**

  $w \leftarrow \texttt{QWI.cntyN4.EarnBeg}_{c,k}(\texttt{cntyN4.m1emp}_{c,k} + \texttt{cntyN4.m2emp}_{c,k} + \texttt{cntyN4.m3emp}_{c,k})$;

  $\texttt{wage}_{c,k} \leftarrow \max(w, \Sigma^{(6,4)}_{k,c})$; /* lower bound is $\Sigma^{(6,4)}_{k,c}$ */

/* Adjust based on an upper bound for the wage value */

foreach $(c,k) \in \mathcal{I}_{/\texttt{cntyN4.qp1}}$ **do**

  $u_{c,k} \leftarrow$ Wage Upper Bound$(c, k,$ CBP.cntyN3.qp1, CBP.cntySct.qp1, CBP.cnty.qp1$)$ where Wage Upper Bound is Algorithm 4;

  $\texttt{wage}_{c,k} = \min(u_{c,k}, \texttt{wage}_{c,k})$

cntyN4.wage is a vector of $\texttt{wage}_{c,k}$ for $(c,k) \in \mathcal{I}$;

Result: cntyN4.wage

Algorithm 4: *(Wage Upper Bound)* Using the CBP data, we get an upper bound on the quarterly wages for county c and NAICS-4 code k. Let the sum of qp1 values over an empty set be set equal to 0.

Input: (c,k); /* county and NAICS-4 code */

CBP.cntyN3.qp1,CBP.cntySect.qp1, CBP.cnty.qp1; /* quarterly payroll values from CBP data */

i and j are the Sector and the NAICS-3 codes corresponding to NAICS-4 code k, respectively;

if $\texttt{CBP.cntySect.qp1}_{c,i}$ *is not missing* **then**

  Let $\mathcal{I}^{(3,2)}_{c,k} = \{j' \in$ NAICS-3 codes within Sector code i : $\texttt{CBP.cntyN3.qp1}_{c,j'}$ is not missing$\}$;

  $u_{c,k} \leftarrow \texttt{CBP.cntySect.qp1}_{c,i} - \sum_{j' \in \mathcal{I}^{(3,2)}_{c,k}} \texttt{CBP.cntyN3.qp1}_{c,j'}$

if $\texttt{CBP.cntyN3.qp1}_{c,j}$ *is not missing* **then**

  Let $\mathcal{I}^{(4,3)}_{c,k} = \{k' \in$ NAICS-4 codes within NAICS-3 code j : $\texttt{CBP.cntyN4.qp1}_{c,k'}$ is not missing$\}$;

  $u_{c,k} \leftarrow \texttt{CBP.cntyN3.qp1}_{c,j} - \sum_{k' \in \mathcal{I}^{(4,3)}_{c,k}} \texttt{CBP.cntyN4.qp1}_{c,k'}$

if $\texttt{CBP.cntyN3.qp1}_{c,j}$ *and* $\texttt{CBP.cntySect.qp1}_{c,i}$ *are missing* **then**

  Let $\mathcal{I}^{(2)}_{c,k} = \{i' \in$ NAICS Sectors codes that appear in county c : $\texttt{CBP.cntySect.qp1}_{c,i'}$ is not missing$\}$;

  $u_{c,k} \leftarrow \texttt{CBP.cnty.qp1}_{c} - \sum_{i' \in \mathcal{I}^{(2)}_{c,k}} \texttt{CBP.cntySect.qp1}_{c,j'}$

Result: $u_{k,c}$

H.3 Synthetic County by NAICS-6 Data

We use the NAICS-4 by county data to create a NAICS-6 by county dataset (Algorithm 5). First we identify any NAICS-4 and county combinations which only have one NAICS-6 code within them according the CBP data. For these cells we set `cntyN6.m1emp`, `cntyN6.m3emp`, and `cntyN6.wage` to be the values from the corresponding NAICS-4 level and add a column for the number of establishments `CBP.cntyN6.estnum`. Then we focus on the NAICS-6 by county cells that have not been filled by the NAICS-4 by county values.

Since we are going to use the CBP imputed `imputeCBP.cntyN6.emp` value for the `cntyN6.m3emp` value at the NAICS-6 by county level, we need to adjust our `imputeCBP.cntyN6.emp` values to match the scale of the `QWI.cntyN4.EmpEnd` values from the QWI that inform some of the NAICS-4 `cntyN4.m3emp` values. We create a column in the NAICS-4 by county data,

$$\texttt{EmpScale} = \begin{cases} \frac{\texttt{imputeCBP.cntyN4.emp}}{\texttt{cntyN4.m3emp}} & \text{if } \texttt{cntyN4.m3emp} > 0 \\ 1 & \text{otherwise} \end{cases}.$$

For the NAICS-6 codes that did not automatically inherit the NAICS-4 value, we set `cntyN6.m3emp` to be the NAICS-6 `imputeCBP.cntyN6.emp` from the imputed CBP over the `EmpScale` for the corresponding county by NAICS-4 code. For example, in some county, consider the NAICS-4 code 2361, if the CBP `imputeCBP.cntyN4.emp` for NAICS-4 2361 is 1000 and the QWI `QWI.cntyN4.EmpEnd` = 1250, let our `cntyN4.m3emp` = 1250 for that county and NAICS-4 combination. Then the `EmpScale` = 0.8. If the county has CBP imputed `imputeCBP.cntyN6.emp` = 700 for NAICS-6 code 236115 and `imputeCBP.cntyN6.emp` = 300 for NAICS-6 code 236116, then our adjusted `cntyN6.m3emp` for that county are 875 for NAICS 236115 and 375 for NAICS 236116.

Algorithm 5: *(County by NAICS-6)*

Input: `cntyN4.m1emp`, `cntyN4.m2emp`, `cntyN4.m3emp`, `cntyN4.wage`, ; /* from Algorithms 1 and 3 */

`CBP.cntyN6.qp1`,`CBP.cntyN6.estnum`,`imputeCBP.cntyN6.emp` ; /* county by NAICS-6 values */

$\mathcal{I}_{\texttt{cntyN6}} = \{(c,k,\ell)$: county, NAICS-4, and NAICS-6 code combination corresponding to `imputeCBP.cntyN6` rows$\}$,

$\mathcal{I}_{\texttt{cntyN4}} = \{(c,k)$: county, NAICS-4 code combination corresponding to `cntyN4.m3emp` rows$\}$

foreach $(c,k) \in \mathcal{I}_{\texttt{cntyN4}}$ **do**

  $\mathcal{I}_{c,k} = \{\ell : (c,k,\ell) \in \mathcal{I}_{\texttt{cntyN6}}\}$;

  if $|\mathcal{I}_{c,k}| = 1$ **then**

    $\texttt{m1emp}_{c,\ell} = \texttt{cntyN4.m1emp}_{c,k}$ for $\ell \in \mathcal{I}_{c,k}$;

    $\texttt{m3emp}_{c,\ell} = \texttt{cntyN4.m3emp}_{c,k}$ for $\ell \in \mathcal{I}_{c,k}$;

    $\texttt{wage}_{c,\ell} = \texttt{cntyN4.wage}_{c,k}$ for $\ell \in \mathcal{I}_{c,k}$

  else

    if $\texttt{cntyN4.m3emp}_{c,k} > 0$ *and* $\texttt{imputeCBP.cntyN4.emp}_{c,k} > 0$ **then**

      $\texttt{EmpScale}_{c,k} = \frac{\texttt{imputeCBP.cntyN4.emp}_{c,k}}{\texttt{cntyN4.m3emp}_{c,k}}$;

      $\texttt{m3emp}_{c,\ell} = \texttt{imputeCBP.cntyN6.emp}_{c,\ell}/\texttt{EmpScale}_{c,k}$ for all $\ell \in \mathcal{I}_{c,k}$

    else

      $\texttt{m3emp}_{c,k} = \texttt{imputeCBP.cntyN6.emp}_{c,\ell}$ for all $\ell \in \mathcal{I}_{c,k}$

    $(\texttt{m1emp}_{c,\ell} : \ell \in \mathcal{I}_{c,k})^T \leftarrow$ Dirichlet Divider$((\texttt{m3emp}_{c,\ell} : \ell \in \mathcal{I}_{c,k})^T, \texttt{cntyN4.m1emp})$ where Dirichlet Divider is Algorithm 6;

    $\mathcal{I}^{(6,4)}_{c,k} = \{\ell \in \mathcal{I}_{c,k} : \texttt{CBP.cntyN6.qp1}_{c,\ell}$ is not missing$\}$;

    $\Sigma^{(6,4)}_{c,k} = \sum_{\ell \in \mathcal{I}^{(6,4)}_{c,k}} \texttt{CBP.cntyN6.qp1}_{c,\ell}$;

    $\texttt{wage}_{c,\ell} = \texttt{CBP.cntyN6.qp1}_{c,\ell}$ for $\ell \in \mathcal{I}^{(6,4)}_{c,k}$; /* use CBP.cntyN6.qp1 if available */

    $A_{c,k} = \texttt{cntyN4.wage}_{c,k} - \Sigma^{(6,4)}_{c,k}$;

    $(\texttt{wage}_{c,\ell} : \ell \in \mathcal{I}_{c,k} \cap \ell \notin \mathcal{I}_{c,k})^T \leftarrow$ Dirichlet Divider$((\texttt{m3emp}_{c,\ell} : \ell \in \mathcal{I}_{c,k} \cap \ell \notin \mathcal{I}_{c,k})^T, A_{c,k})$

`cntyN6.wage` is a vector of $\texttt{wage}_{c,\ell}$ for $(c,\ell) \in \mathcal{I}_{\texttt{cntyN6}}$;

`cntyN6.m1emp` is a vector of $\texttt{m1emp}_{c,\ell}$ for $(c,\ell) \in \mathcal{I}_{\texttt{cntyN6}}$;

`cntyN6.m3emp` is a vector of $\texttt{m3emp}_{c,\ell}$ for $(c,\ell) \in \mathcal{I}_{\texttt{cntyN6}}$;

`cntyN6.estnum` $\leftarrow$ `CBP.cntyN6.estnum`;

Result: `cntyN6.m1emp`, `cntyN6.m3emp`, `cntyN6.wage`, `cntyN6.estnum`

To get a `cntyN6.m1emp` for each NAICS-6 by county we use an algorithm based on the Dirichlet distribution which we call the Dirichlet Divider (Algorithm 6). The algorithm uses a concentration vector parameter of size d where d is the desired output size. The algorithm randomly divides an inputted total value into d subtotals proportional to the concentration parameter. In this case, our concentration parameter is `cntyN6.m3emp` at the NAICS-6 level corresponding to a given NAICS-4 code. If all concentration vector elements are 0, the

algorithm sets them all to be 1. The algorithm generates an observation from a Dirichlet distribution that is a random vector of d proportions that sum to 1. A inputted total value is multiplied by the random vector to get d non-negative subtotals which sum to to the total. In this case we use the `cntyN4.m1emp` value from the corresponding NAICS-4 code as the total and the resulting subtotals become the `cntyN6.m1emp` values for the NAICS-6 by county codes. We will not get `m2emp` values again until we are at the establishment level.

Algorithm 6: *(Dirichlet Divider)* Using a concentration input vector $\mathbf{b} \in \mathbb{R}^d_{\geq 0}$, the Dirichlet Divider randomly splits a total value A into d values, $a_1, a_2, \ldots, a_d$ such that $A = \sum_{i=1}^{d} a_i$ and $a_i \geq 0$ for $i = 1, \ldots, d$.

Input: $\mathbf{b} = (b_1\ b_2\ \ldots\ b_d)^T \in \mathbb{R}^d$ such that $b_i \geq 0$, total value $A \geq 0$
if $b_1 = b_2 = \ldots b_d = 0$ **then**
  $\mathbf{b} \leftarrow \mathbf{1} \in \mathbb{R}^d$
Generate $\mathbf{p} \sim$ Dirichlet($\mathbf{b}$) ;
$(a_1\ a_2\ \ldots\ a_d)^T \leftarrow \mathbf{p}A$;
Result: $\mathbf{a} = (a_1\ a_2\ \ldots\ a_d)^T$

Finally, we just need to derive wage values at the NAICS-6 by county level. We get the sum of NAICS-6-level `CBP.cntyN6.qp1` for a specific NAICS-4 code, $\Sigma^{(6,4)}_{c,k}$. Then we get the difference between that sum and the `cntyN4.wage` value at the NAICS-4 level. For NAICS-6 codes within the specified NAICS-4 code that do not have `CBP.cntyN6.qp1` values we use the Dirichlet Divider with the `cntyN6.m3emp` values as the concentration vector elements and $\texttt{CBP.cntyN4.qp1}_{c,k} - \Sigma^{(6,4)}_{c,k}$ as the total.

H.4 Establishment-level Synthetic Data

The last stage uses the County by NAICS-6 cell values and the Dirichlet Divider Algorithm 6 from previous steps (Algorithm 7). We randomly generate the concentration parameters for the Dirichlet Divider as independent and identically distributed random variables from a Gamma distribution with a shape of 10 and a scale parameter of 200 for each establishment in a NAICS-6 by county cell. We use these randomly generated concentration parameters across three applications of the Dirichlet Divider, one for wages, month 1 employment, and month 3 employment. The Gamma distribution produces positive-valued random variables with probability 1 which matches the parameter space of the Dirichlet distribution. Establishment data for wages and employment is often right-skewed [39, 45] and the gamma distribution has a right skew can be easily adjusted using the shape and scale parameters, making it an appropriate choice. Using the same shape parameters for the three applications of the Dirichlet Divider means an establishment should get roughly similar proportions of the wages and employment totals while still allowing the proportion to vary across the wage and employment values. This preserves a relationship between the number of employees and the quarterly wages, but allows the wage per employee to vary across establishments. The shape and scale parameters of the gamma distribution were chosen through trial and error to find values that would result in generated sets of proportions did not vary too much across replications and resulted in values that where not so close to zero that they would be often be rounded down to zero when applied to the total value. A low shape value such as 1 results in low and near zero proportions being generated and varies heavily across applications of the Dirichlet divider. In Figure 6, the proportions up to 0.1 across repetitions. High shape and scale parameters do not produce much variation across the establishments.

Using month 1 and 3 employment values from the Dirichlet Divider, month 2 employment is generated with Algorithm 2 and noise parameter 0.5. The state and higher levels of industry codes can be extracted from the county and NAICS-6 codes. In Table 6, five variables take constant values in our synthetic QCEW data. We complete the generation of the synthetic establishment dataset by adding these constant-valued variables to the dataset.

Potential improvements to this method would be to utilize the size class information from the CBP data to inform the generation of monthly employment values at the establishment level. Additionally, a more sophisticated imputation technique could be employed rather than simulating from regression models. The regression models could also be improved by incorporating more publicly-accessible county-level data or by using more subject-area informed variable selection.

I An additional result for QCEW

In Section 8.1, the quartiles and mean difference between the sanitized and original aggregate cell values for six aggregation levels are reported (Table 3). For a more complete picture of the distribution, the corresponding boxplots for four of these aggregation levels are include in Figure 7.

We can also compare what happens when groupings get more fine-grained. Figure 8 shows boxplots (median, 1st and 3rd quartile, and outliers) of the signed differences for the group by queries "3-digit NAICS prefix" (left) and "3-digit NAICS prefix within each county" (right). Thus, each group in the right-most plot (all establishments with the same 3-digit NAICS prefix) gets subdivided in the left-most plot (a group consists of all establishments in the same county having the same 3-digit NAICS prefix). Figure 8 shows high utility (the quartiles and median are near 0), except for the outliers caused by confidentiality concerns as explained above. By comparing to Figure 2, we see that outliers are caused when the true employment total within a group is small (and hence would raise confidentiality problems if the answer in

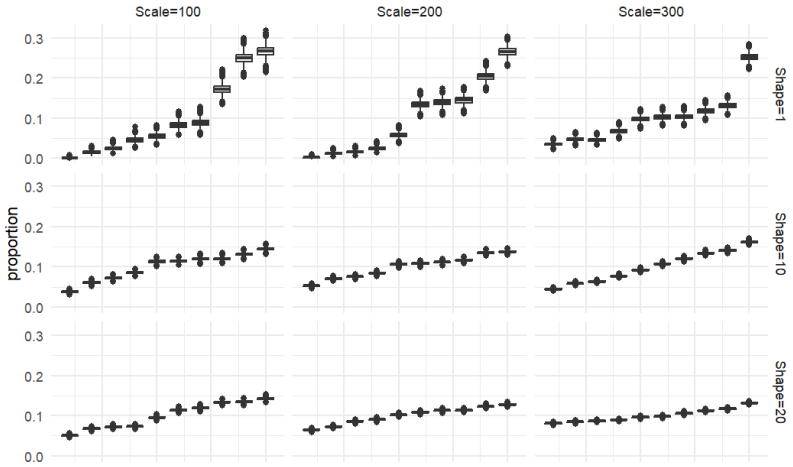

Figure 6: For each of nine different combinations of shape and scale parameters of a Gamma distribution a set of concentration parameters is generated for ten establishments. With each set of concentration parameters, the corresponding proportions are generated over 10,000 replications.

Algorithm 7: *(Employment and Wages Microdata) This algorithm uses the Dirichlet Divider (Algorithm 6* and the Month 2 Employment (Algorithm 2 to generate synthetic establishment-level monthly employment and quarterly wage values based on the output from County by NAICS-6 (Algorithm 5). Then from the NAICS-6 code and county code the remaining variable described in Table 6 are generated.

Input: `cntyN6.m1emp`, `cntyN6.m3emp`, `cntyN6.wage`, `cntyN6.estnum` ; /* From Algorithm 5 output */

1 $\eta > 0$; /* noise parameter for m2emp */

2 $\alpha_{prior}, \theta_{prior}$; /* shape and scale parameters for the Dirichlet prior respectively */

3 $n = 0$; /* initialize index for establishments */

4 **foreach** $(c, \ell) \in$ `cntyN6` **do**

5 $d \leftarrow$ `cntyN6.estnum`$_{c,\ell}$;

6 $\mathbf{a} = (a_1, \ldots, a_d)$ where $a_j \overset{i.i.d.}{\sim} \text{Gamma}(\alpha_{prior}, \theta_{prior})$;

7 `m1emp`$_{n+1}, \ldots$ `m1emp`$_{n+d} \leftarrow$ Dirichlet Divider($\mathbf{a}$, `cntyN6.m1emp`) where Dirichlet Divider is Algorithm 6;

8 `m3emp`$_{n+1}, \ldots$ `m3emp`$_{n+d} \leftarrow$ Dirichlet Divider($\mathbf{a}$, `cntyN6.m3emp`) ;

9 `wage`$_{n+1}, \ldots$ `wage`$_{n+d} \leftarrow$ Dirichlet Divider($\mathbf{a}$, `cntyN6.wage`) ;

10 `cnty`$_{n+j} = c$ for $j = 1, \ldots, d$;

11 `naics`$_{n+j} = \ell$ for $j = 1, \ldots, d$;

12 `primary_key`$_{n+j} = n + j$ for $j = 1, \ldots, d$;

13 `m2emp`$_{n+j} \leftarrow$ Month 2 Employment(η, `m1emp`$_{n+j}$, `m3emp`$_{n+j}$) for $j = 1, \ldots, d$ where Month 2 Employment is Algorithm 2;

14 $n = n + d$;

/* fill in other variables from naics and cnty values */

15 **foreach** `id` $\in$ `primary_key` **do**

16 Extract `supersector`$_{\texttt{id}}$, `sector`$_{\texttt{id}}$, `naics3`$_{\texttt{id}}$, `naics4`$_{\texttt{id}}$, and `naics5`$_{\texttt{id}}$ from `naics`$_{\texttt{id}}$;

17 Extract `state`$_{\texttt{id}}$ from `cnty`$_{\texttt{id}}$;

18 Set constant-valued variables `year`$_{\texttt{id}} = 2016$, `qtr`$_{\texttt{id}} = 1$, `own`$_{\texttt{id}} = 5$, `can_agg`$_{\texttt{id}} =$ "Y", and `rectype`$_{\texttt{id}} =$ "C"

Result: dataset with a row for each `primary_key` and columns for each variable from Table 6

such a group was too accurate, especially if the group only had 1 establishment). For this reason, plots of confidentiality-preserving vs. true value (e.g., Figure 2) appear to be more informative than the distribution of raw differences across groups (e.g., Figure 8) as they present context that is not present in Figure 8.

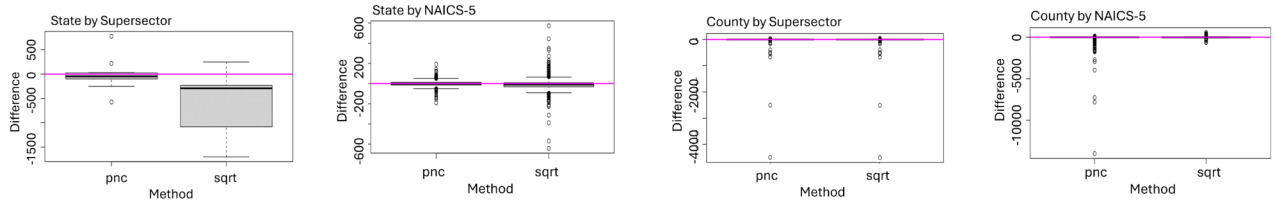

Figure 7: Comparison of the differences of the output from the two disclosure limitation methods and the original values for the state at different levels of aggregation. Going from left to right we can see the distribution of the difference for the totals of cells by super-sector (aggregation level 53), 5-digit NAICS (aggregation level 57), super-sector by county (aggregation level 73), and 5-digit NAICS by county (aggregation level 57).

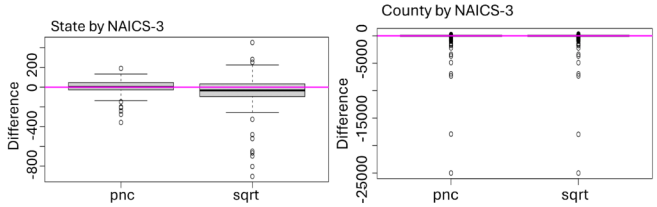

Figure 8: (Employment) Box-plot of signed difference for group-by 3-digit NAICS prefix (left) vs. group-by county and 3-digit NAICS prefix (right). Box-plots show median, 1st and 3rd quartiles, and the outliers. Both plots show downward bias. Median and quartiles for the right plot are all near 0.

q	Query	μ_q	μ_q^2/μ^2	$\mu_{q,e}$	$\mu_{q,w}$	$\mu_{q,w}^2/\mu_q^2$
I	Identity	1.22	.281	0.70	0.15	.015
Σ	State Total	0.36	.024	0.20	0.10	.077
$N5$	State NAICS-5	1.05	.207	0.60	0.15	.020
C	County Total	1.05	.207	0.60	0.15	.020
$CxN5$	County by NAICS-5	1.22	.281	0.70	0.15	.015

Table 9: The baseline privacy loss budget allocation to queries. The overall privacy budget $\mu \approx 2.31$ is allocated to five queries. The query level budget is denoted μ_q for query q. Within each query q, the budget is allocated to each month's employment (denoted $\mu_{q,e}$) and quarterly wage (denoted $\mu_{q,w}$). Columns with headings $\mu_{q,e}$ and $\mu_{q,w}$ are exact values; everything else is computed from them using the composition rule (square root of the sum of squares) and rounded.

J Additional results from synthetic evaluation data

For our experiments, we focus on synthetic data for New Jersey and look at error metrics for quarterly wage and month 3 employment level query values. The creation of this synthetic dataset is described in Section H. We use two additional utility metrics in the appendix: absolute difference and relative absolute difference. We define absolute difference as $|x - \widetilde{x}|$ where x is a confidential cell value of a query and $\widetilde{x}$ is the corresponding sanitized cell value. The relative absolute difference is the absolute difference scaled by the original value (i.e. $|x - \widetilde{x}|/(x + 1)$). When we summarize the utility of cells across a query we use the mean for absolute difference and the median for the relative absolute difference. Median metrics are more robust to outliers than mean metrics so the combination of the two can be helpful to capture error behavior. However, the state total (51) and state by domain (52) queries only have one and two cells respectively, so these summary statistics should be identical in these cases. Additionally, we may be interested in how the distribution of the cell values within a query and the number of establishments aggregated into a cell will influence the utility of the queries. In Table 10 and Table 11, for each query we show the number of cells and some summary statistics for the number of establishments in each cell. The 3-,4-, and 5- digit NAICS codes at both the state and county level have a wide range of establishments per cell with a minimum of 1 and the maximum over 1,000. Most queries result in the standard deviation of establishments per cell that is larger than the mean number of establishments. All this suggests highly variable cells in terms of the number of establishments.

In Section 8.2, the GDP mechanism is omitted from several plots (Center and right plots of Figure 3 and Figure 4) because the scale of its error is so large it obscures the differences between the pnc-mechanism and $\sqrt{\ }$-mechanism. The plots including the GDP mechanism is shown in Figures 10 and Figures 11.

Synthetic New Jersey State-level Queries					
	Min	Median	Max	Mean	Std.Dev.
State by Domain (52) has 2 cells					
n	28476	103884	179292	103884	106643
emp	389796	1513961	2638126	1513961	1589809
w	56870451	65671540	74472630	65671540	12446620
State by Supersector (53) has 10 cells					
n	250	22542	44429	20777	14141
emp	3457	236286	648521	302792	245261
w	1164792	5928548	51028965	13134308	16676271
State NAICS Sector (54) has 22 cells					
n	67	5070	29073	9444	9538
emp	1261	97925	561919	137633	129971
w	34977	2829067	24074883	5970140	7101329
State NAICS 3-digit (55) has 74 cells					
n	1	641	29073	2808	5253
emp	9	18359	298154	40918	61407
w	97	596856	19955782	1774906	3123361
State NAICS 4-digit (56) has 258 cells					
n	1	144	16950	805	1716
emp	1	3679	200556	11736	21559
w	39	161771	9017144	509082	1250695
State NAICS 5-digit (57) has 571 cells					
n	1	79	16950	364	1075
emp	1	1530	200556	5303	14205
w	39	40925	8276039	230023	747805
State NAICS 6-digit (58) has 853 cells					
n	1	33	8172	244	692
emp	0	819	140800	3550	10231
w	0	21351	8276039	153978	591643

Synthetic New Jersey County Queries					
	Min	Median	Max	Mean	Std.Dev.
County Total (71) has 21 cells					
n	999	9385	28689	9894	6968
emp	15581	157835	370634	144187	103475
w	938253	5954245	16382122	6254432	5017341
County by Supersector (73) has 208 cells					
n	1	676	6170	999	1069
emp	5	6844	99154	14557	18051
w	279	131295	9177485	631457	1344169
County NAICS Sector (74) has 454 cells					
n	1	177	4146	458	638
emp	1	3280	77820	6669	9323
w	56	77536	8096919	289302	777652
County NAICS 3-digit (75) has 1324 cells					
n	1	33	4146	157	348
emp	0	694	44129	2287	4458
w	5	13925	8096919	99202	440438
County NAICS 4-digit (76) has 3992 cells					
n	1	12	2110	52	121
emp	0	209	28438	758	1704
w	0	3920	5935630	32902	201878
County NAICS 5-digit (77) has 8026 cells					
n	1	7	2110	26	79
emp	0	84	25265	377	1159
w	0	1180	5935630	16365	137917
County NAICS 6-digit (78) has 10067 cells					
n	1	5	1289	21	55
emp	0	63	25265	301	933
w	0	834	5929093	13047	120849

Table 10: Summary statistics for number of establishments in a cell (*n*), month 3 employment count (*emp*), and quarterly wages (*w*) for each aggregate level code for the synthetic New Jersey. Overall, synthetic New Jersey has 207,768 establishments.

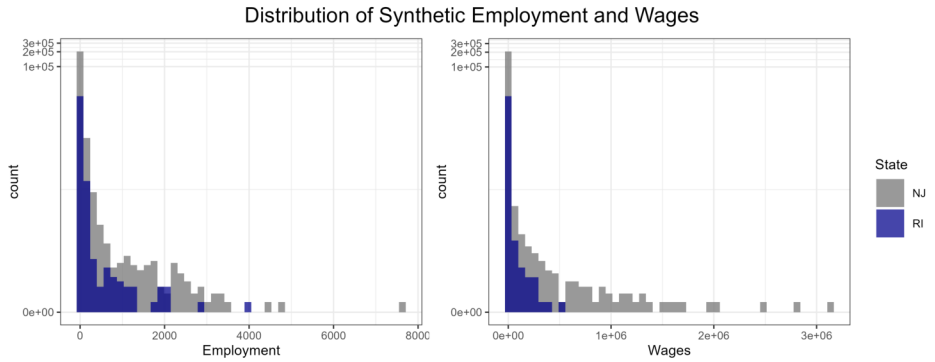

Figure 9: The distributions of wages and employment of the establishments in the Synthetic New Jersey and Rhode Island Datasets features a heavy tail. The skew of the data is so intense that the scale of the y-axis have to be transformed so the counts of the higher values bins are visible.

For all additional experiments, we use a baseline pnc-mechanism implementation with $\zeta = 0.01$ and $\sqrt{}$-mechanism implementation for comparison. This baseline matches the parameters used in Section 8. The baseline overall privacy budget is $\mu = 2.31$ and the privacy distance for monthly employment is $\gamma_{emp} = 0.5$ and for quarterly wages is $\gamma_w = 50$. The baseline implementations use the queries and privacy parameters described in Table 9

Synthetic Rhode Island State-level Queries					
	Min	Median	Max	Mean	Std.Dev.
State by Domain (52) has 2 cells					
n	4439	12710	20981	12710	11697
emp	57050	188016	318981	188016	185213
w	5985933	5989127	5992321	5989127	4517
State by Supersector (53) has 10 cells					
n	47	2994	4903	2542	1570
emp	297	34772	100286	37603	30426
w	64248	548588	5309640	1197825	1606737
State NAICS Sector (54) has 22 cells					
n	14	687	3272	1155	1165
emp	111	14654	79298	17092	17778
w	35043	240442	3079880	544466	769112
State NAICS 3-digit (55) has 68 cells					
n	4	109	2986	374	634
emp	8	2604	42028	5530	7873
w	104	73695	1250070	176151	247533
State NAICS 4-digit (56) has 224 cells					
n	1	32	2409	113	227
emp	1	674	36147	1679	3292
w	27	19038	670281	53474	104387
State NAICS 5-digit (57) has 484 cells					
n	1	16	2409	53	144
emp	1	200	36147	777	2278
w	1	4243	670281	24748	71079
State NAICS 6-digit (58) has 663 cells					
n	1	10	1144	38	94
emp	1	147	20274	567	1617
w	1	2604	670281	18067	58239

Synthetic Rhode Island County Queries					
Variable	Min	Median	Max	Mean	Std.Dev.
County Total (71) has 5 cells					
n	1102	3354	14414	5084	5340
emp	10955	38191	237653	75206	92665
w	699819	1959523	4255187	2395651	1470599
County by Supersector (73) has 50 cells					
n	1	274	2764	508	634
emp	1	2936	68784	7521	12431
w	1232	63424	2346055	239565	432476
County NAICS Sector (74) has 107 cells					
n	1	104	1957	238	384
emp	1	1462	54369	3514	6694
w	87	25122	1314614	111946	241131
County NAICS 3-digit (75) has 303 cells					
n	1	17	1698	84	197
emp	0	328	22958	1241	2834
w	0	6722	1175186	39532	109617
County NAICS 4-digit (76) has 854 cells					
n	1	8	1323	30	73
emp	0	104	19551	440	1204
w	0	2232	510039	14026	48852
County NAICS 5-digit (77) has 1620 cells					
n	1	4	1323	16	49
emp	0	42	19551	232	861
w	0	702	510039	7394	34446
County NAICS 6-digit (78) has 1996 cells					
n	1	4	610	13	34
emp	0	35	12845	188	674
w	0	498	510039	6001	30892

Table 11: Summary statistics for number of establishments (*n*), employment count (*emp*), and quarterly wages (*w*) for each aggregate level code for the synthetic Rhode Island. Overall, synthetic Rhodes Island has 25,420 establishments.

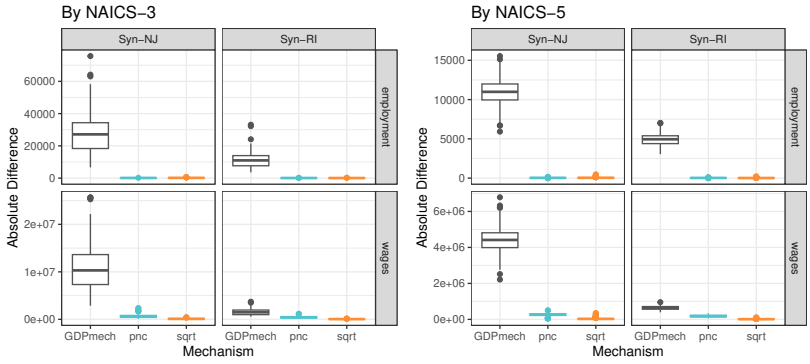

Figure 10: Absolute error for each query group averaged across 34 replicates including the GDP mechanism.

When we vary privacy allocation with a fixed overall budget, we scale the privacy share of a query or confidential value type (i.e. wages or employment). This share is described as the squared ratio of the privacy cost to its next highest cost level. In other words, for query k the wage privacy share is the squared privacy cost to wage ($\mu^2_{Q,w}$) over the query-level privacy budget (μ^2_Q). For the baseline parameters this value is in the last column of Table 9. When we look at privacy allocation to queries, we look at the squared query budget μ^2_Q over the squared total budget μ^2. For the baseline parameters, this is fourth column labeled μ^2_Q/μ^2 in Table 9.

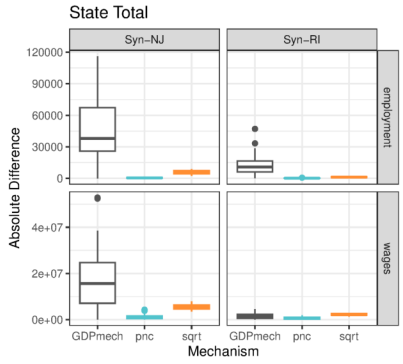

Figure 11: Absolute error across 34 replicates including the GDP mechanism.

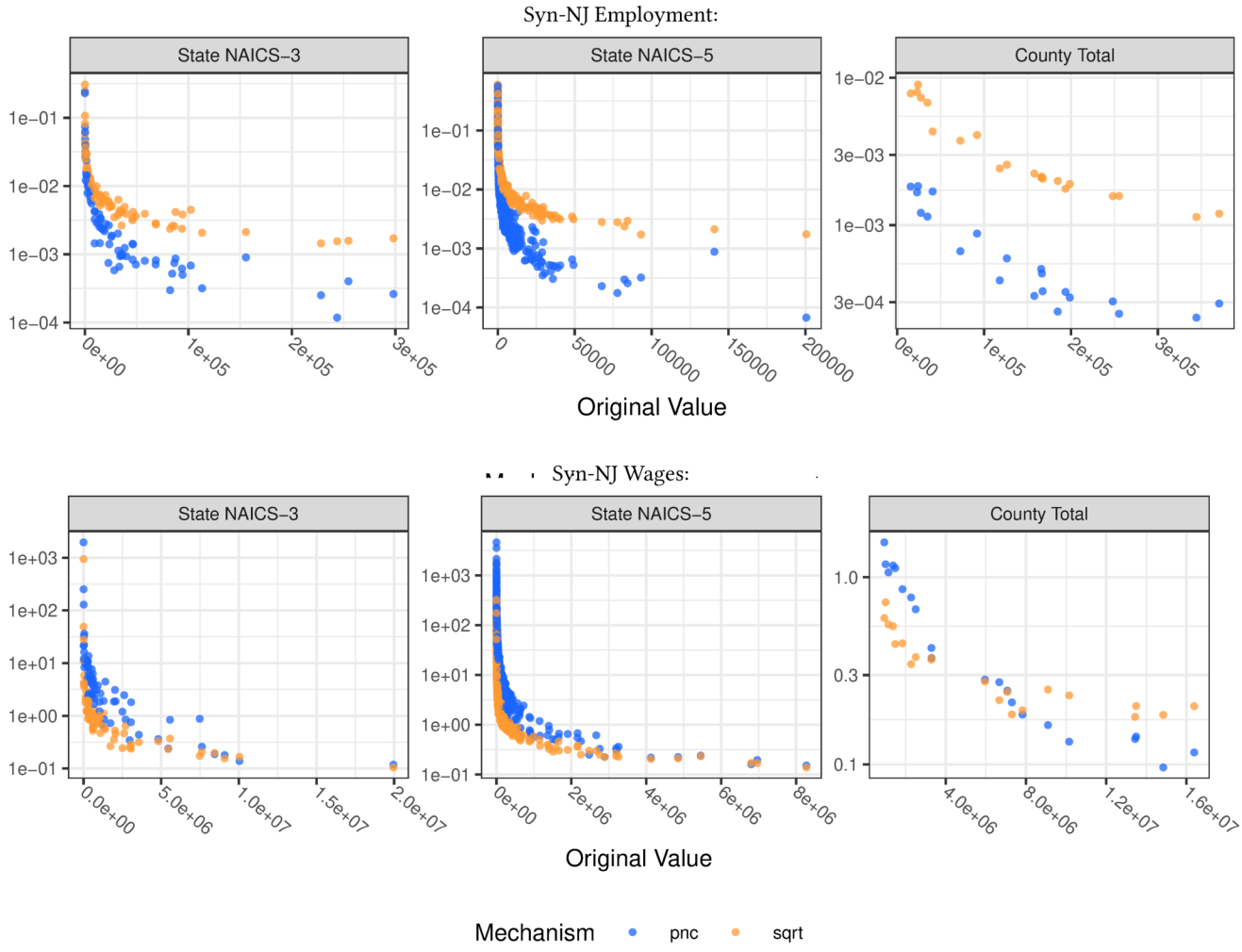

Figure 12: Comparison of pnc-mechanism in blue vs. $\sqrt{\ }$-mechanism in orange on Synthetic NJ for employment and wages, using the baseline configuration from Table 9 ($\mu \approx 2.31$). The first column is the query that groups by 3-digit NAICS prefix (not a measurement query), while the other two columns evaluate accuracy on 2 measurement queries. x-axis: ground truth answer for a group, y-axis: corresponding relative error for that group.

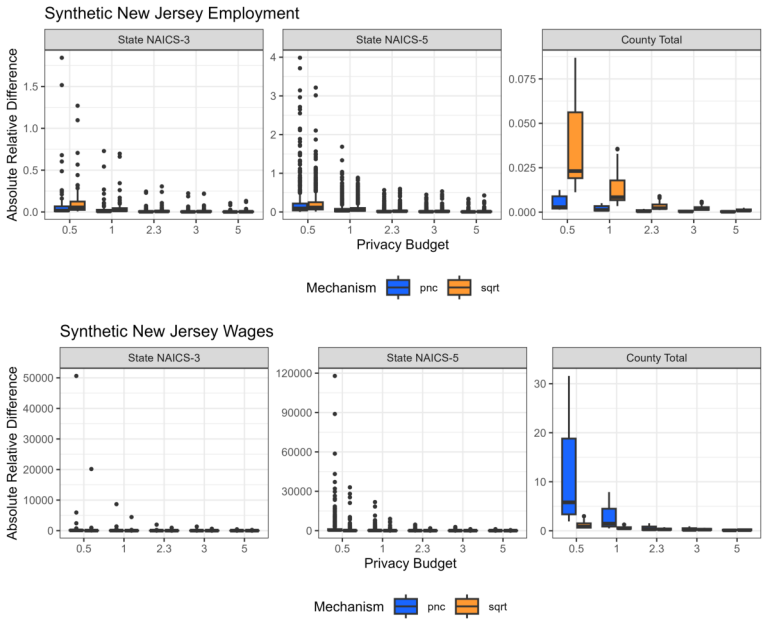

Figure 13: Employment and Wages relative errors, as overall privacy budget varies, using the base configuration from Table 9.