Differential Privacy in Travel Data

MPI Camp 2021

Professors: Manuchehr Aminian, Sunil K. Dhar, Tobin A. Driscoll, David A. Edwards, Brooks Emerick, Pak-Wing Fok, Andrew O. Hall, Taras Lakoba, Karl Wimmer

Grad Students: Xing Fan, Leah Gibson, Elizabeth Grimes, Gess Iraji, Carlos Rojas Mena, Zhen Shao, Yuqi Su, Sheng Wang, Andrea Weaver, Lingyi Yang

University of Delaware, Newark

June 18th, 2021

Overview

- Introduction
- Objective

Coin Flip Perturbation:

- Differential Privacy
- Vehicle Perturbation
- Who Drives That Lotus

Census Block Perturbation:

- Uniform Block Perturbation
- Non-Uniform Block Perturbation
- Tiling of Polygons
- References
- Closing Remarks/Questions

Household Travel Study

Household Travel Study (HTS) are collections of demographic data and detailed travel diary data from household members[1].

Common data collected:

- Household - income, age, relationships, race/ethnicity
- Important locations - home, work, school
- Reasons for trips, who you traveled with
- Trip Traces - extra stops along the way

This data is commissioned by public agencies for such things as traffic light scheduling, road redesign and other transportation efficiency/optimization.

[1] Cited from [LDGE19]

Data Collection

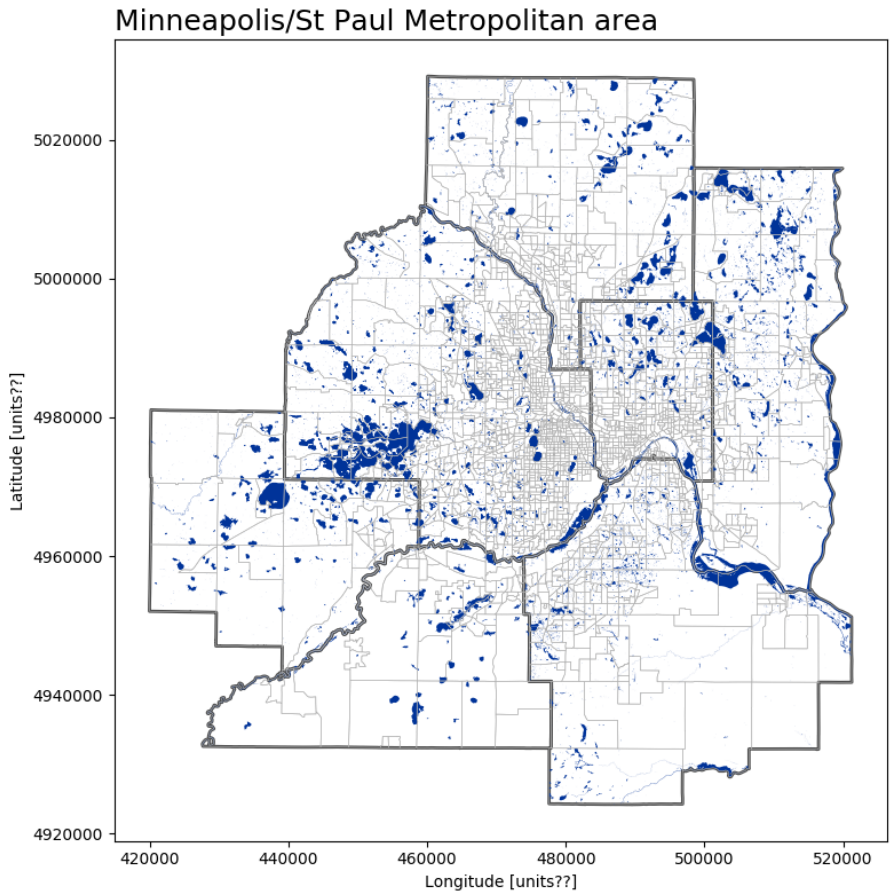
Three methods of data collection:

- Smartphone Application
- Call Center
- Online Application

Participants surveyed in 2019 [2]:

Total: 16,152
App: 11,405; Call/Online: 4,747

Divided into subcategories:
days, households, locations,
persons, trips, and vehicles.



[2] All data collected from [LDGE19]

Category Name	Explanation	Number of Observations
Day	Basic information with dates of surveys, number of trips, and reasons.	84,562
Household	Information on type of residence, renting or owning status, income, and duration of time at residence.	7,837
Location	Latitude and longitude of travel destinations.	1,048,575
Person	Demographic information and data collection type (smartphone app on online/phone).	16,152
Trip	More detailed information on trips taken throughout each day.	240,449
Vehicle	Information about vehicle type.	13,432

Figure: Summary of Data

Data Exploration

Like mentioned before, the data collected can tell a variety of information. Consider the ride-hailing service trips (1244 trips)[3].

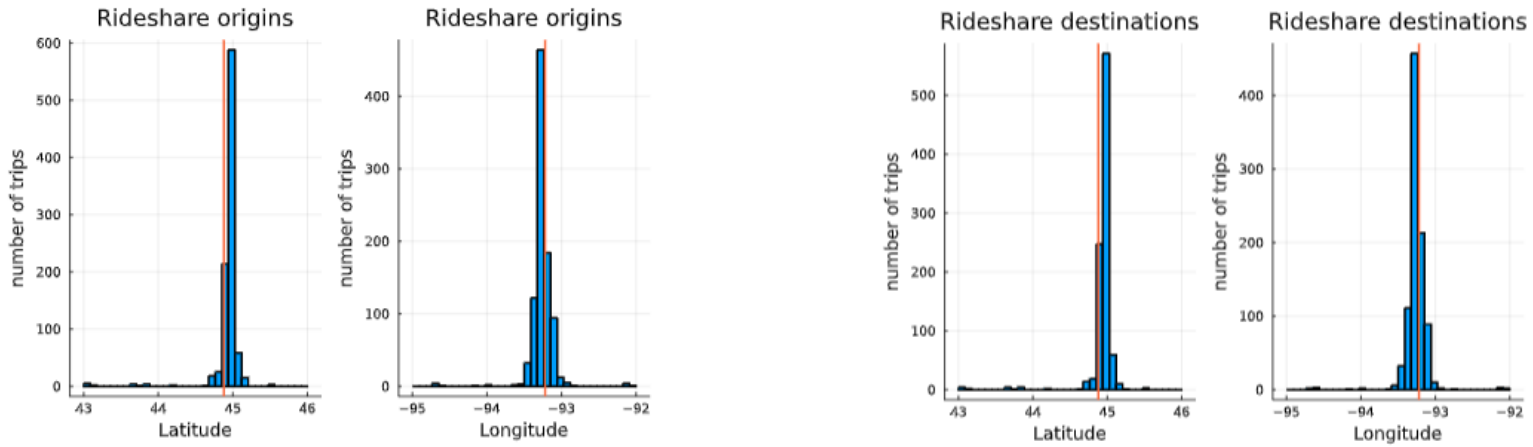


Figure: Latitude/Longitude of Ride-Share

[3]Looked at a large box enclosure, outliers were removed.

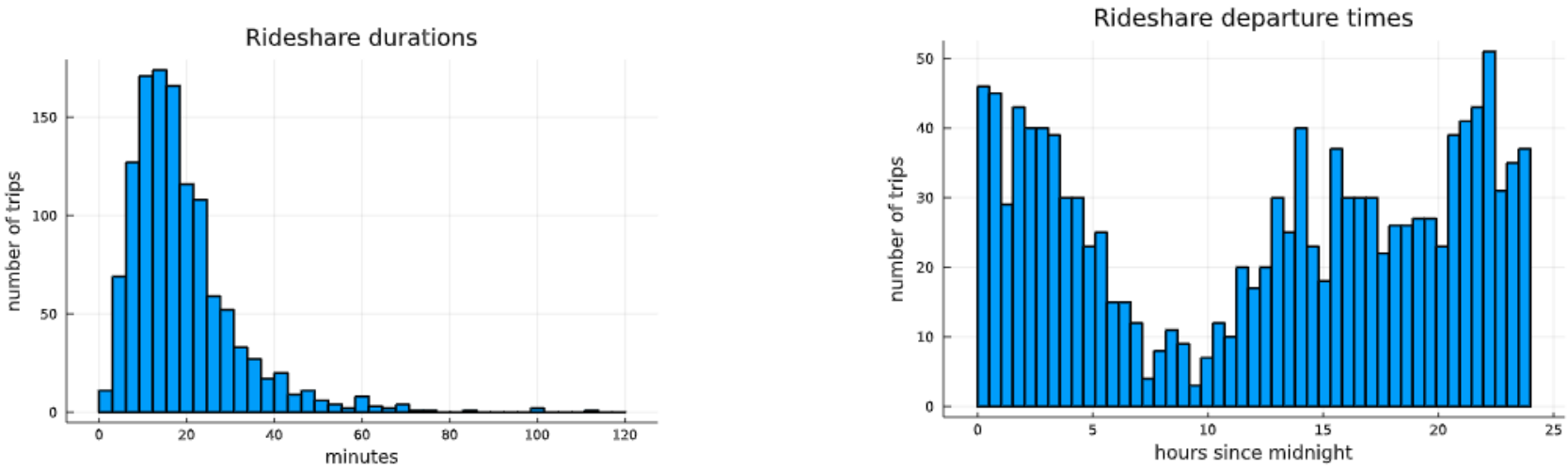


Figure: Extra Data on Ride-Share

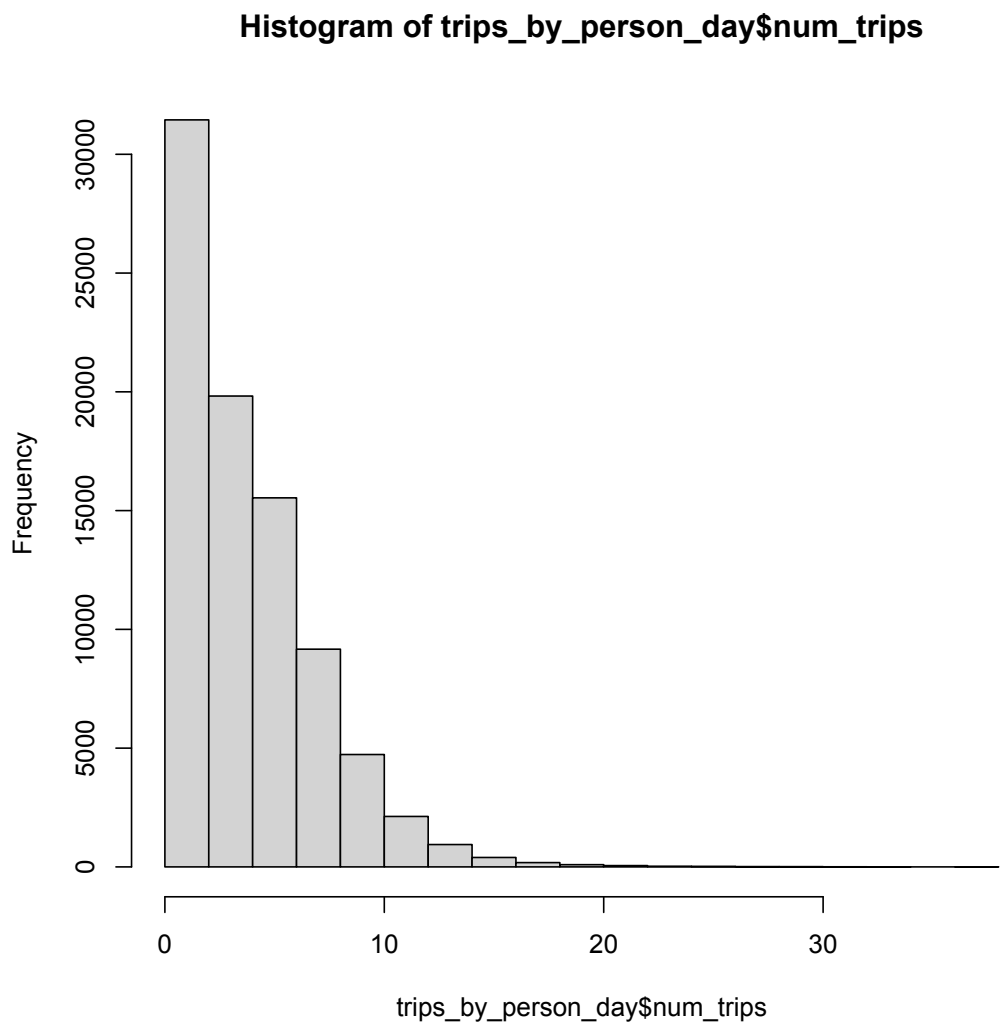


Figure: Histogram of the number of trips made by each person per day

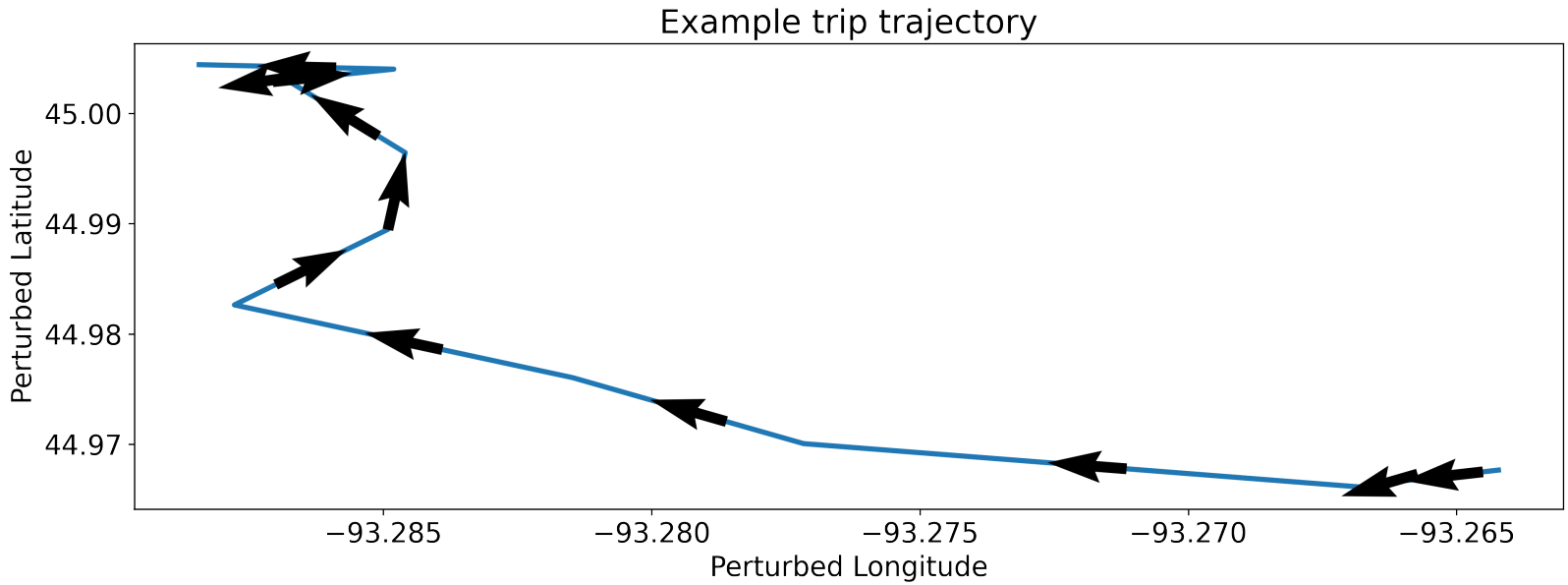
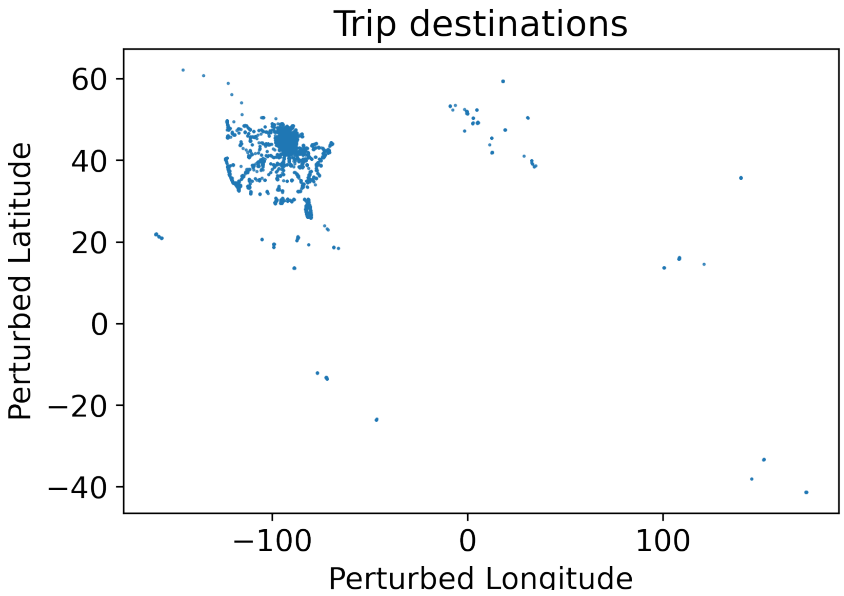


Figure: An example trajectory



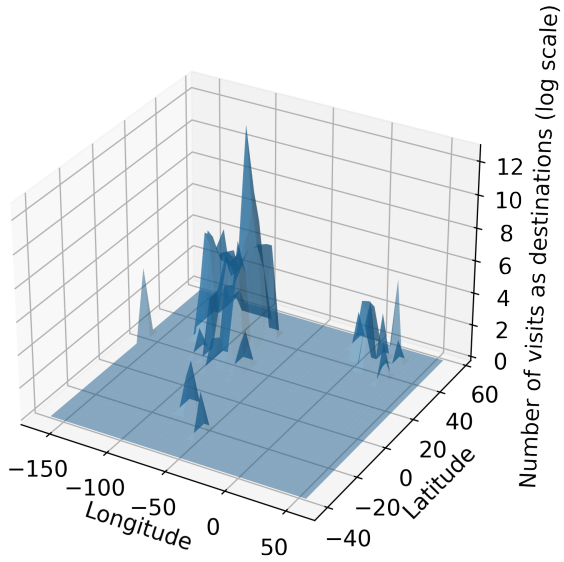


Figure: Visualization of the destination distributions

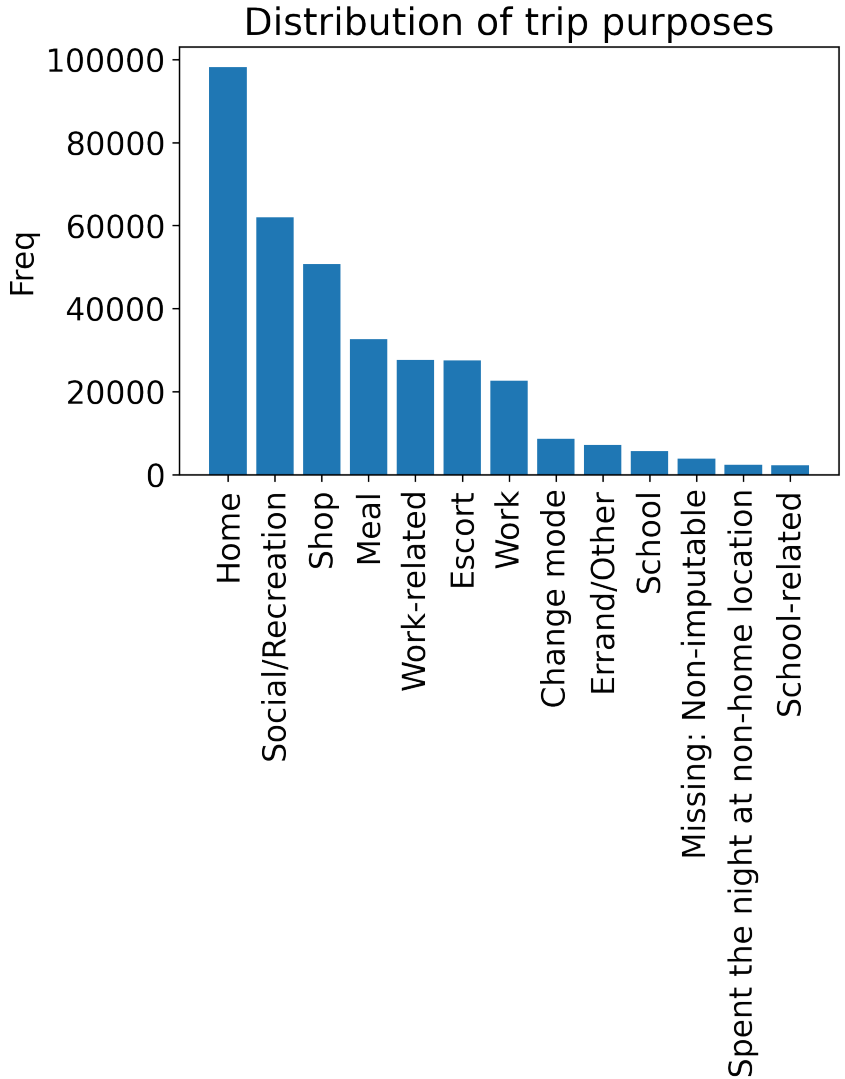


Figure: Histogram of the reasons for making trips

Our Objective

Privacy is the Goal!

Famous example of privacy concern: Netflix simple anonymized data was not enough to protect its participants[4].

[4] For more information: [NS06]

Differential Privacy

Differential Privacy is a method of perturbing data using a randomized algorithm that adds noise to ensure that any individual will not be positively or negatively impacted by their information being in the data[5].

- Students mean test scores is not differentially private.
- Coin flip example is differentially private. (Will see later!)

[5] For more detail: [DR14]

Formal Definition of Differential Privacy

Let M be a mechanism, for two neighboring data sets D_1 and D_2, and S is a subset of the image of M we get:

Theorem (Differential Privacy)

$$\frac{Pr[M(D_1) \in S]}{Pr[M(D_2) \in S]} \leq e^{\epsilon}$$

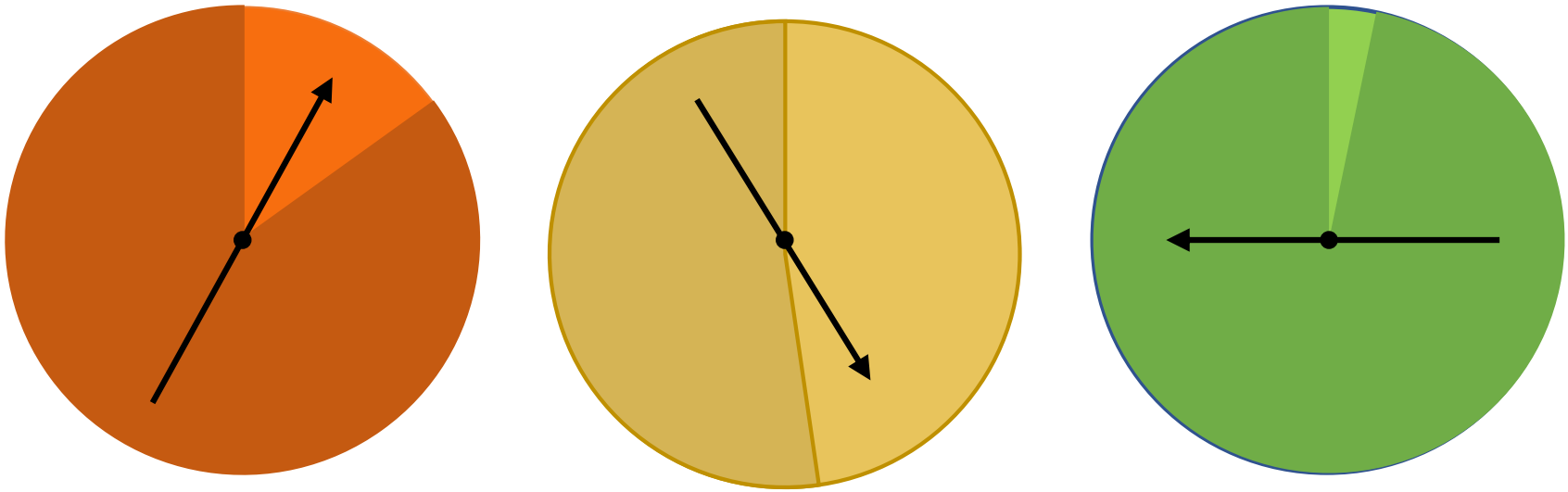
Important Note: Based on numeric queries which are maps f from data bases to real numbers. (Example - sum or mean)

Data Types

data type	categorical	numerical
example	vehicle model	latitude and longitude data
method	randomized response	perturb and aggregate

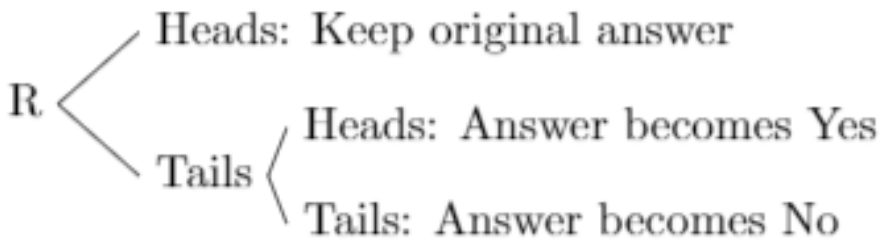
Coin Flip Perturbation

Randomized Responses



Coin Flip Perturbation

Suppose participants are asked "Do you like math?" To protect the privacy of everyone we are going to introduce a coin-flip perturbation.



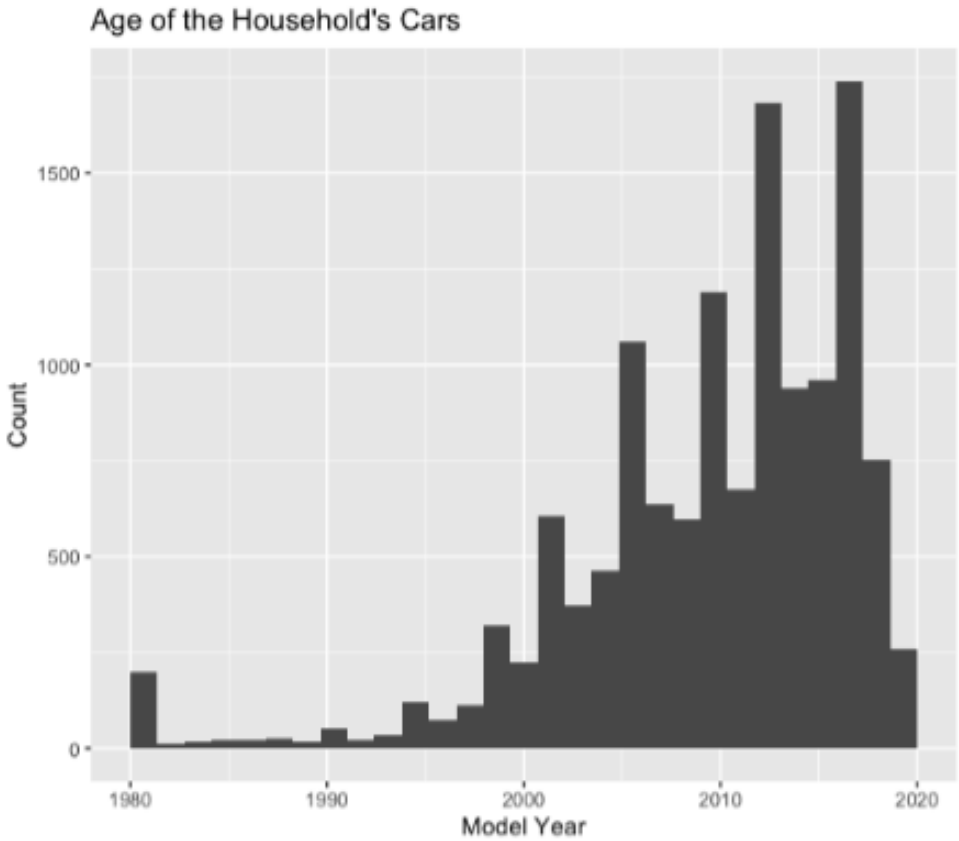
Is this method DP?

$$\frac{Pr[R = \text{"yes"} \,|\text{Truth} = \text{"yes"}]}{Pr[R = \text{"yes"} \,|\text{Truth} = \text{"no"}]} = \frac{3/4}{1/4} = 3 \leq e^{\epsilon}$$

Thus the coin flip algorithm is $\ln(3)$-differentially private.

Vehicle Data Summary

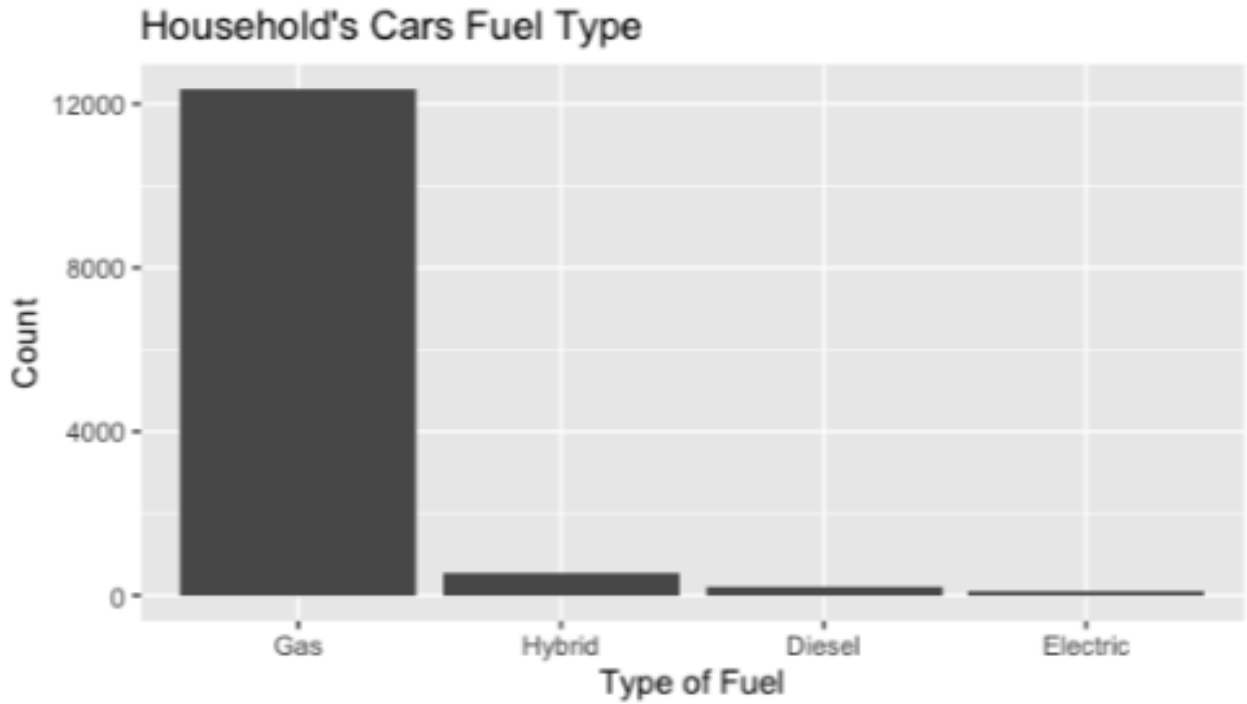
Total number of Vehicles: 13, 431[6]



[6]Data Cleaning Applied

Objective/Motivation

Perturb the data to increase privacy while keeping the accuracy of the vehicle fuel type, using the coin flip technique.



Vehicle Perturbation Comparison

Using the following coin flip method:

- 6,772 (50.42%) vehicle names remained the same
- After reassignment: 6,788 (50.54%) vehicle names are the same
- 16 vehicles were randomly assigned to the same vehicle: 9-gas,4-hybrid, and 3-electric

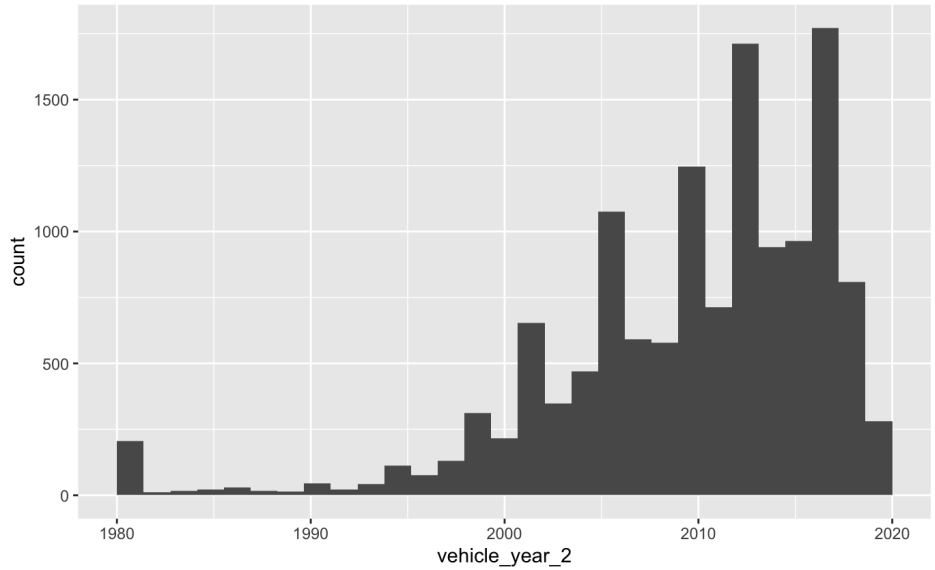
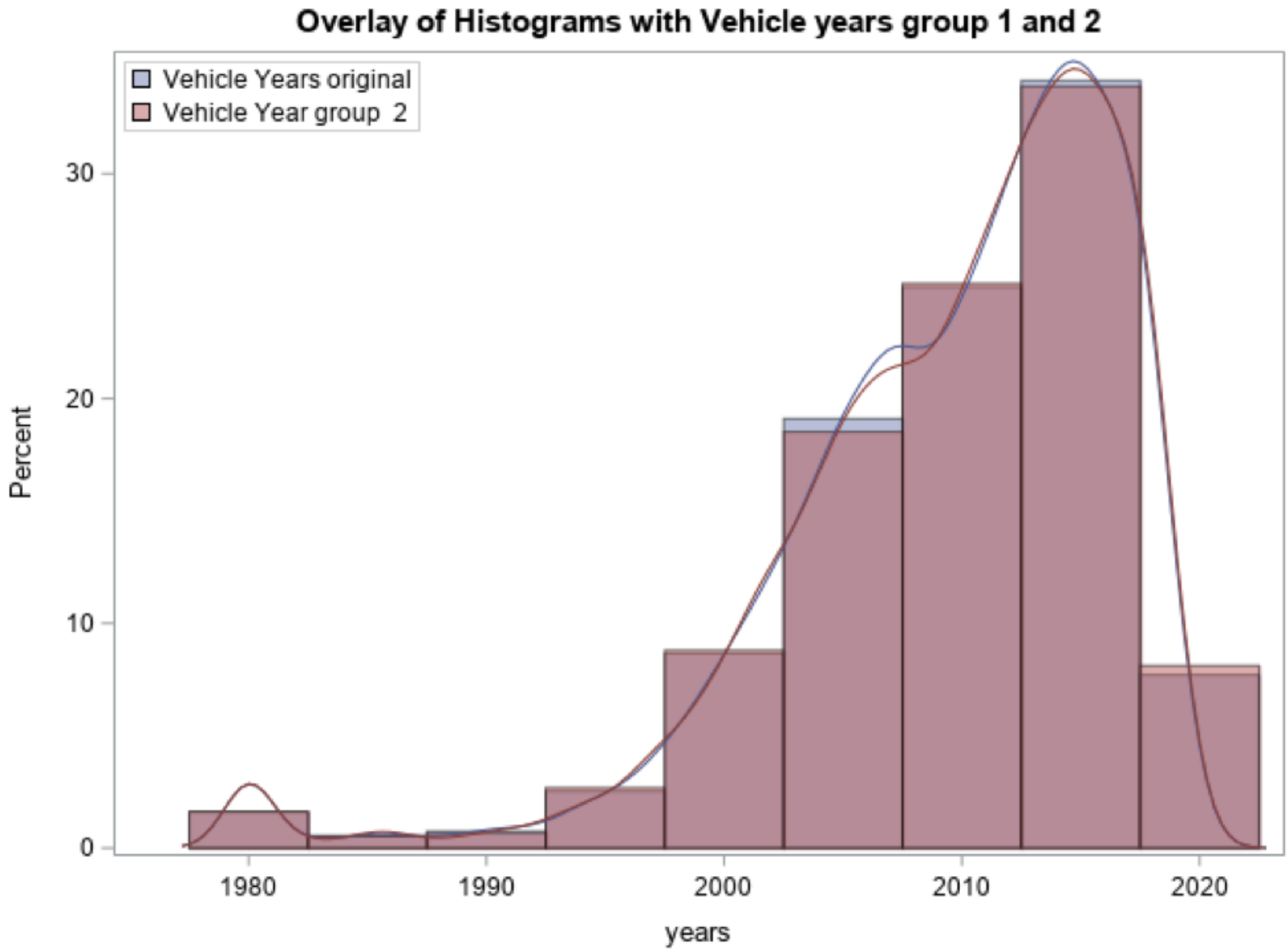


Figure: Distribution of Perturbed Vehicle Fuel Types


Overlay of Histograms with Vehicle years group 1 and 2
Vehicle Years original
Vehicle Year group 2
Percent
0
10
20
30
1980
1990
2000
2010
2020
years

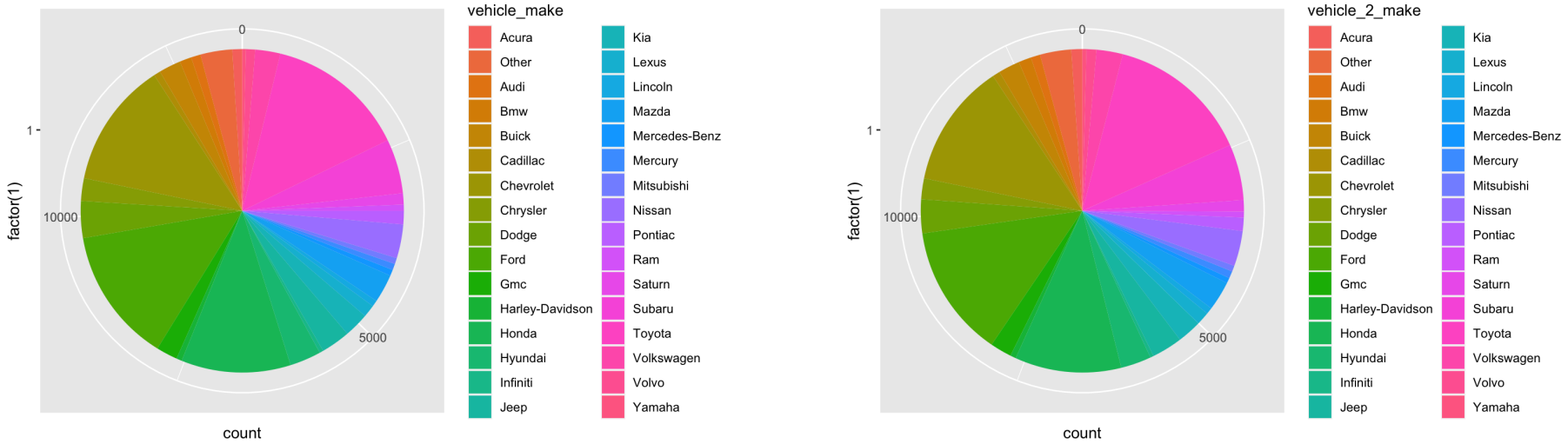
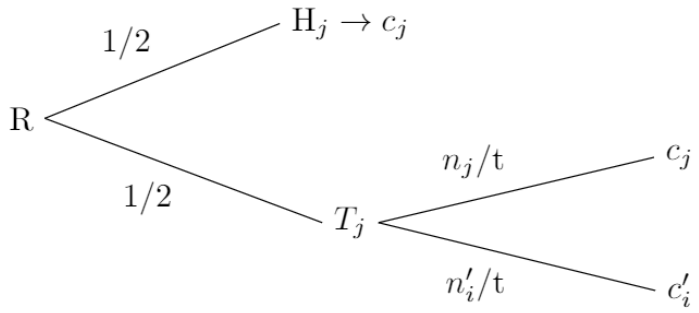


Figure: Comparison of Vehicle Makes Between Original and Perturbed

Modified Coin Flip Algorithm

Differential privacy to subcategories of the vehicle data.



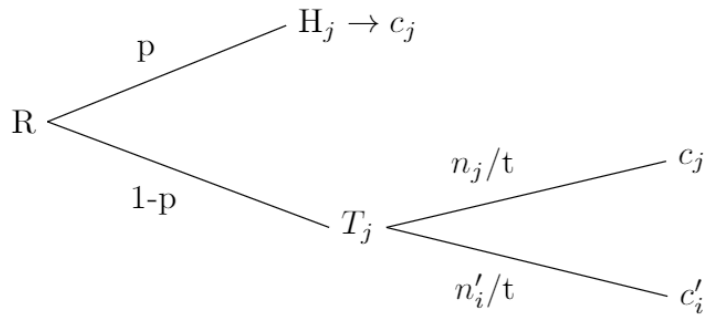
$$\max\left\{\frac{\Pr[R=c_j|\ \text{Truth}=c_j]}{\Pr[R=c_j|\ \text{Truth}=c'_i]}\right\} = \max\left\{\frac{\frac{1}{2}+\frac{n_j}{2t}}{\frac{1}{2}\cdot\frac{n_j}{t}}\right\}$$

$$= \max\left\{\frac{t+n_j}{n_j}\right\}$$

$$= \frac{t+n}{n}$$

Generalized Probability Algorithm

Generalization of modified coin flip algorithm for any desired probability.

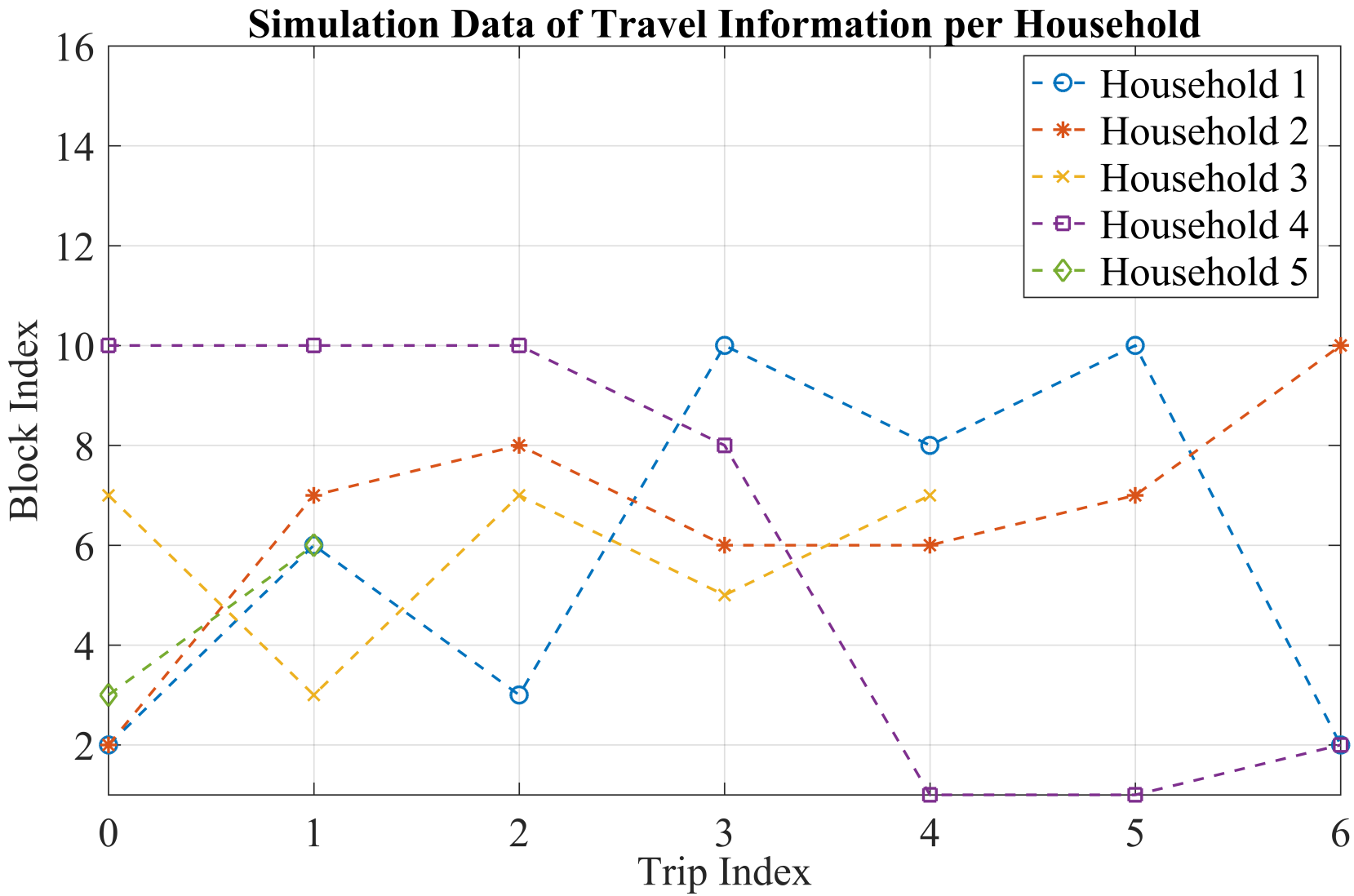


$$\max\left\{\frac{\Pr[R = c_j \text{ Truth} = c_j]}{\Pr[R = c_j | \text{ Truth} = c_i']}\right\} = \max\left\{\frac{p + (1-p)\frac{n_j}{t}}{(1-p)\cdot\frac{n_j}{t}}\right\}$$

$$= \max\left\{\frac{pt + (1-p)n_j}{(1-p)n_j}\right\}$$

$$= \frac{pt + (1-p)n}{(1-p)n}$$

Data Simulation

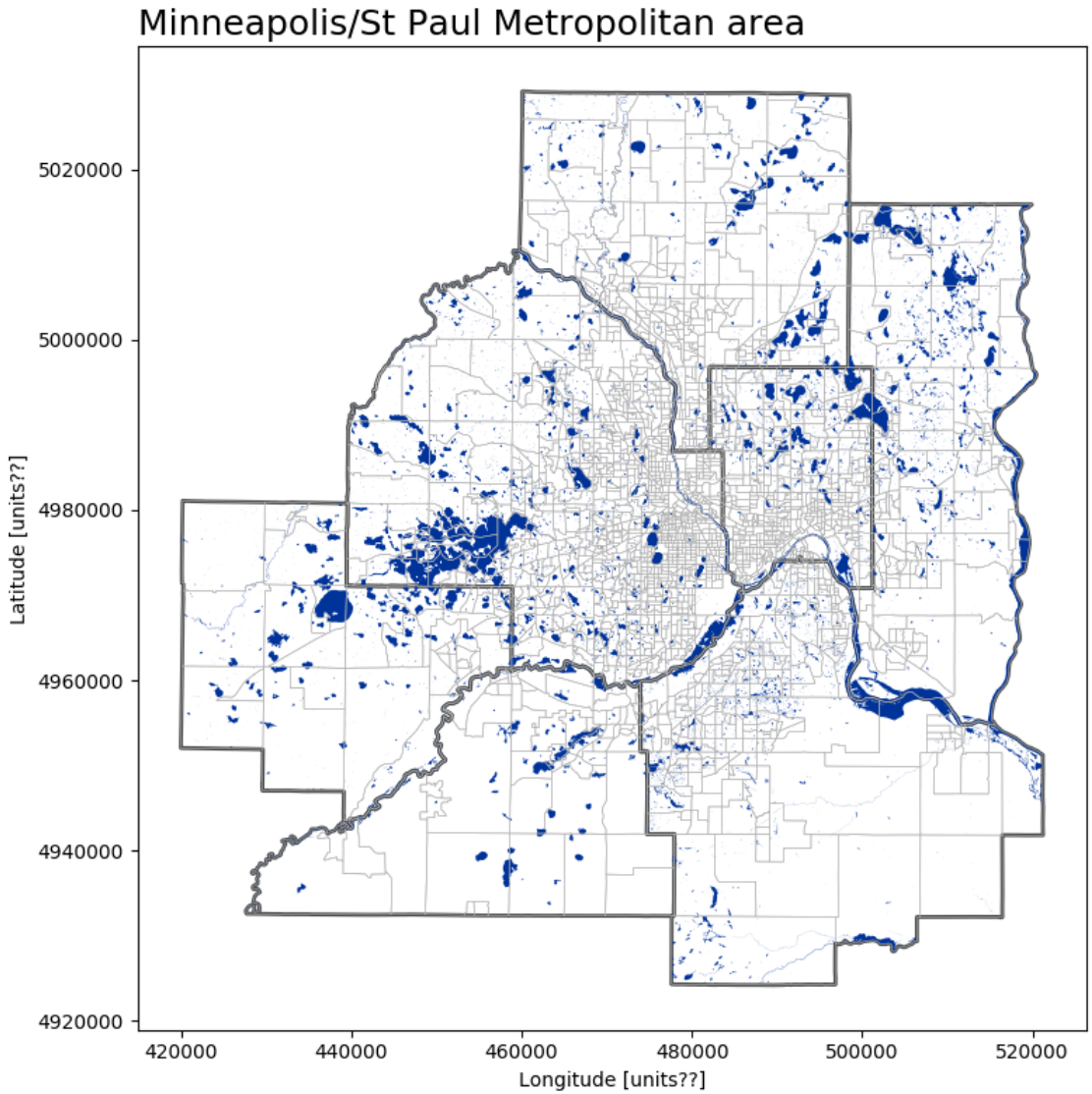


Block Perturbation

Data Perturbation and Privacy

- Data perturbation and aggregation is a common approach to provide privacy.
- $y = x + n$ where x is original data and n is random noise.
- When data is published it is often published in aggregated form e.g. data is posted about census blocks not individuals.
- Aggregation provides some protection. Data perturbation adds to this protection.

Census Block



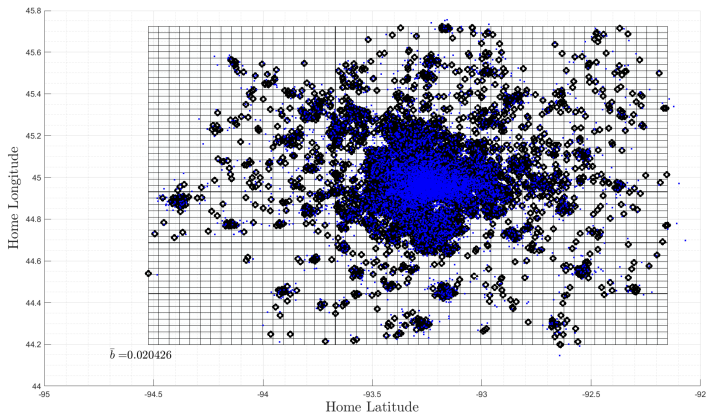
Modeling The Effects of Data Perturbation

- Trade-off between the protection and the inaccuracy data perturbation provides.
- If data is highly perturbed, then any statistic on a particular entity cannot be deduced from published data. Data is useless but highly private.
- If location data is perturbed and then labeled according to its census blocks we want to guarantee that after perturbation some data is mislabeled. Otherwise, perturbation has no effect on the published data.
- We modeled the proportion of data points that change census block after perturbation.

Modeling the effects of data perturbation

- We can formulate this problem by considering a polygon (census block) with (location) points uniformly distributed inside. After adding noise, we ask what proportion of the points escaped the polygon.
- Developed program to compute this quantity in a simulation for a uniform and non-uniform grid, and for a general polygon shape.

Escaped points in uniform grid



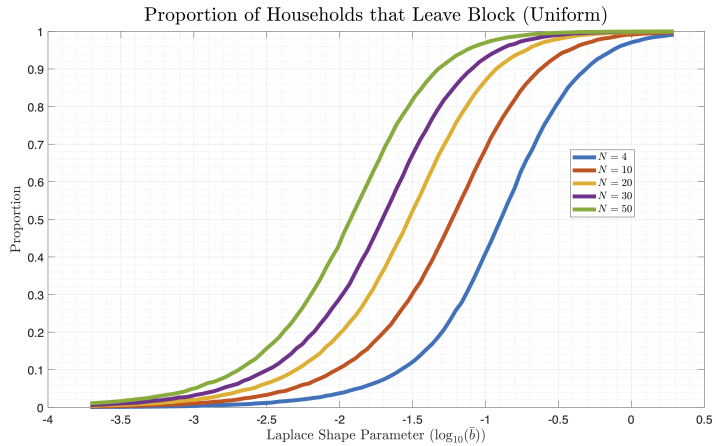
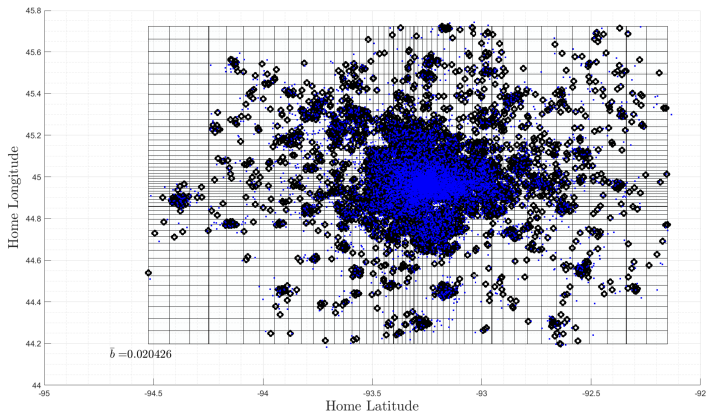


Figure: Left: Household location data overlaid on uniform grid. Black represents the actual data; the blue dots correspond to the perturbed data.
Right: Proportion of perturbed points that left original block versus $\log_{10}(b)$ where $2b^2$ is the variance of Laplacian distribution.

Escaped points non-uniform grid



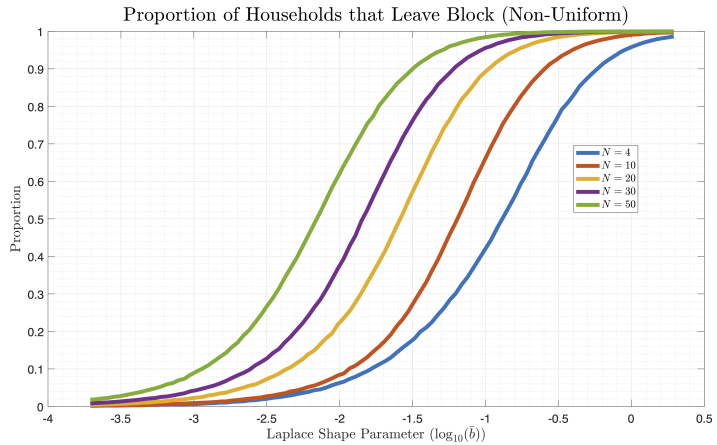


Figure: Left: Household location data overlaid on non-uniform grid. Center with high concentration of census blocks. Black represents the actual data; the blue dots correspond to the perturbed data.
Right: Proportion of perturbed points that left original block versus $\log_{10}(b)$ where $2b^2$ is the variance of Laplacian distribution.

Escaped points from general polygon

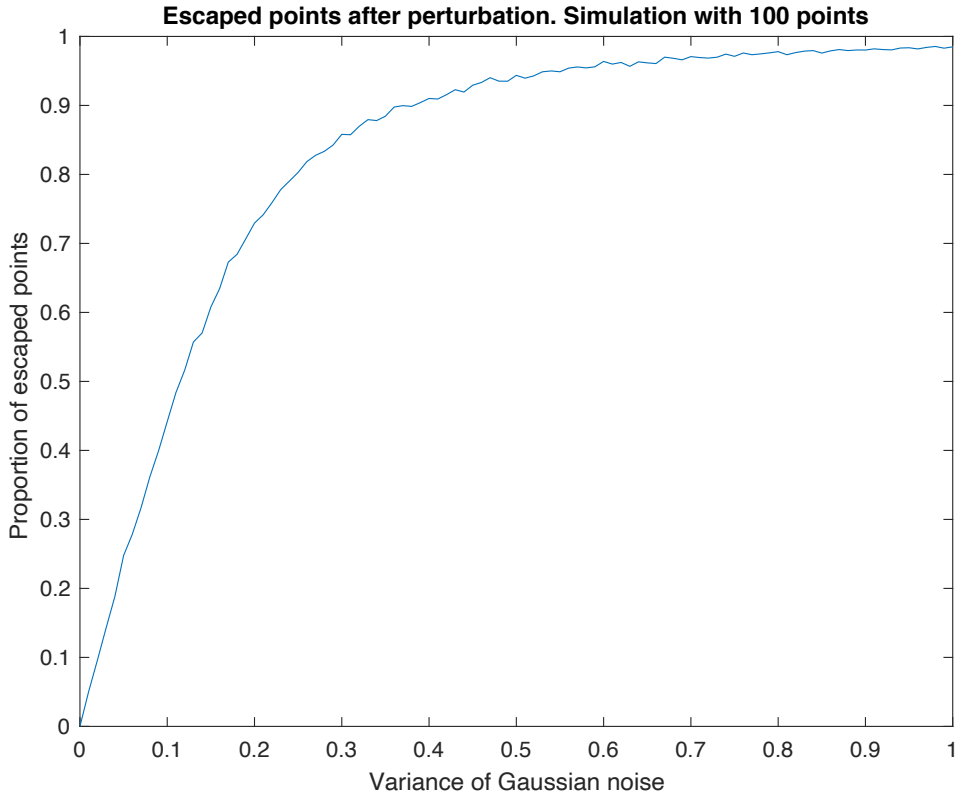


Figure: Proportion of escaped points in a pentagon vs. variance of Gaussian noise. Simulated with 100 uniformly distributed points inside pentagon.

From One Polygon to a Tiling of Polygons

Suppose we have a tiling of N polygon P_i each with R_i points inside. Let ρ_i be the escape proportion. It can be shown that the proportion of total points that leave their polygon after perturbation is

$$\frac{\sum_{i=1}^{N} R_i \rho_i}{\sum_{i=1}^{N} R_i}$$

Therefore, we can apply our algorithm for polygons on each census block independently to obtain the proportion of location points whose census block is changed.

Conclusions

A summary of our accomplishments are

- Analyzed the general structure of the data.
- Modified the coin flip to make vehicle data more private.
- Added perturbations to latitude and longitude data and checking for the proportion that left the original block.

Thank you for your time! Questions?

References


Cynthia Dwork and Aaron Roth.
The algorithmic foundations of differential privacy.
Foundations and Trends in Theoretical Computer Science, 9, 2014.


Joann Lynch, Jeffrey Dumont, Elizabeth Greene, and Jonathan Ehrlich.
Use of a smartphone gps application for recurrent travel behavior data collection.
Transportation research record, 2673(7):89–98, 2019.


Arvind Narayanan and Vitaly Shmatikov.
How to break anonymity of the netflix prize dataset.
CoRR, abs/cs/0610105, 2006.