Differential Privacy and Census Data

Walter Schwarm, Jonathan Buttle
California Department of Finance
Demographic Research Unit

Why Differential Privacy?

- Title 13 requires the Census Bureau to ensure that responses to surveys remain confidential and no publication allow for the identification of any establishment or individual;

- Based on simulations and testing, the Census Bureau determined that data protection techniques used in prior Censuses were no longer sufficient to meet Title 13 confidentially requirements.

What is Differential Privacy?

- Differential Privacy (DP) is a mathematical technique that allows for the formal quantification of the risk of data disclosure;
- Formally, DP is a property of algorithms for answering queries. An algorithm is considered differentially-private for a given epsilon (ε) if, for two databases that differ by one record, it satisfies:

$$\Pr[A(D) \in T] \leq \exp(\varepsilon) \Pr[A(D') \in T]$$

- If the algorithm satisfies this definition, the expression provides a bound on how much information can be inferred from adding or deleting a record in the database and prevents learning about a specific record by examining two datasets.

What is Differential Privacy (con’t)

- As a result, DP allows for mathematically quantifying the risk of identifying a specific element in a dataset;

- Specifically, differentially private algorithms provide formal bounds as to how many queries can be made before the probability of learning specific information about a database increases beyond acceptable levels.

The Components of Differential Privacy

- The privacy loss budget. The privacy loss budget is typically represented by epsilon (ε).
- When $\varepsilon = 0$, the resulting data would be random and essentially useless (perfect privacy).
- When $\varepsilon = \infty$, the resulting data would allow for full identification of survey participants (perfect accuracy).
- Values of epsilon between $0 \text{ and } \infty$ represent a trade off between privacy and accuracy.

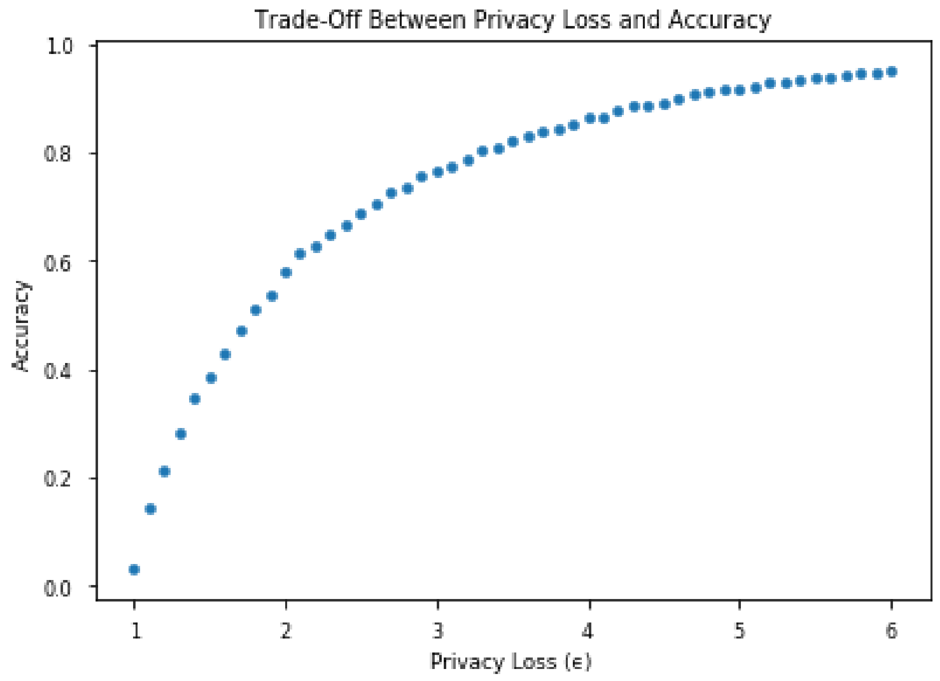
The Privacy Budget

- An alternative interpretation of epsilon is that of a “privacy budget”.
- If only a single query on the data is expected to be performed, that query might use up the entirety of the budget;
- However, performing a series of queries on the data requires allocation of the budget over all the queries;
- There are two methods of allocating the privacy budget – sequential and parallel.

The Privacy-Accuracy Tradeoff

- This graph illustrates the privacy-accuracy trade off for a privacy mechanism with epsilon values between 1 and 6:

Accuracy is defined as $1 - [\frac{\sum abs(obs_n - obs_o)}{length(obs_o)}]$



The DP Mechanism

- The DP mechanism works by injecting statistically calibrated ”noise” into the data;
- The amount of noise injected is determined by epsilon and by sensitivity – sensitivity being the amount one or more individuals can influence the output of the mechanism;
- Statistical “noise” is typically derived from two distributions:
  - ➢ The Laplace distribution, or the
  - ➢ The Geometric distribution;
- The geometric distribution has the advantage of returning integer values, while the Laplace distribution does not, and so the geometric mechanism has been employed in the Census Bureau’s DP engines.

Sequential Composition

- Sequential composition is where information from a database is released on an overlapping set of individuals;
- Example – a query to generate the population total for a county and a separate query generating the total by age group for that same county;
- In this case, the total privacy budget is the sum of the privacy budgets for the overlapping queries;
- In other words, the analyst must account for all the operations performed on the data to ensure the global privacy for the dataset.

Parallel Composition

- Parallel composition is where a series of queries on a database release information on a disjoint set of individuals;
- Example – a query generates the number of persons in all counties in one county while another query returns the number of persons by age category who reside in a second county;
- The total privacy budget would be the max of the individual query budgets;

Post-Processing

- One important characteristic of DP is that once a dataset has been privatized through a DP algorithm, processing on the privatized dataset maintains the differential privacy;
- Therefore, additional data processing can address issues such as:
  - ➢ Counts less than zero;
  - ➢ Ensuring the sum of counts for lower geographies are equal to counts for higher geographies (i.e. the sum of the counts for all counties in a state equal the total count for the state).

Census Bureau and DP

- Early implementation
  - 2008 – OnTheMap/LEHD
- Post-Secondary Employment Outcomes
  - Earnings Distributions
- 2020 Census

DP and the 2020 Census

- Original test implementation – 1940 Census Dataset
  - Top-Down Methodology;
  - Creates a histogram of demographic attributes (total population, voting age, race/ethnicity, group quarters type);
  - Assigns them iteratively to various geographies (nation, state, county, enumeration district);
  - Applies ‘noise’ to the attributes by adding results from random number generator to the attribute counts;
  - Post-process the resulting noisy data subject to ‘invariants’ – total population at the state level and total housing unit and group quarters counts at the block level.

DP and the 2020 Census (con't)

- 1940 Census Dataset
  - The Census Bureau released the source code (in python) and the 1940 Census dataset was made available through IPUMS;
  - The Census Bureau also released a series of DP runs for various epsilon levels (0.25, 0.5, 0.75, 1, 2, 4, and 6);
- Analysis of the results
  - Low privacy loss budget (epsilon) – 0.25 – resulted in significant distortions in smaller geographic areas and attributes such as race/ethnicity relative to original data;

DP and the 2020 Census (con't)

- 2010 Demonstration Data Products Disclosure Avoidance System (DAS) release -
  - Updated DP applied to the Census Edited File used in the 2010 Census to generate person and housing tables from the PL94 and SF1;
  - DP process employed a global epsilon of 6.0 – 4.0 allocated to person tables and 2.0 allocated to housing tables;
  - Geographies expanded to include tract groups, tracts, block groups and blocks;
  - Tables expanded to include age by groupings by sex and households by race/ethnicity, sex, and presence of persons age 60 plus;

DP and the 2020 Census (con't)

- Analysis of the resulting tables found:
  - Transfer of population counts from larger geographic areas to smaller geographic areas as a result of invariants and post-processing error;
  - Significant distortions in demographic categories such as 5-year age groups;
  - Distortions in population counts for American Indian and Alaska Native Tribal areas and in off-spline geographic areas;
  - Distortions in housing statistics (vacant and occupied housing units) and persons per household ratios.

DP and the 2020 Census (con't)

- The Census Bureau identified the following issues:
  - ➢ Measurement error due to DP noise;
  - ➢ Post-processing error from creating internally consistent, non-negative integer counts from noisy measurements;
  - ➢ Of those errors, post-processing errors tend to be larger than DP error;

DP and the 2020 Census (con't)

- How Census plans to address these issues:
  - ➢ Select epsilon to reduce measurement error while maintaining privacy;
  - ➢ Adopt a revised post-processing mechanism –
    - o Multi-pass hieratical post-processing
  - ➢ Updated DAS development cycle consisting of 4-week development sprints followed by 2-week evaluation windows;
  - ➢ Revised accuracy metrics released periodically to coincide with evaluation windows;

Demonstration Products – Metrics Tables

- Starting in March 2020, Census released updated metrics designed to use cases and stakeholder feedback;
- The purpose is to allow users/stakeholders to see improvements from changes to the DAS mechanism;
- The metrics will include measures of accuracy, bias, and outliers;
- Census plans to add AIAN and off-spline geographies, and to improve race metrics and outlier measures;

Demonstration Products – Metrics Tables

- Measures of accuracy.
  - Accuracy is measured by comparing the post-disclosure protected tabulations to the original, publicly available tabulations from the 2010 Census and the internal pre-disclosure avoidance microdata from the 2010 Census.
- Proposed accuracy measures include –
  - Mean/Median Absolute Error (MAE);
  - Mean/Median Numeric Error (ME) ;
  - Root Mean Squared Error (RMSE);
  - Mean/Median Absolute Percent Error (MAPE); and
  - Coefficient of Variation (CV)

Demonstration Products – Metrics Tables

- Measures of bias.
  - Related to accuracy, but bias measures the direction of change and whether it varies with population size or some other characteristic.
- Proposed bias measures include –
  - Mean/Median Numeric Error (ME); and
  - Mean/Median Percent Error (MALPE)

Demonstration Products – Metrics Tables

- Sample metrics table with measures of accuracy, bias, and outliers (5/27/2020 compared with the 3/25/2020 release):

Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers – 5/27/2020 release
Universe: Total population
Geography: Summary Level 050 - State-County

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of counties where the absolute percent difference is 5% to 10%	Count of counties where the absolute percent difference exceeds 10%
All counties	3,143	15.95	21.15	0.14	0.02	0.02	2	2
Counties with total population less than 1,000	35	13.51	17.19	2.72	2.50	(0.03)	2	2
Counties with total population 1,000 to 4,999	268	14.40	19.42	0.52	0.64	0.14	-	-
Counties with total population 5,000 to 9,999	395	15.51	20.72	0.21	0.28	0.07	-	-

Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers – 3/25/2020 release
Universe: Total population
Geography: Summary Level 050 - State-County

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV	MALPE (%)	Count of counties where the absolute percent difference is 5% to 10%	Count of counties where the absolute percent difference exceeds 10%
All counties	3,143	82.18	141.39	0.78	0.14	0.69	31	17
Counties with total population less than 1,000	35	76.49	128.60	28.49	18.71	28.35	13	13
Counties with total population 1,000 to 4,999	268	62.11	74.27	2.35	2.43	2.31	18	4
Counties with total population 5,000 to 9,999	395	58.77	71.60	0.81	0.95	0.75	-	-
Counties with total population 10,000 to 49,999	1,469	58.53	73.59	0.29	0.29	0.20	-	-
Counties with total population 50,000 to 99,999	398	63.99	86.08	0.09	0.12	(0.03)	-	-
Counties with total population of 100,000 or more	578	180.45	287.70	0.07	0.07	(0.06)	-	-

Questions/Discussion

Contact Information

- Walter Schwarm - walter.schwarm@dof.ca.gov
- Jonathan Buttle – jonathan.buttle@dof.ca.gov

- California Department of Finance
- Demographic Research Unit
- (916) 323-4086