Comparison of Poisson-Gamma and Laplace Mechanisms for Differential Privacy

Harrison Quick (Drexel University)

Table of Contents

Motivating Use-Case: CDC WONDER

Generating Differentially Private Synthetic Data

Illustrative Example: Cancer Mortality in Pennsylvania

Summary

Table of Contents

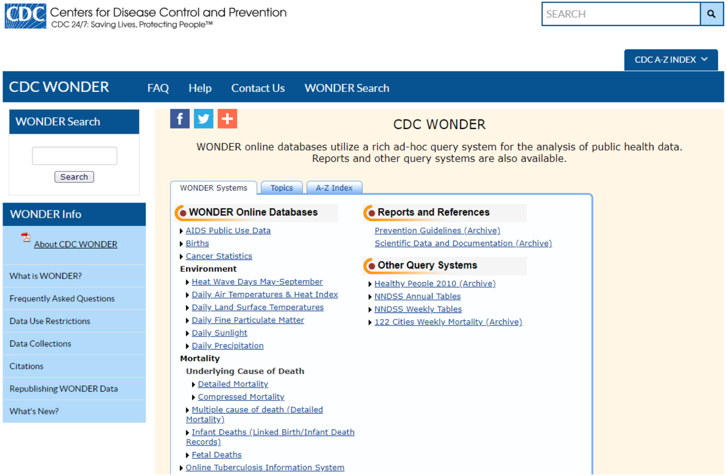
Motivating Use-Case: CDC WONDER

Generating Differentially Private Synthetic Data

Illustrative Example: Cancer Mortality in Pennsylvania

Summary

CDC WONDER



CDC WONDER

WONDER online databases utilize a rich ad-hoc query system for the analysis of public health data. Reports and other query systems are also available.

WONDER Online Databases

- AIDS Public Use Data
- Births
- Cancer Statistics

Environment

- Heat Wave Days May-September
- Daily Air Temperatures & Heat Index
- Daily Land Surface Temperatures
- Daily Fine Particulate Matter
- Daily Sunlight
- Daily Precipitation

Mortality

Underlying Cause of Death

- Detailed Mortality
- Compressed Mortality
- Multiple cause of death (Detailed Mortality)
- Infant Deaths (Linked Birth/Infant Death Records)
- Fetal Deaths
- Online Tuberculosis Information System

Reports and References

Prevention Guidelines (Archive)

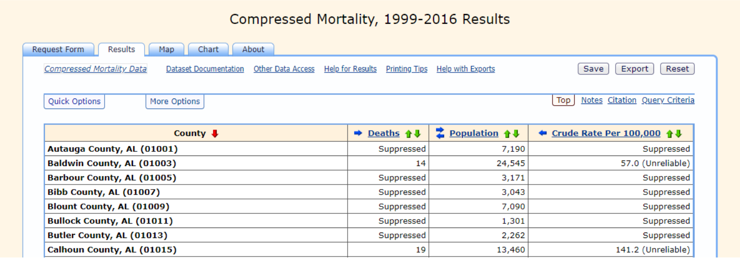
Scientific Data and Documentation (Archive)

Other Query Systems

- Healthy People 2010 (Archive)
- NNDSS Annual Tables
- NNDSS Weekly Tables
- 122 Cities Weekly Mortality (Archive)

CDC WONDER

County-level heart disease-related death counts for ages 35–44 in 2016 from all races and all genders



Compressed Mortality, 1999-2016 Results

County	Deaths	Population	Crude Rate Per 100,000
Autauga County, AL (01001)	Suppressed	7,190	Suppressed
Baldwin County, AL (01003)	14	24,545	57.0 (Unreliable)
Barbour County, AL (01005)	Suppressed	3,171	Suppressed
Bibb County, AL (01007)	Suppressed	3,043	Suppressed
Blount County, AL (01009)	Suppressed	7,090	Suppressed
Bullock County, AL (01011)	Suppressed	1,301	Suppressed
Butler County, AL (01013)	Suppressed	2,262	Suppressed
Calhoun County, AL (01015)	19	13,460	141.2 (Unreliable)

All counts less than 10 are suppressed in public-use datasets

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy
  - Targeted attacks by clever intruders can overcome data suppression to uncover the true counts

Is there a way that CDC can address these issues?

Table of Contents

Motivating Use-Case: CDC WONDER

Generating Differentially Private Synthetic Data

Illustrative Example: Cancer Mortality in Pennsylvania

Summary

Synthetic Data

One option to address the issue of data suppression would be to release *synthetic data*: e.g., if

- $\mathbf{y} = (y_1, \ldots, y_I)^T$ denotes a restricted-use dataset of I observations,
- $p(\mathbf{y} \mid \phi)$ is an appropriate statistical model for $\mathbf{y}$ with parameters ϕ, and
- $p(\phi \mid \psi)$ is a prior distribution for ϕ given hyperparameters, ψ,

then we can generate a synthetic dataset, $\mathbf{z} = (z_1, \ldots, z_I)^T$, from the posterior predictive distribution,

$$p(\mathbf{z} \mid \mathbf{y}, \psi) = \int p(\mathbf{z} \mid \phi)\, p(\phi \mid \mathbf{y}, \psi)\, d\phi.$$

That is, we can sample ϕ^* from $p(\phi \mid \mathbf{y}, \psi)$ and then sample $\mathbf{z}$ from $p(\mathbf{z} \mid \phi^*)$.

- Natural next question: How do we know if synthetic data generated from $p(\mathbf{z} \mid \mathbf{y}, \psi)$ are sufficiently protective?

Differential Privacy (Dwork, 2006)

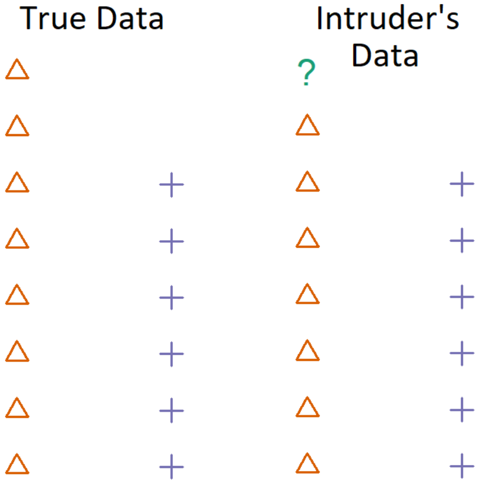
The standard typically used for demonstrating formal privacy guarantees is the concept of *differential privacy* (Dwork, 2006).
In this context, $p\left(\mathbf{z} \,|\, \mathbf{y}, \boldsymbol{\psi}\right)$ is ϵ-differentially private if for any similar[1] dataset, $\mathbf{x}$,

$$\left| \log \frac{p\left(\mathbf{z} \,|\, \mathbf{y}, \boldsymbol{\psi}\right)}{p\left(\mathbf{z} \,|\, \mathbf{x}, \boldsymbol{\psi}\right)} \right| \leq \epsilon. \tag{1}$$

While $\boldsymbol{\psi}$ *can* be viewed as a vector of model parameters, *in practice* the elements of $\boldsymbol{\psi}$ are merely specified to satisfy ϵ-differential privacy.

[1] $\|\mathbf{x} - \mathbf{y}\| = 2$ and $\sum_i x_i = \sum_i y_i$ — i.e., there exists i and i' such that $x_i = y_i - 1$ and $x_{i'} = y_{i'} + 1$ with all other values equal

What is Differential Privacy? Simple (Conventional) Example



Suppose we want to release the proportion of triangles in this dataset without disclosing any individual shape.

- Worst case scenario: Intruder knows all but one shape
  - Releasing the true value $(8/14)$ compromises the remaining shape
- Let's add noise to the proportion such that

$$\frac{8+\text{noise}}{14} \approx \frac{7+\text{noise}}{14}$$

where the amount of noise depends on the level of protection desired (measured by ϵ)

  - e.g., noise $\sim \text{Lap}\,(0, 1/\epsilon)$

Laplace Mechanism

In theory, the Laplace mechanism is pretty straightforward:

$$z_i = y_i + e_i, \quad \text{where } e_i \sim \text{Lap}\left(0, 1/\epsilon\right)$$

but because the Laplace mechanism can produce *negative* values, some post processing is required to produce sensible values.

- Suppose we desire synthetic values $z_i \geq 0$, for $i = 1, \ldots, I$ such that $\sum_i z_i = z.$, then we can let

$$z_i^* = \lceil y_i + e_i \rceil^+, \quad \text{where } e_i \sim \text{Lap}\left(0, 1/\epsilon\right),$$
$$z_i = z_i^* \times \frac{z.}{\sum_i z_i^*}$$

  - Note: This approach will produce *non-integer* values, but that feels less awkward than *negative* values...

Laplace Mechanism with Hierarchical Structure

Now suppose our data are indexed by multiple factors (e.g., age, race, county), denoted y_{ijk}. While we could simply let z_{ijk} be defined by

$$\begin{aligned} z^*_{ijk} &= \lceil y_{ijk} + e_{ijk} \rceil^+, \ \ \text{where } e_{ijk} \sim \text{Lap}\,(0, 1/\epsilon), \\ z_{ijk} &= z^*_{ijk} \times \frac{\sum_{ijk} y_{ijk}}{\sum_{ijk} z^*_{ijk}}, \end{aligned}$$

we may instead *try* to preserve utility at certain aggregate levels; e.g.,

$$\begin{aligned} z_{i\cdot\cdot} &= z^*_{i\cdot\cdot} \times \frac{z_{\cdot\cdot\cdot}}{\sum_i z^*_{i\cdot\cdot}}, && \text{where } z^*_{i\cdot\cdot} = \lceil y_{i\cdot\cdot} + e_{i\cdot\cdot} \rceil^+ && \text{and } e_{i\cdot\cdot} \sim \text{Lap}\,(0, 1/\epsilon_1) \\ z_{ij\cdot} &= z^*_{ij\cdot} \times \frac{z_{i\cdot\cdot}}{\sum_j z^*_{ij\cdot}}, && \text{where } z^*_{ij\cdot} = \lceil y_{ij\cdot} + e_{ij\cdot} \rceil^+ && \text{and } e_{ij\cdot} \sim \text{Lap}\,(0, 1/\epsilon_2) \\ z_{ijk} &= z^*_{ijk} \times \frac{z_{ij\cdot}}{\sum_k z^*_{ijk}}, && \text{where } z^*_{ijk} = \lceil y_{ijk} + e_{ijk} \rceil^+ && \text{and } e_{ijk} \sim \text{Lap}\,(0, 1/\epsilon_3) \end{aligned}$$

such that $\epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon$ and where $z_{\cdot\cdot\cdot} = y_{\cdot\cdot\cdot} = \sum_{ijk} y_{ijk}$.

A Working Hypothesis...

The hypothesis underlying my work in data privacy is that synthetic data generated from the true data generating process will outperform a noisy version of the true data.

Thus, my strategy is to specify a model that (a) is statistically appropriate and (b) can be proven to satisfy differential privacy, and then hope the posterior predictive distribution can approximate the true data generating process.

- In the next few slides, I'll discuss two simple model specifications — the multinomial-Dirichlet model and the Poisson-gamma model — that have been proven to satisfy differential privacy.
- Along the way, I will discuss the appropriateness of these models for synthesizing public health data — e.g., county-level death counts.

Multinomial-Dirichlet model (Machanavajjhala et al., 2008)

Let $\mathbf{y}$ be a vector of sensitive count data of length $I \geq 2$ with $\sum_i y_i = y.$ and assume

$$\mathbf{y} \,|\, \boldsymbol{\theta} \sim \text{Mult}\left(y., \boldsymbol{\theta}\right) \text{ and } \boldsymbol{\theta} \sim \text{Dir}\left(\boldsymbol{\alpha}\right).$$

To generate a synthetic data vector, $\mathbf{z}$, with a given $\sum_i z_i = z. = y.$:

1. Sample $\boldsymbol{\theta}^*$ from its posterior, $\boldsymbol{\theta} \,|\, \mathbf{y} \sim \text{Dir}\left(\mathbf{y} + \boldsymbol{\alpha}\right)$
2. Sample $\mathbf{z}$ from the posterior predictive distribution, $\mathbf{z} \sim \text{Mult}\left(z., \boldsymbol{\theta}^*\right)$

It can (but won't) be shown that if

$$\min \alpha_i \geq z. / \left[\exp\left(\epsilon\right) - 1\right],$$

the multinomial-Dirichlet synthesizer, $p\left(\mathbf{z} \,|\, \mathbf{y}, \boldsymbol{\alpha}\right)$, will satisfy ϵ-differential privacy.

- If our $\text{Dir}\left(\boldsymbol{\alpha}\right)$ prior is informative enough, it can sufficiently mask the data...
  - ... but it will do so by allocating events *uniformly*, which is bad.
  - e.g., if ϵ is small, the model will try to assign the same number of deaths to Small Town, PA as it would to Philadelphia.

Poisson-Gamma model (Quick, 2021)

Motivated by the field of disease mapping — where death data are typically modeled as being Poisson distributed — Quick (2021) proposed assuming

$$y_i \mid \lambda_i \sim \text{Pois}\left(n_i \lambda_i\right) \text{ and } \lambda_i \sim \text{Gamma}\left(a_i, b_i\right)$$

which implies $\lambda_i \mid y_i \sim \text{Gamma}\left(y_i + a_i, n_i + b_i\right)$. Now recall that if the y_i are (conditionally) independent Poisson random variables, then

$$\mathbf{y} \mid \boldsymbol{\lambda}, \sum_i y_i = y_{\cdot} \sim \text{Mult}\left(y_{\cdot}, \left\{\frac{n_i \lambda_i}{\sum_j n_j \lambda_j}\right\}\right)$$

Thus, we can generate synthetic data by:

1. Sampling λ_i^* from $\text{Gamma}\left(y_i + a_i, n_i + b_i\right)$ for $i = 1, \ldots, I$
2. Sampling $\mathbf{z} \sim \text{Mult}\left(z_{\cdot}, \left\{n_i \lambda_i^* / \sum_j n_j \lambda_j^*\right\}\right)$

But under what conditions will this satisfy ϵ-differential privacy?

Poisson-Gamma model — ϵ-differential privacy

It *can* (but won't) be shown that the Poisson-gamma synthesizer, denoted $p\left(\mathbf{z} \,|\, \mathbf{y}, \mathbf{a}, \mathbf{b}\right)$, will satisfy ϵ-differential privacy if

$$a_i \geq \frac{z_{\cdot}}{e^{\epsilon}/\nu_i - 1} \tag{2}$$

where $\nu_i \in [1, 2]$ denotes what amounts to a *penalty* term associated with the additional information gained from using the Poisson-gamma model compared to the multinomial-Dirichlet model.

- It would take too much time/space to write out the expression for ν_i, but it's a function of the group-specific population sizes and prior event rates.
- If the group-specific population sizes and prior event rates are equal, then $\nu_i = 1$ for all groups, thus making the M-D and P-G models *mathematically* equivalent.

Drawback of the Poisson-Gamma model of Quick (2021)

Unlike the multinomial-Dirichlet model, the Poisson-gamma model behaves fairly well when ϵ is small.

- i.e., the model will allocate events based on the population sizes, n_i, and the prior expected event rates, $\lambda_{i0} = a_i/b_i$, thus if these values were chosen "wisely", we won't get *terrible* synthetic data like Small Town, PA $\approx$ Philadelphia

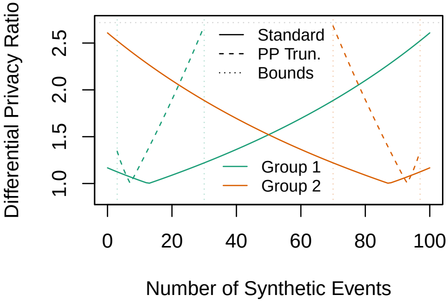
Unfortunately, another problem with the multinomial-Dirichlet model that *is* shared by the Poisson-gamma model of Quick (2021) is that when the total number of events, $y_\cdot$, is large, *very* informative priors are required to satisfy even *moderate* values of ϵ.

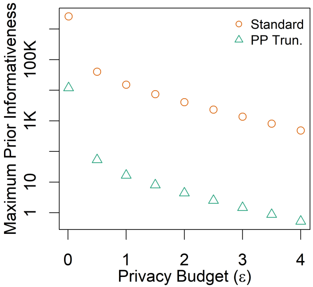
- While it is unlikely that Small Town, PA would be assigned as many deaths as Philadelphia, our privacy protections are designed to guard against this possibility, and that's where the issues arise.

As a result, the synthetic data typically just reflect the prior information.

Prior Predictive Truncated Poisson-Gamma model (Quick, 2022)

To combat this, Quick (2022) proposed using the *prior predictive distribution* to truncate the synthetic data to a "reasonable" range of values.





Example on the left has $E\left[\mathbf{y} \mid \mathbf{a}, \mathbf{b}\right] = (15, 85)^T$ and $\epsilon = 1$

- Standard approach: Risk increases *slowly* up to $e^1 = 2.71$ at $z_1 = 100$.
- Truncated approach: Risk increases *quickly* up to $e^1 = 2.71$ at $z_1 = 30$.

The speed of the risk increase is driven by how *not* informative the prior distribution is

- Plot on the right is from the cancer example I will talk about momentarily...

Table of Contents

Motivating Use-Case: CDC WONDER

Generating Differentially Private Synthetic Data

Illustrative Example: Cancer Mortality in Pennsylvania

Summary

Cancer-related Deaths in Pennsylvania Counties in 1980

Attribute	**Levels**
County	$i = 1, \ldots, 67$ Counties in Pennsylvania
Cancer Type	$c = 1, \ldots, 9$ Forms of Cancer Cancers of the lip, oral cavity, and pharynx (ICD-9: 140–149); Cancers of the digestive organs and peritoneum (ICD-9: 150–159); Cancers of the respiratory and intrathoracic organs (ICD-9: 160–165) Cancers of the breast (ICD-9: 174–175); Cancers of the genital organs (ICD-9: 179–187); Cancers of the urinary organs (ICD-9: 188–189); Cancers of all other and unspecified sites (ICD-9: 170–173, 190–199); Leukemia (ICD-9: 204–208); and all other cancers of the lymphatic and hematopoietic tissues (ICD-9: 200–203)
Age	$a = 1, \ldots, 13$ Levels Ages under 1; Ages 1–4; Ages 5–9; Ages 10–14; Ages 15–19; Ages 20–24; Ages 25–34; Ages 35–44; Ages 45–54; Ages 55–64; Ages 65–74; Ages 75–84; and Ages 85 and older
Race	$r = 1, \ldots, 3$ Levels (Black, White, and Other)
Sex	$s = 1, 2$ Levels (Male and Female)

In total, there were $y. = \sum_{icars} y_{icars} = 26{,}116$ cancer-related deaths in PA in 1980 belonging to these $67 \times 13 \times 9 \times 3 \times 2 = 47{,}034$ strata.

Cancer-related Deaths in Pennsylvania Counties in 1980

Attribute	**Levels**
County	$i = 1, \ldots, 67$ Counties in Pennsylvania
Cancer Type	$c = 1, \ldots, 9$ Forms of Cancer Cancers of the lip, oral cavity, and pharynx (ICD-9: 140–149); Cancers of the digestive organs and peritoneum (ICD-9: 150–159); Cancers of the respiratory and intrathoracic organs (ICD-9: 160–165) Cancers of the breast (ICD-9: 174–175); Cancers of the genital organs (ICD-9: 179–187); Cancers of the urinary organs (ICD-9: 188–189); Cancers of all other and unspecified sites (ICD-9: 170–173, 190–199); Leukemia (ICD-9: 204–208); and all other cancers of the lymphatic and hematopoietic tissues (ICD-9: 200–203)
Age	$a = 1, \ldots, 13$ Levels Ages under 1; Ages 1–4; Ages 5–9; Ages 10–14; Ages 15–19; Ages 20–24; Ages 25–34; Ages 35–44; Ages 45–54; Ages 55–64; Ages 65–74; Ages 75–84; and Ages 85 and older
Race	$r = 1, \ldots, 3$ Levels (Black, White, and Other)
Sex	$s = 1, 2$ Levels (Male and Female)

In total, there were $y_{\cdot} = \sum_{icars} y_{icars} = 26{,}116$ cancer-related deaths in PA in 1980 belonging to these $67 \times 13 \times 9 \times 3 \times 2 = 47{,}034$ strata.

- Over 42,000 of the death counts are zero

How Good is our Prior Information?

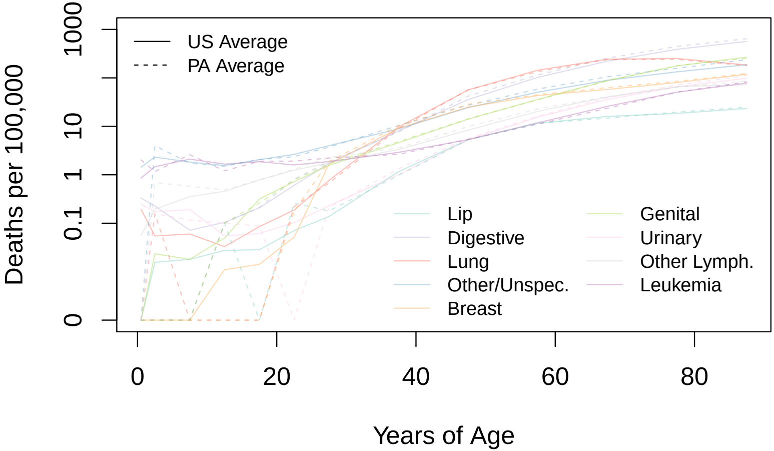
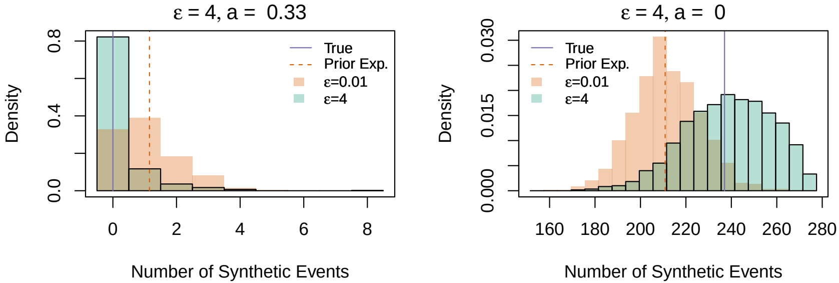


Figure 1: Cause-specific death rates at the national level and for the state of Pennsylvania. National-level rates are used as prior information for estimating the proper allocation of deaths at the state and county level.

What do the Poisson-Gamma Synthetic Data Look Like?



(a) Group with small y, $E\left(y \mid \mathbf{a}, \mathbf{b}\right)$

(b) Group with large y, $E\left(y \mid \mathbf{a}, \mathbf{b}\right)$

Figure 2: Posterior predictive distribution for various levels of ϵ. In Panel (a), the prior predictive expected value is $E\left[y \mid \mathbf{a}, \mathbf{b}\right] = 1.15$ and the true death count is $y = 0$. In Panel (b), the prior predictive expected value is $E\left[y \mid \mathbf{a}, \mathbf{b}\right] = 211$ and the true death count is $y = 237$.

- As $\epsilon \to 0$, the synthetic values shift away from y toward $E\left[y \mid \mathbf{a}, \mathbf{b}\right]$.

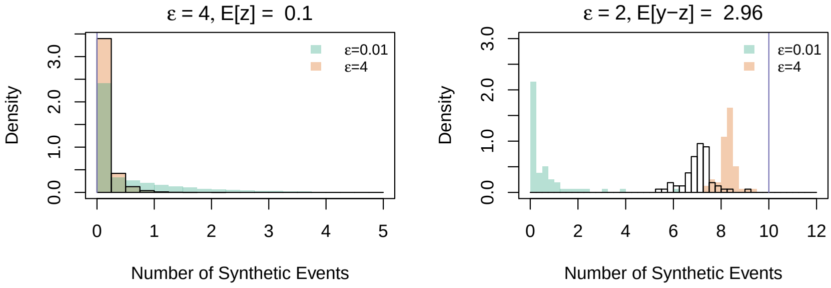
Hierarchical Strategy for Laplace mechanism

To do our comparison, we will consider:

- No hierarchy
- One-level hierarchy
  - Add noise to the [County], [Age], [Cause], [Race], or [Sex] specific totals
  - Most of the privacy budget allocated to top level of hierarchy
  - I plan(ned) to explore different privacy budget allocations, but I haven't yet :/
- Selected two-level hierarchies
- Selected three-level hierarchies
- Selected four-level hierarchies
  - Multilevel hierarchies selected for each inferential question (you'll see what I mean)

Note: All designs preserve the state-level total number of deaths (i.e., it's invariant).

What do the Laplace-Sanitized Values Look Like?



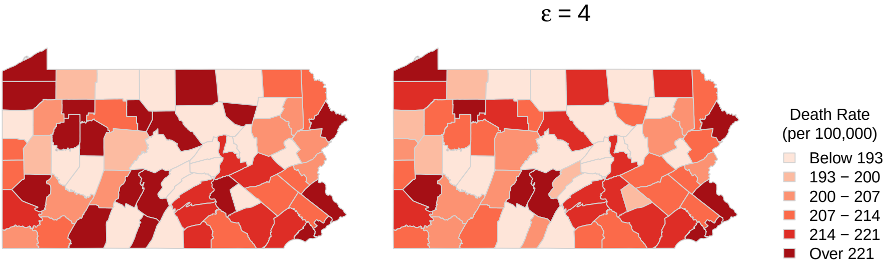
(a) $y = 0$, Various ϵ

(b) $\epsilon = 2$, Various y

Figure 3: Sampling distribution of the Laplace-sanitized values (with no hierarchical structure). Panel (a) shows the distribution of values when is $y = 0$ for various levels of ϵ, while Panel (b) shows the distribution of values for $\epsilon = 2$ for $y \in [0, 10]$.

- As $\epsilon \to 0$, the bias for $E\left[z \mid y = 0\right]$ increases
- Because of the disproportionate number of zeros in our dataset, all other values of $E\left[z \mid y\right]$ are also biased

Age-Adjusted Cancer Death Rates — Poisson-Gamma



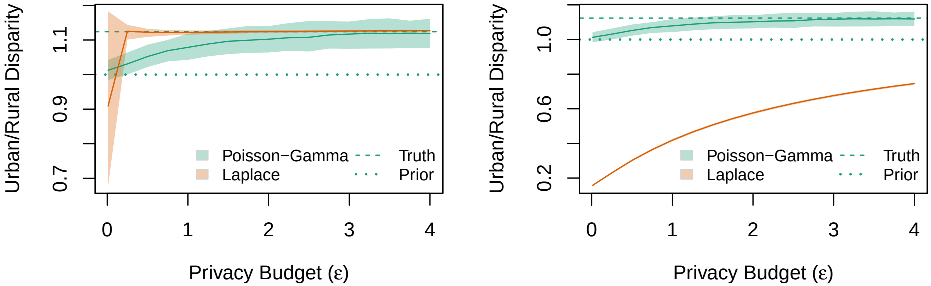
(a) True Age-Adjusted Rates

(b) Synthetic Age-Adjusted Rates

Figure 4: Degradation in utility for the age-adjusted rates as ϵ decreases.

- For large ϵ, geographic disparities in the data are largely preserved
- As $\epsilon \to 0$, the prior — which does not account for geographic disparities — becomes more influential and the rates all converge toward the statewide average

Urban/Rural Disparities in Cancer Death Rates



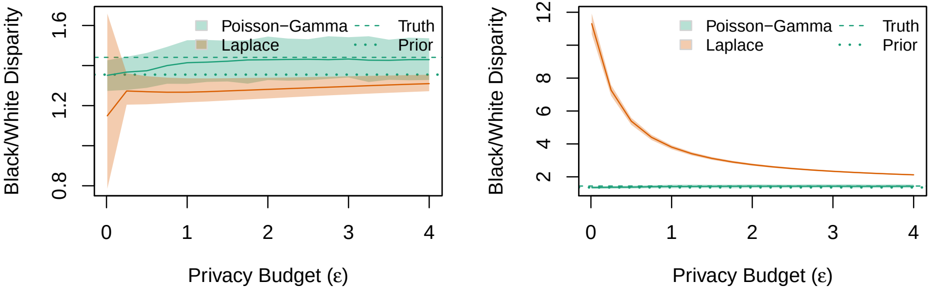
(a) “Best” Case Scenario

(b) Other Scenarios

Figure 5: Estimated urban/rural disparities. Values based on the true data (dashed lines) and the prior information (dotted lines) are provided for reference, while the shaded bounds represent the variability of the synthetic data.

- Poisson-gamma approach generally provides estimates between truth and the prior
- Laplace performs best when county-level totals are targeted, otherwise... :/

Black/White Disparities in Cancer Death Rates Among Males



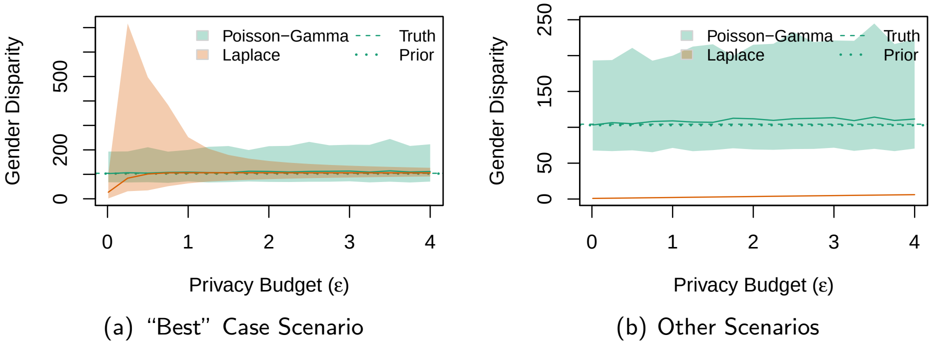
(a) “Best” Case Scenario

(b) Other Scenarios

Figure 6: Estimated black/white disparities. Values based on the true data (dashed lines) and the prior information (dotted lines) are provided for reference, while the shaded bounds represent the variability of the synthetic data.

- Poisson-gamma approach generally provides estimates between truth and the prior
- Laplace performs decently when race+age totals are targeted, otherwise it’s mixed or bad

Gender Disparities in Breast Cancer Death Rates



(a) “Best” Case Scenario

(b) Other Scenarios

Figure 7: Estimated female/male disparity in breast cancer death rates. Values based on the true data (dashed lines) and the prior information (dotted lines) are provided for reference, while the shaded bounds represent the variability of the synthetic data.

- Poisson-gamma approach generally provides estimates between truth and the prior
- Laplace performs decently when cause+sex+age totals are targeted, otherwise it’s BAD

Table of Contents

Motivating Use-Case: CDC WONDER

Generating Differentially Private Synthetic Data

Illustrative Example: Cancer Mortality in Pennsylvania

Summary

Summary

Based on this work, I claim:

- Using additive noise / "output perturbation" approaches to satisfy differential privacy — while convenient — will be inferior to sampling from the true data generating process
    - While hierarchical designs can help preserve inference on certain quantities (e.g., urban/rural disparities), it quickly devolves into a game of whack-a-mole as improving inference on one quantity will likely erode inference on many others.
- Samples from a posterior predictive distribution that aims to *approximate* the true data generating process can still perform well
    - Good prior information can keep your synthetic values in the right ballpark
    - Minimizing the informativeness of the priors allows the data to dictate what happens