MATHEMATICAL PROBLEMS IN INDUSTRY, 2021
DIFFERENTIAL PRIVACY IN TRAVEL DATA

INDUSTRIAL PRESENTER
JEFF DUMONT, RSG

ACADEMIC PARTICIPANTS
MANUCHEHR AMINIAN, CAL POLY POMONA
ANNE-SOPHIE CHAREST, UNIVERSITÉ LAVAL
SUNIL K. DHAR, NEW JERSEY INSTITUTE OF TECHNOLOGY
TOBIN A. DRISCOLL, UNIVERSITY OF DELAWARE
DAVID A. EDWARDS, UNIVERSITY OF DELAWARE
BROOKS EMERICK, KUTZTOWN UNIVERSITY
XING FAN, UNIVERSITY OF CENTRAL FLORIDA
PAK-WING FOK, UNIVERSITY OF DELAWARE
LEAH D. GIBSON, COLORADO STATE UNIVERSITY
ELIZABETH GRIMES, UNIVERSITY OF CENTRAL FLORIDA
ANDREW O. HALL, MARYMOUNT UNIVERSITY
GESS KELLY, BRANDEIS UNIVERSITY
TARAS LAKOBA, UNIVERSITY OF VERMONT
CARLOS ROJAS MENA, UNIVERSITY OF ROCHESTER
ZHEN SHAO, UNIVERSITY OF OXFORD
YUQI SU, NORTH CAROLINA STATE UNIVERSITY
SHENG WANG, UNIVERSITY OF OXFORD
ANDREA WEAVER, UNIVERSITY OF DELAWARE
KARL WIMMER, DUQUESNE UNIVERSITY
LINGYI YANG, UNIVERSITY OF OXFORD

Abstract. The report below describes the breadth of work done during the 2021 Mathematical Problems in Industry workshop project provided by Resource Systems Group, Inc (RSG) investigating techniques to ensure privacy in large-scale travel behavior datasets. In working with such large, publicly-funded data, the importance and need of making actionable information available to the public and policymakers needs to be balanced with the need to protect the privacy of individuals included within. During the workshop, we investigated algorithms and theoretical guarantees in the Differential Privacy framework, and the subtleties therein when applied to a complex data set. We report some preliminary results with extensions of DP algorithms, such as the "Random Response" algorithm, to a setting with categorical variables and uneven weighting (such as when dealing with vehicle makes and models). This extension introduces a parameter which directly allows one to balance between privacy and utility of the data. Lastly, we investigate some approaches to perturbation of trip start/end location data using the Laplace mechanism, and provide recommendations given observations during the workshop as well as directions for future work.

Contents

1. Background

1.1. **Household Travel Studies.** Household travel studies or surveys (HTS) are collections of detailed travel data representing who is traveling, when they are traveling, where they are going and how they get there [LDGE19]. A variety of questions may be answered based on the data from an HTS: How many car trips does the average resident make on a typical week? What are the peak travel hours, and what are the purposes of the trips taken during these hours? This data is used to inform investment and urban planning decision around transportation infrastructure [McG18]. More specifically, it can be used for traffic light scheduling, road redesign and other transportation optimization efforts [HG05]. For instance, a company may consider expanding a road based on the frequency that a road is traveled to help with the traffic flow. In recent years, the HTS data collection has became easier to obtain with the increasing population of people with smartphones [NWB+14]. Previously, the primary collection process consisted of paper surveys, and has now evolved to phone, online, and smartphone application surveys. With the expansion of the smartphone applications, participants' data can be collected in larger samples with a larger range of variants between parameters [LDGE19]. The collection of this data is usually commissioned by public agencies who often publish their results and collected data to the public. Examples of such transportation companies include the Ohio Department of Transportation [VS15] and the Metropolitan Transportation Commission [mtc12]. Since this data often has to be published to the public due to the nature of government/taxpayer funding, there is a major concern for the privacy of the participants [GBM15].

1.2. **Data Privacy.** The rise of privacy concerns in recent years has led to ever-developing policies that seek to protect the digital users' privacy [gov21, GV16] including HTS survey participants. In the most recent years, data collection companies have to comply with the General Data Protection Regulations [Tea17], California Consumer Privacy Act/California Privacy Rights Act [Buk19], and numerous other state and county privacy regulations.

Implementing and maintaining privacy relating to individual's data has been an ongoing challenge of the past decade. Past privacy breaches attest to the challenging nature of achieving true privacy for digital data which the public can access. An example of such incidents involves the streaming services company Netflix [Dun15, McN11]. In this incident, the company held a contest to improve its recommendation algorithm. Before making it publicly available, they cleansed the data by removing or encrypting personally identifying information such as name or address. Later, experts showed how by comparing this data with other reference data accessible to the public, the attackers can identify individuals in the data. This example tells us of the importance of testing against potential privacy attacks and the need for advancing data privacy.

1.3. **Summary of Report.** In this report, we study a few methods studying anonymization and differential privacy for the HTS data collected by RSG (the sponsor for the project) understanding this and similar data sets need to be published with some assurances relating to privacy. These methods include perturbing and aggregating the geographical data and applying generalized random response [Hol16] to the categorical data. First, we provide an overview of the type of data that we work with. Then, we define the random response

method and generalize it to best fit this data. Lastly, we discuss applying perturbation to the portion of this data that involves locations. Throughout these procedures, we use principles from differential privacy to measure our success at achieving data privacy in the following sections. We report that assessing the efficiency of these methods in light of differential privacy requires further investigation.

2. Exploratory Data Analysis

This section shows the result of our exploratory data analysis and preprocessing of the data set as provided – prior to exploring approaches to anonymization and differential privacy.

2.1. **Data collection process.** The data set in question, a household travel study (HTS) is intended to collect granular information about travel habits to better inform

Data collection took place through three main methods, smartphone data, call center data, and online data. Smartphone users downloaded an app that collected travel data for a continuous 7 day period. At the end of each day, the app would collect addition information such as reasons for traveling and why a person did not travel that day [LDGE19]. Those completing the survey through a call center or online only reported a single day travel diary.

The HTS which was provided to us by RSG contained, in total, 16,152 participants. Since participants filled in much of their data, there are places where the data is missing information. Out of the total participants, 11,405 used the app and 4,747 used the call center on answered online. The various aspects of this study summarized in table 1; aside from the "location" data set which has detailed GPS observations, the other categories contain summary information such as dates/number of trips by household ("Day"), socioeconomic information ("Household") and vehicle information ("Vehicle") among others. The multiple tables contain sufficient information by "observation" (row in a table) to connect the various aspects of the data to one another if needed.

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

Table 1. An explanation of each category of data and the number of observations within each category.

2.2. **Removal of geographic outliers.** Some trips venture or stay completely outside the Twin Cities metro area, as evidenced by locations in the `locations.csv` file. These may have resulted, for instance, when a participant flew to a different city and continued to track trips. They do not add much to the understanding of the Twin Cities area, and they interfere with statistical summaries.

A Julia code was written to group each location datum by trip ID number, then find the extreme values for longitude and latitude. A large box enclosing the metro area was drawn at latitudes (N44.1,N46.1) by longitude (W91,W94.3). Any trip ID whose extreme values were outside of this box was noted in a file, ending with 22,258 flagged trips. We note that the analysis was performed on locations that may have been perturbed by up to 1 km – but we expect the number of "false positives" in flagging outliers as a result of this were extremely small, if none at all.

2.3. **"Locations" data.** Figure 1 visualizes one example of a trip trajectory (chosen arbitrarily from the data). The direction of movement is indicated by arrows drawn. The figure shows that GPS data are collected with relatively high frequency, such that identifying a specific route taken is possible. From a preliminary standpoint, it is clear that this kind of information would be highly important to privatize – but the process of creating realistic synthetic data at this fine scale, which still has some value, will be a significant challenge.

While this fine-grained information about the routes taken in individual trips can be very valuable, it poses great challenges from the perspective of privatization. A straightforward simplification – to only report the start/end of a trip without reference to the specific route – also was reported in the "trips" dataset, which we explore next.

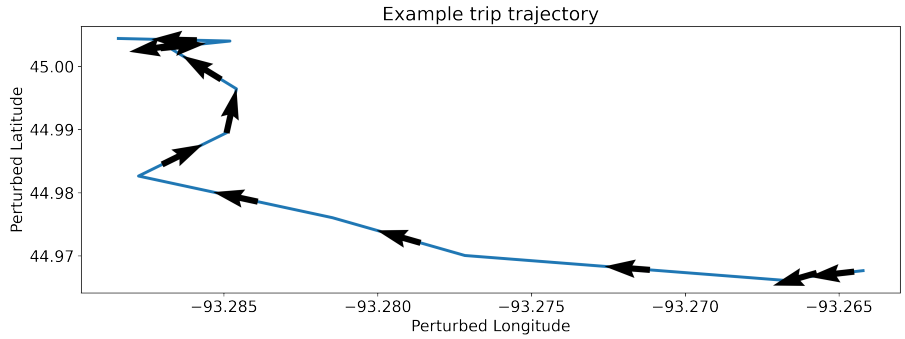

FIGURE 1. An example of a trip trajectory, the direction is indicated by an arrow (which are plotted at the midpoint of the interpolation between trip reference points).

2.4. **"Trips" data.** For HTS, we consider trips that individual people in each household makes as particularly sensitive data, so understanding aspects of this component of the data important.

Figure 2 shows the histogram of how many trips are made by each person on each day that they were part of the survey for. We see that most people on a typical day make very few trips, and there is a rapid decay in the trip frequency when the number of trips increase. We make note of this as the first, of several aggregate aspects of the overall data that may be valuable to account for if and when synthetically generated data is used.

One aspect of the data reported in each row of this data is the trip start/destination coordinates. We considered this distribution another element of note to preserve during the privatization process. In Figure 3 illustrate the spatial distribution of all trips in relative coordinates. The left panel shows a scatter plot of all trips distributions, where we see that most trips are focused on a tight cluster which is the Twin Cities region. The right panel visualizes the relative frequency of the same data distributed over the city. It is difficult to identify relative importance in this side view; however, as the z-coordinate (height) is the log-transformed frequency, it is clear the enormous majority of trips occur in the city center.

2.4.1. *Ride-hailing trips.* Trips coded specifically with `mode_type=6` are designated as ride-hailing service trips, such as those provided by Uber and Lyft. There are 1244 such trips in `trips.csv`. Histograms of the endpoint locations of these trips (distributions of all trips per latitude, or per longitude) are shown in Figure 4. The strong single-mode distribution suggests that approximately 1000 of the 1244 ride-hailing trips have Minneapolis-Saint Paul International Airport (MSP airport) as one of the trip endpoints, based on a manual cross-reference against the airport's coordinates.

To further explore the nature of ridesharing component of the data, Figure 5 (right panel) illustrates the distribution of departure times, by time of day. The lowest rates for these trips occur between roughly 8am to 12 noon, at an amount roughly 3-5 times lower than the rest of the day. Figure 5 (left panel) illustrates associated trip durations for rideshare trips, which appear to have a mean of 15-20 minutes, and the large majority of trips lasting less than 60 minuets.

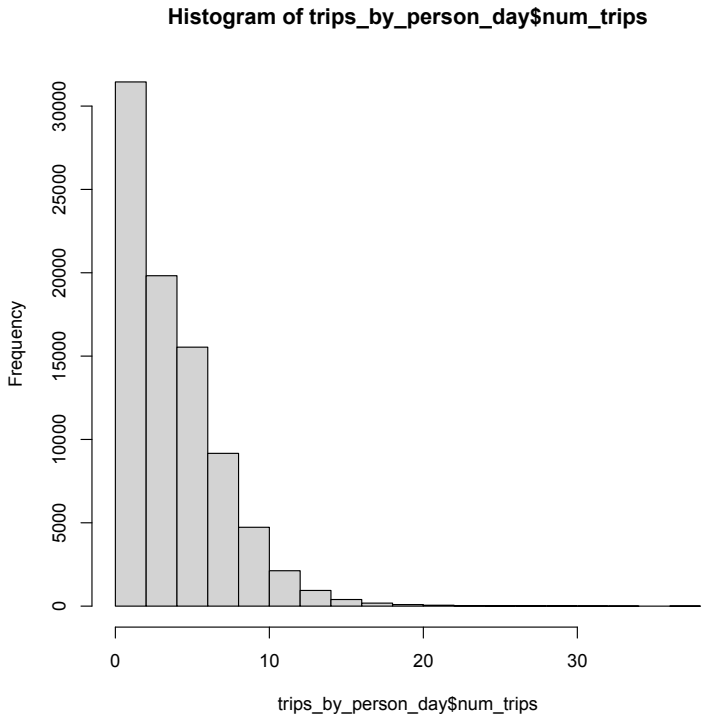

FIGURE 2. Histogram illustrating the number of trips taken by one person in a day. The large majority of people take fewer than 10 trips.

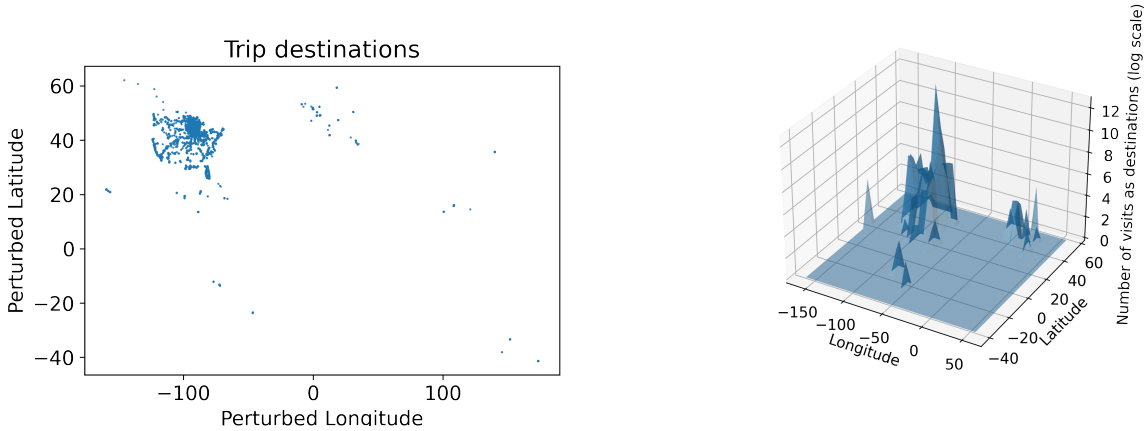

FIGURE 3. Visualization of the destination distributions

3. Differential Privacy

3.1. **The basics of DP and examples.** Our broad goal is to investigate approaches, and feasibility of, applying differential privacy algorithms to minimize a while still preserving the usefulness of the data. A differential privacy (DP) framework provides a way of ensuring such privacy with an underlying mathematical framework with some guarantees [Kur21]. Differential privacy is often implemented by of random perturbation of data, for the purposes of individual privatization, in such a way such that numerical bounds on statistics of the data information can be obtained [DR14].

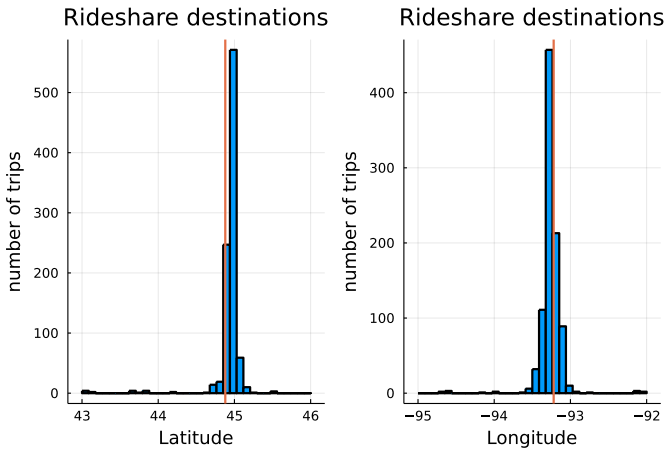

FIGURE 4. Destination locations (longitude and latitude) for all 1244 ride-hailing service trips. The vertical lines show the location of the MSP airport.

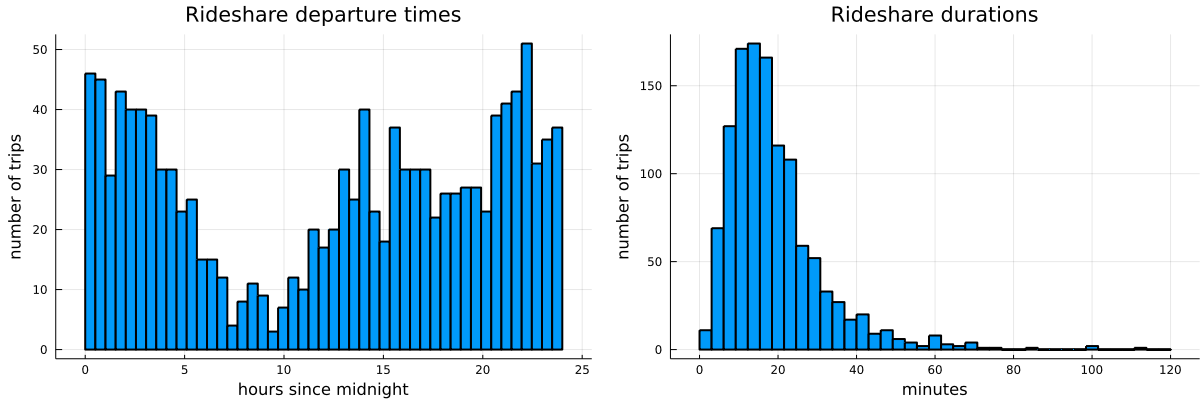

FIGURE 5. Departure times-of-day and trip durations for all 1244 ride-hailing service trips.

Differential privacy is beyond a hypothetical concern. Major companies, such as Netflix, found out the hard way that simply anonymizing data by obscuring strictly sensitive fields is not enough to maintain the privacy of study subjects [DR14, Dun15, McN11]. What's more, a first attempt at a "fix" – such as releasing seemingly harmless data such as the mean of a data set – is also not truly private, and is certainly not differentially private.

For example, suppose a teacher has a class of 17 students and tells them that the mean on their last test was a 80%. In the extreme, a motivated agent interested in knowing an individual student's score could take this mean, and the remaining 16 of the 17 students their individual scores on the test and use this information to determine the score of the last individual who wished to remain anonymous. In isolation, this example may seem a little absurd. However, typical approaches to de-anonymizing data work along similar principles – starting with an "anonymized" data set, then combining with information from other data sets and de-anonymize information about individuals (or close enough the individual level for all practical purposes).

3.2. **Formal Definition.** We now more formally define differential privacy. Differential privacy is used to answer numeric queries which we understand as functions from data bases to real numbers [DR14]. For example, a numeric query could return a sum or mean. We let $\mathcal{M}$ be a mechanism, an algorithm that takes a data set as an input and returns some output [DR14]. We call two data sets D_1 and D_2 "neighboring" if they differ by a single element. Then, we say that a randomized algorithm $\mathcal{M}$ is ϵ-differentially private if,

for two neighboring data sets D_1 and D_2, and all subsets S of the image of the function $\mathcal{M}$,

$$\exp(-\epsilon) \leq \frac{\Pr[\mathcal{M}(D_1) \in S]}{\Pr[\mathcal{M}(D_2) \in S]} \leq \exp(\epsilon). \tag{1}$$

Differential privacy is also often written as

$$\Pr[\mathcal{M}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(D_2) \in S], \tag{2}$$

understanding that the definition will be "symmetric" when switching the roles of D_1 and D_2.

If we think of $R \subset S$ as being the numerical result of a query, then $\mathcal{M}$ is ϵ-DP if the probability of $\mathcal{M}(D_1)$ producing the result R is at most $\exp(\epsilon)$ more likely than the probability of $\mathcal{M}(D_2)$ producing the result of R [Kur21]. The hope is for ϵ to be small, thus ensuring more privacy for the participants [Kur21]. If ϵ is small, then there is only a very small chance that the differing data sets D_1 and D_2 will produce a different result, and thus the difference between including and excluding an individual's information in the data is small, meaning they are protected [Kur21].

3.3. **Random Response Algorithm.** We will now explore one of the classic ways of implementing DP, called "Random Response." Suppose participants are asked a question that could revel information about their health such as "Do you eat peanuts?" To protect the privacy of everyone, the answers to this question can be recorded in the following way. A participant flips a coin. If the coin shows heads, they answer truthfully. If the coin shows tails, the participant flips the coin again. Now if it shows heads, they answer "yes" regardless of their true answer and if the coin shows tails, they answer "no." Intuitively, we understand the data resulting from this process is "private" because of a degree of "plausible deniability" of any outcome. We will see that applying this algorithm on a data set such as this is differentially private [DR14].

We will begin by fixing the answer of a respondent. Suppose their truthful answer to "Do you eat peanuts?" is "yes." However, using the random response algorithm, they will not always answer yes. Working through the space of possibilities, there is a 3/4 chance that the randomized response is "yes": there is a 1 in chance they tell the truth based on getting a tails in the first coin flip (choosing to give a randomized response), and a 1/4 chance they say "yes" based on first getting a tails and then flipping a heads. Similarly, there is a 1/4 chance that the respondent answers "yes" even though the truth is "no," as the result of flipping a tails (randomized response) getting tails then heads (the random response is "yes"). Using the definition of DP, we write

$$\frac{\Pr[R = \text{“yes”} \mid \text{Truth} = \text{“yes”}]}{\Pr[R = \text{“yes”} \mid \text{Truth} = \text{“no”}]} = \frac{3/4}{1/4} = 3 \leq e^{\epsilon} \tag{3}$$

where R represents the response. The case when the true response is "no" follows the same logic. From here, we see that $\epsilon = \log(3)$. So the coin flip algorithm is $\log(3)$-differentially private.

3.3.1. *Varying the truth reporting rate in Random Response.* Continuing the example in this section, suppose we wanted to determine the proportion of the population that eats peanuts. If this mechanism is applied for every individual's response, then we can compute an empirical estimate, simply by reporting the number of "yes" answers. If the true population proportion is p, and working with the knowledge that the data was modified by Random Response, the proportion of people that report that they eat peanuts is $\frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{1}{2}p$. Thus, if our observed fraction of "yes" responses is p', our estimate of the true proportion is obtained by solving the equation $p' = \frac{1}{4} + \frac{1}{2}p$ for p. Note that we multiplied our estimate for p' by 2 in order to construct our estimate of the true proportion. This governs how closely we should estimate p' in order to truly estimate p, in analogy to a sensitivity parameter. For example, if we want an estimate for p that is correct to within $\pm\alpha$ (with high probability), we need an estimate of p' that is correct to within $\pm\frac{1}{2}\alpha$ (with the same probability).

There is a tradeoff between this scaling parameter and the privacy guarantee. Suppose we work with a weighted coin with probability θ that coin lands heads (and the respondent reports truthfully), and the random response is still based on a random coin flip. Repeating the calculation from above,

$$\frac{\Pr[R = \text{“yes”} \mid \text{Truth} = \text{“yes”}]}{\Pr[R = \text{“yes”} \mid \text{Truth} = \text{“no”}]} = \frac{\theta + (1-\theta)\frac{1}{2}}{(1-\theta)\frac{1}{2}} = \frac{1+\theta}{1-\theta}. \tag{4}$$

so that controlling the random response rate with probability θ gives a $\log\left(\frac{1+\theta}{1-\theta}\right)$-differentially private algorithm. We see that for $\theta = 1/2$, we have the same $\log(3)$ differential privacy. When $\theta = 0$, every response

is replaced with a random response, and the bound is $\log(1) = 0$. This is an optimal value for differential privacy, but clearly the data set has no usefulness anymore. In the opposite limit $\theta \to 1$, vanishingly few responses are randomized, but the differential privacy *bound* grows to infinity, providing no guarantees for the algorithm applied to the data set.

3.3.2. *Tradeoff of privacy and utility from the perspective of measuring a population proportion.* Let p_t represent the fraction of responses that end up reporting the truth; whether they decide to report the truth or report a randomized response. For example, $p_t = 3/4$ when a fair coin is used in both stages of Random Response (RR). An important question is what and how the population proportion is measured after the RR algorithm is applied to the data.

If the true population proportion of "yes" responses is p, then using Bayes' rule, we expect to see a

$$\widehat{p} = p_t p + (1 - p_t)(1 - p) \tag{5}$$

ratio of "yes" responses in our sample. We would like to know how close $\widehat{p}$ is to the true population proportion p. Solving for p, we get that our estimate of the true proportion is

$$\frac{\widehat{p} + p_t - 1}{2p_t - 1}. \tag{6}$$

We can think of $\frac{1}{2p_t - 1}$ as a "scale parameter" in analogy to the scale parameter for the Laplace mechanism discussed in the following section. As p_t approaches 1, all responses are reported truthfully; the "scale parameter" $\frac{1}{2p_t - 1}$ approaches 1, the estimate of the proportion goes to $\widehat{p}$ (the measured response proportion), and the privacy parameter goes to ∞. As p_t approaches $1/2$, the "scale parameter" approaches ∞, while the privacy parameter goes to 0.

3.4. **The Laplace Mechanism.** One of the most common ways to add noise to real-valued data (in contrast with binary responses) from the perspective of differential privacy is through what is called the "Laplace mechanism." This method is appealing because it can be demonstrated to preserve ϵ-differential privacy in a predictable fashion. We first define the Laplace Distribution, a distribution centered at 0 with scale parameter $b > 0$ and the probability density function

$$\text{Lap}(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right). \tag{7}$$

The Laplace mechanism computes the query f, and perturbs using noise from the Laplace distribution [DR14]. Given a function f, the Laplace Mechanism is

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where the Y_i are random variables from $\text{Lap}(x|\Delta f/\epsilon)$ [DR14].

3.5. **Is DP useful for our data?** In Section 3.4, we see that the result of the Laplace Mechanism is to randomize the answer to a query, which could be a mean, a maximum, or even the entire database itself. When ϵ is small, differential privacy algorithms produce perturbed data that protects the privacy of the individuals of the study. For example, releasing a mean test score that has gone through the Laplace Mechanism would prevent 16 students from figuring out the score of the last student. However, in the strict sense, applying differential privacy in the does not return a perturbed and private data set; rather it returns perturbed means, maximums, etc. In Section 4.1, we look at one possible way that the Randomized Response algorithm could be applied to individual categories in the data set to make the overall data set private. There are likely ways to use modified approaches to differential privacy to produce a private data set, but there are also other spatially-based perturbation methods that we explore in Section 5.

4. Application of Random Response methods on categorical data

4.1. **Who drives that Lotus?** The provided data set offers interesting information on the types of vehicles for each household in the survey. Some details are very recognizable, for example the vehicle make and fuel type. However, some features are harder to identify, such as year and specific model.

It's not a surprise to see that vehicle data has a direct impact on city planning. As an example, the rise of electric vehicles demands increasing infrastructure (e.g. charging stations) to support them [Glo21]. More generally, knowing the properties of vehicles can inform expectations about road use, stress, etc.

While it is important to have data about the types of cars in a particular area, knowing the number, make, and model of vehicles driven by members of a particular household can also offer insight into sensitive information about the household's members. For example, if it is known that a household member drives an expensive vehicle, such as a Lotus, a Tesla, or a Hummer, then someone with access to this data can speculate as to the household income and other sensitive information. If a household has more than two vehicles, then one can speculate that there are more than three licensed drivers in a household. Driving a minivan can signal that there are children in a household. All of this information can contribute to data revealing the identity of members of a household. For this reason, we chose to investigate methods of making vehicle data private. Here we explore extensions of the basic differential privacy algorithms while considering attempting to preserve useful aspects of the data set; such as proportions of car types.

In this data, respondents were asked to provide the year, make, and model of each vehicle in their household using pull-down menus. For vehicles built prior to 1980, respondents were instructed to answer "1980 or before."

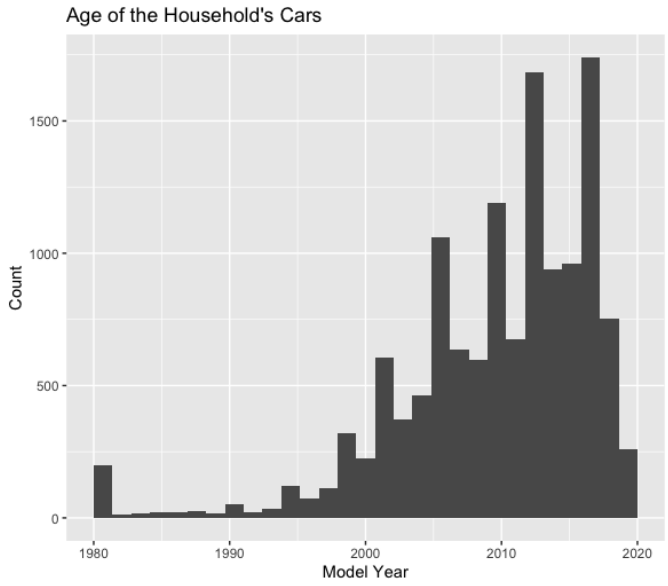

FIGURE 6. Age of the Household's Cars in the Greater Minnesota Community.

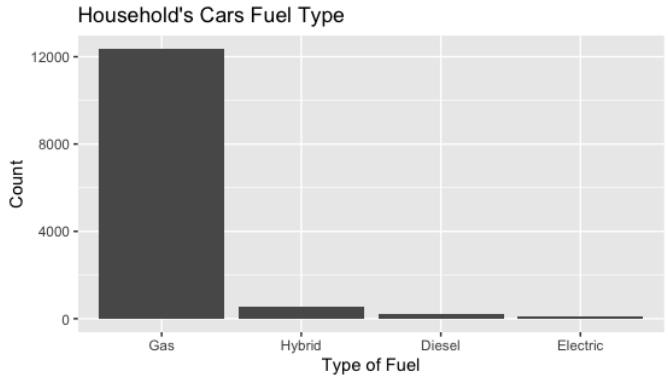

FIGURE 7. Fuel Types of the Household's Cars in the Greater Minnesota Community.

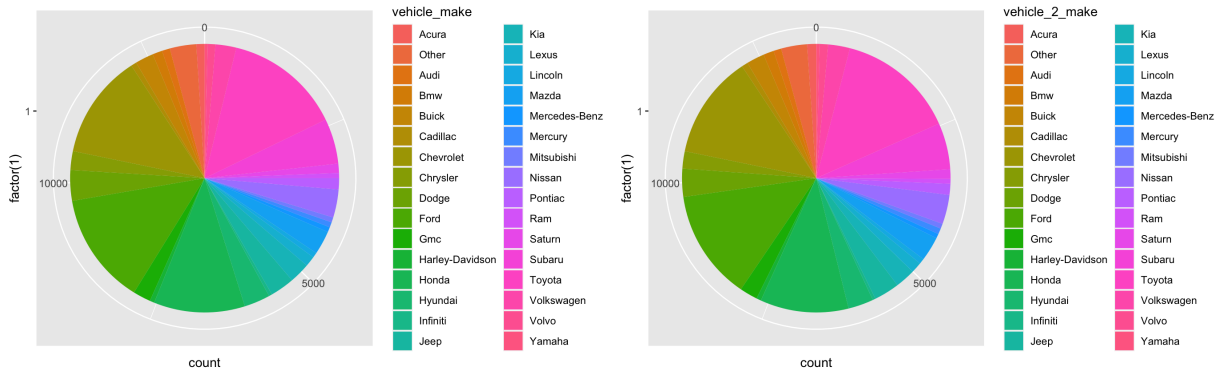

FIGURE 8. Distributions of vehicle makes before and after implementation of coin flip algorithm

4.1.1. *Data Cleaning.* Data cleaning and processing was done in several ways to extend usefulness and avoid the need to throw away data where reasonable imputation could be done. We manually examined data entries where the fuel type was missing but the vehicle name was there and imputed correct data on what fuel type that vehicle is.

Aside from this, data missingness and non-uniformity was an issue throughout this data. For example, some car names only consist of a year, are missing the model, or have spelling errors or non-uniform naming. For example, some people called their vehicle a "Chevy" while other people wrote "Chevrolet." We handled this and similar algorithms using a basic text reassignment, though we discovered the need for additional data cleaning while examining the relationship between the original and perturbed data after performing the reassignment algorithm.

4.1.2. *Random Replacement applied to car types.* From Section 3, we know the Random Replacement (RR) algorithm is differentially private. In this section we will apply this method to the vehicle categories to try to increase the privacy of the users, while keeping the integrity of the vehicle fuel type (i.e. keep the same amount of electric vehicles, vs gas, diesel and hybrid fuel types). Note that because we are maintaining the integrity of the fuel types (i.e. a person with an electric car will still report that they drive some sort of electric car), the overall data set, by definition, might not be fully differentially private.

We first create subcategories of the vehicle models based on the fuel types gas, diesel, hybrid, and electric. Refer to Figure 7 for the current distribution of vehicles based on fuel types. Each fuel subcategory is filled with the already existing vehicles, taking year, make, and model as as a single unit (for example, 2011 Subaru Outback is an element of the gas subcategory). For each entry, a coin will be flipped. If the coin lands on heads, the truthful make and model remains. If the coin lands on tails, a randomly selected vehicle make and model would be selected from the same subcategory as the original.

Perturbing the data within the subcategories based on fuel ensures that the ratio of gas, diesel, hybrid, and electric cars is preserved. However, more specific details such as whether a person owns a gas consuming motorcycle or a gas consuming truck will be lost. Different subcategories would be required to preserve such data. In Figure 8, the distribution of vehicle makes in the original data and the perturbed data are displayed.

Of the 13,431 vehicles observed in the data, 6772 vehicle names (50.42% of total) remained the same using RR. The remaining 6659 had their vehicle names randomly reassigned from the set of all vehicle names with the same fuel type. After performing reassignment, a total of 6788 (50.54% of total) vehicle names were the same after reassignment as they were in the original data set. Thus only 16 vehicles total–9 gas, 4 hybrid, and 3 electric–were randomly assigned to the same vehicle that the household actually owns. Figure 9 illustrates this comparison of the distribution of vehicles based on year in the original and perturbed data. From a glance, the overall distribution is fairly similar, which is useful to a city if they are interested in the age of cars driven in the city.

4.1.3. *Revisiting Random Response with multiple categories.* Intuitively, we suspect that this modified Random Response method should be differentially private with respect to vehicle names. However, it will be useful to follow through with a calculation to produce a relevant differential privacy style bound. Finding

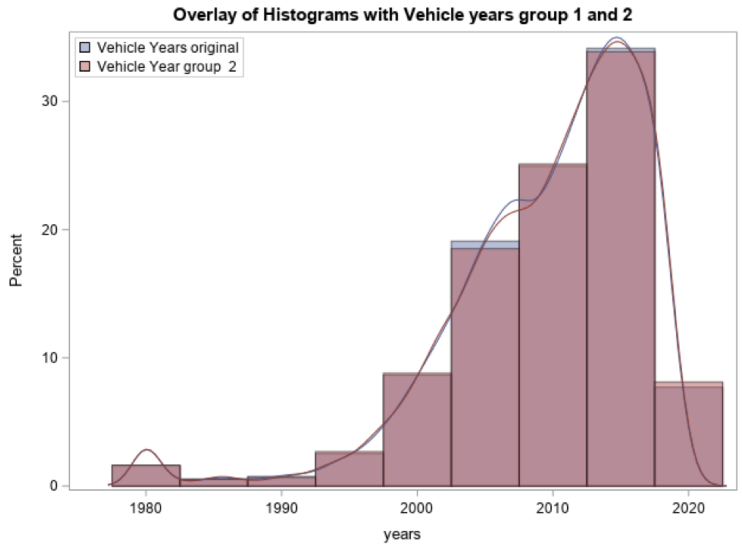

FIGURE 9. A comparison between vehicle year distribution in the original data versus the perturbed data.

a formula for the associated value ϵ in this scenario is an extension of the proof in section 3.3 when the response was binary ("yes" or "no").

In the modified RR algorithm, the reporting of a person's vehicle is based on a coin flip, where a "heads" means the true car is reported, and a "tails" means that the reported car is randomly selected from within the same subcategory of vehicles. We will look at the differential privacy within a fixed fuel type subcategory. We begin by fixing a respondent who drives car type c. We want to find the probability that they report car c as their vehicle given that the truth is car c, and we look at the probability that they report that they drive car c given that they drive a different car c':

$$\frac{\Pr[R = c|\ \text{Truth} = c]}{\Pr[R = c|\ \text{Truth} = c']}. \tag{8}$$

We must run through all possible cars that a person could and could not drive, so we now consider c_j and c_i' as the car that someone owns and the car that someone does not own respectively. For example, if c_j is a 2011 Subaru Outback, then we look at the ratio when c' is a 2018 Toyota Corolla, when c' is a 2017 Honda CR-V, etc. We then must also consider the case when c_j is the 2018 Toyota Corolla compared to c_i being a 2011 Subaru Outback, 2017 Honda CR-V, and so on. This ratio takes the worst case scenario value; i.e. the value that yields the largest ϵ. We calculate the probabilities in the following way. The probability that the respondent reports c_j given that the truth is c_j is given by the sum that the probability that a heads is flipped and the probability of car c_j being selected from the total vehicles in the sub category. The probability that the respondent reports c_j given that the truth is c_i' is given by the probability that they flip a tails and select the other car type. To take into account that there could be multiple of the same car c_j, we'll let n_j be the total number of c_j, n_i' be the total number of car type c_i', and t the total number of cars in the subcategory. Refer to Figure 10 for a visual of this scenario. From here we can find the probabilities and take the worst case scenario - the maximum - of the ratios

$$\max_j \left\{ \frac{\Pr[R = c_j|\ \text{Truth} = c_j]}{\Pr[R = c_j|\ \text{Truth} = c_i']} \right\} = \max_j \left\{ \frac{\frac{1}{2} + \frac{n_j}{2t}}{\frac{1}{2} \cdot \frac{n_j}{t}} \right\} = \max_j \left\{ \frac{t + n_j}{n_j} \right\} = \frac{t+n}{n} \tag{9}$$

where n, the value that maximizes the ratio, is the minimum of the of owned vehicles c over all n_j. From here, we conclude that this modified RR algorithm is differentially private with

$$\epsilon = \log\left(\frac{t+n}{n}\right). \tag{10}$$

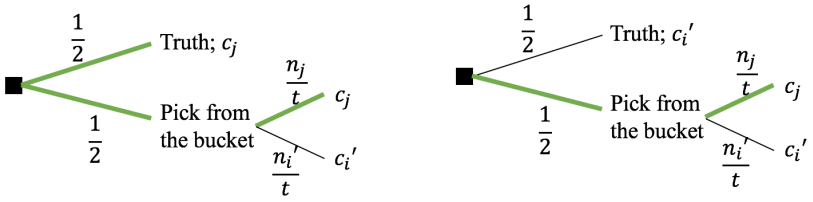

FIGURE 10. The left diagram shows the method of calculating the probability for the numerator of the ratio for the modified coin flip algorithm, and the right diagram shows the method of calculating the probability of the denominator.

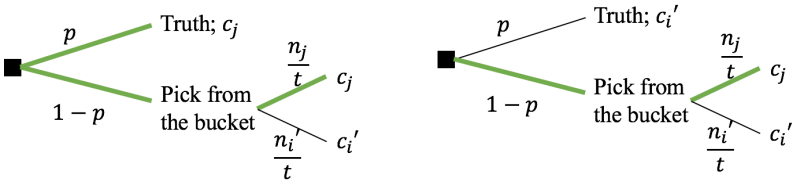

FIGURE 11. The diagram on the left shows the method to calculate the probability of the numerator for the generalized probability algorithm, and the diagram on the right shows the method to calculate the probability of the denominator.

Since t is the total number of vehicles in the sub category and t will likely be large whereas $n < t$ (or possibly $n \ll t$), we expect ϵ to be large, and hence provides little or no guarantees about differential privacy. The precise values here as well as studying how this modified RR depends on parameters in the algorithm is an area for future study.

4.1.4. *Random Response with weighted coin flips and multiple categories.* Suppose we want to further generalize the modified coin flip algorithm to have better control of the balance between privacy and data utility. Suppose the probability that the truth is reported occurs with probability p (note: this corresponds to θ in Section 3.3.1). The generalized diagram is shown in Figure 11. The ratio of probabilities is now

$$\max_j \left\{ \frac{\Pr[R = c_j |\ \text{Truth} = c_j]}{\Pr[R = c_j |\ \text{Truth} = c_i']} \right\} = \max_j \left\{ \frac{p + (1-p)\frac{n_j}{t}}{(1-p) \cdot \frac{n_j}{t}} \right\} = \max_j \left\{ \frac{pt + (1-p)n_j}{(1-p)n_j} \right\} = \frac{pt + (1-p)n}{(1-p)n} \tag{11}$$

Again where n is the minimum over all values n_j. This situation gives an expression for differential privacy:

$$\epsilon = \log\left(\frac{pt + (1-p)n}{(1-p)n} \right) \tag{12}$$

The benefit of this algorithm is that we have more control over ϵ than we did in the modified coin flip because of the addition of p controlling the rate of reporting the truth. If p is large, then there is a high probability that a person will report their true vehicle, and thus we expect there is less privacy. If p is small, then there is a small probability that an individual reports their true vehicle, ϵ approaches 0, and differential privacy guarantee is improved.

4.2. **Differential Privacy analysis for trip data.** Suppose we take $n = n' = 1$ in (11). The resulting algorithm corresponds to random response where one replies with his/her actual car with probability p, and selected uniformly at random from a list of t cars with probability $1 - p$. Revisiting the derived DP bound, asymptotically

$$\epsilon = \log\left[\frac{tp + 1 - p}{1 - p} \right] \sim \log t, \tag{13}$$

for $t \gg 1$. For fixed p, the value of ϵ grows as the size of the list grows. One could keep ϵ moderately small by p near zero, but this produces a mostly synthetic data set that very inaccurately models the original.

Recall that the "Trips" data set records starting and finishing locations. Briefly, we propose to first aggregate these start and finishing locations to one of about 2500 census blocks, as is seen in practice (see Fig. 12)Then, we may consider another variation of the Random Response algorithm. Consider using

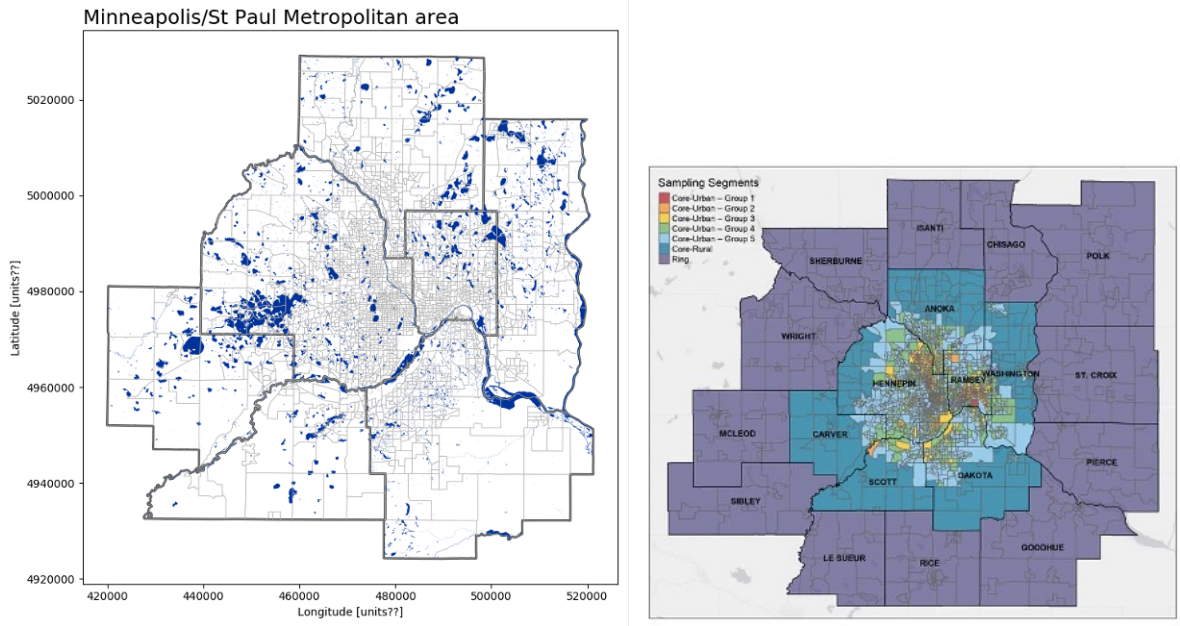

FIGURE 12. Division of Minneapolis/St. Paul Metropolitan area into about 2500 blocks (left) and 19 counties (right).

randomized response to mask a database of two-stop trips by replacing them with a randomly generated pair (i, j) with probability $1 - p$. This method is qualitatively similar to perturbing the longitude and latitude of the starting/finishing points and then aggregating them to blocks. Note that trips can be represented as ordered integers (i, j) where $i, j \in \{1, 2, \ldots, 2500\}$ and there are 2500^2 distinct trips. With $p = 1/2$ and $t = 2500^2$ in (13), we have

$$\epsilon \approx 16.$$

In practice, this estimate for ϵ is quite large. Trips with more stops will have even greater values of ϵ. The intuition is that because the state space of trips is so large, entries are more "unique" than when the state space is small. Removal of any one entry could drastically change the conclusions of any study performed on this data set and therefore this approach to masking the trip data is not very differentially private.

5. Study of home location perturbation with synthetic census blocks.

In this study, the actual data is household location data and it is aggregated into census blocks which are geographical regions whose populations are roughly the same. As one would expect there is a trade off between amount of perturbation and accuracy of data. If only a small perturbation is applied to the data before aggregation it is likely that the census block labels for each household remains unchanged. If a large perturbation is applied to the data before aggregation it is likely that the census block labels are highly inaccurate and useless. Therefore in practice one must tune the perturbation to the accuracy and protection desired. It is for this reason that we studied the relationship between perturbation and accuracy.

As household locations in a census block can be thought of points sitting inside a polygon, accuracy can be understood as the distribution of points in a particular polygon whose census block label changes after perturbation. In this case we say the point escaped the polygon after perturbation. Therefore, we studied the effects of data perturbation before aggregation on the household location-census block relationship.

Currently data about household location is often associated with a *census block*, which is a region with a population in a particular range. Though there is an R code that will associate latitude and longitude coordinates with a census block, at first we chose to model a much simpler problem working on a simple quasi-rectangular grid.

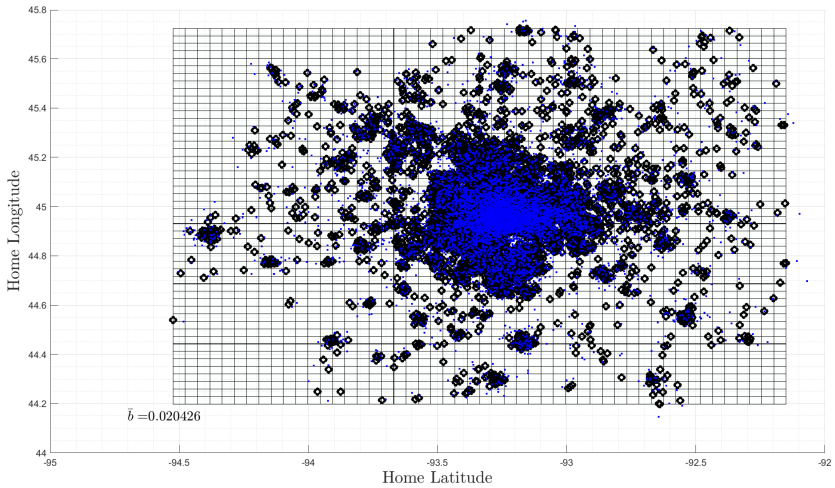

FIGURE 13. Example of household location data overlaid on uniform grid for $N = 50$. The black diamonds represent the actual data; the blue dots represent the perturbed data.

We took the rectangular region enclosing all the household data points and subdivided it into a uniform grid. The range in each direction was subdivided into N intervals, yielding N^2 "blocks." We started with nearly 7400 household locations, and perturbed them using a technique similar to the Laplace mechanism (7). Instead of using (7) for the radius of the perturbation, we used separate Laplace noise in the x- and y-directions with parameters b_x and b_y, where

$$b_x = \frac{\Delta f \text{ (in the } x\text{-direction)}}{\epsilon}, \tag{14}$$

and similarly for b_y. A visualization of the results is shown in Fig. 13.

To analyze the effects of this perturbation scheme, we performed ten different random realizations of the perturbations, and then averaged the results. We did this for various values of the Laplace scales, and computed the proportion of households that left their original block under perturbation. Results are shown in Fig. 14, where the horizontal axis is $\log \bar{b}$, where $\bar{b}$ is just the average of b_x and b_y. As expected, for fixed N, the larger the value of $\bar{b}$ (and hence the larger the perturbation), the larger the proportion that left the original block. Similarly, for fixed $\bar{b}$, the larger the value of N (and hence the smaller the original block size), the larger the proportion that left the original block.

These results, along with the form of (7), suggest a natural relationship between $\bar{b}$ and N. In particular, we expect that what is important is the ratio of the Laplace scale to the grid spacing, which behaves as N^{-1}. Hence in Fig. 15 we plot the same results as in Fig. 14, but with the horizontal axis equal to $\log(\bar{b}N)$. Note that the simulations now lie nearly on top of one another, indicating the appropriateness of the lumped parameter. (Note also that at the upper end, the Laplace parameter is 100 times as large as the grid spacing, which seems unrealistic.)

Another parameter of interest would be preservation of block population. In particular, given our previous results, we know that after perturbation, households will move from block to block. Hence we would expect some blocks to experience a net loss of households, some would have a net gain, and the population of some blocks to remain the same. Since census blocks have roughly the same population and can be a useful way to categorize data, it may be useful to maximize the proportion with no net change.

To analyze this situation, we used the algorithm described above to compute the proportion of blocks that retained the same population. The results are shown in Fig. 16. As expected, for fixed N, the larger

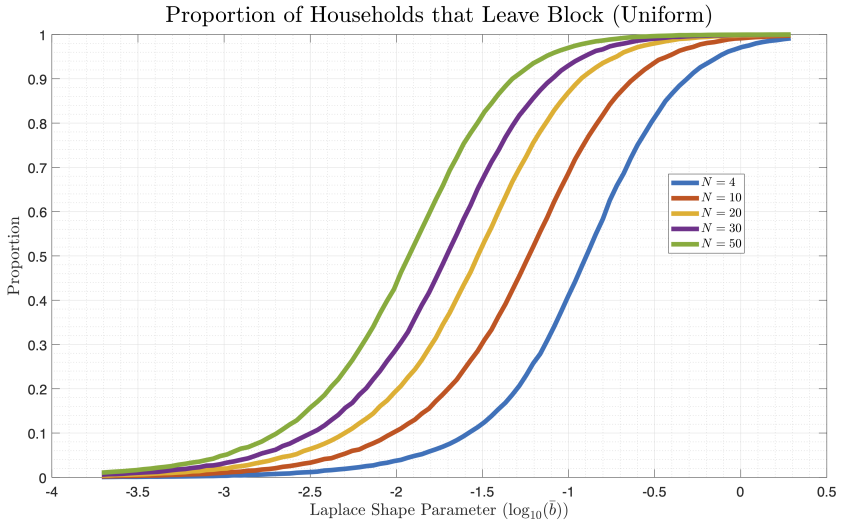

FIGURE 14. Comparison on proportion of locations perturbed outside the original block group.

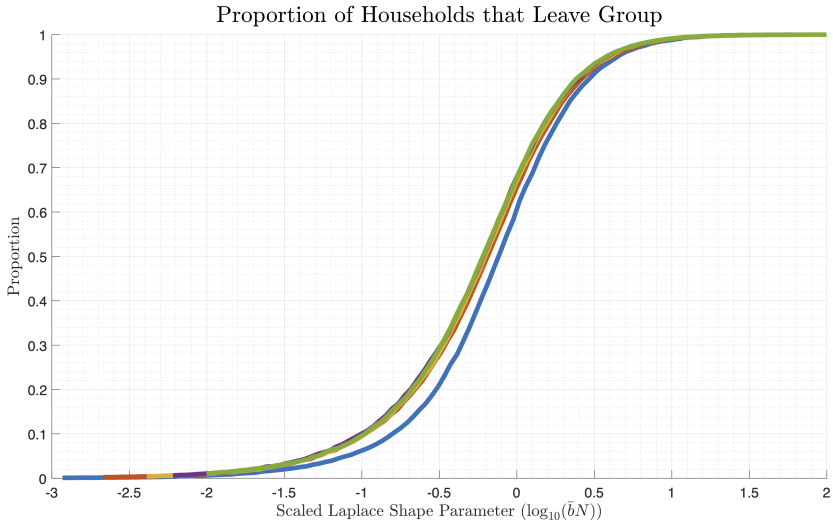

FIGURE 15. Comparison of the proportion of perturbed points that left original block with respect to the scaled parameter $\bar{b}N$, where $\bar{b}$ is the shape parameter of the Laplace distribution and N is the grid size parameter.

the value of $\bar{b}$ (and hence the larger the perturbation), the smaller the proportion of blocks with unchanged population. Similarly, for fixed $\bar{b}$, the larger the value of N (and hence the smaller the original block size), the smaller the proportion of blocks with unchanged population.

To get more information about the distribution than just the proportion of blocks with unchanged population, it would be interesting to create a histogram with the number of blocks *vs.* unsigned relative change in population. However, we were unable to do this before the workshop ended.

Though the uniform grid is easy to analyze, the visualization in Fig. 13 shows a weakness of the model. In particular, recall that the original census groups that we are trying to model contain roughly the same overall population. But in the data at hand, the households are clustered around the center of the region, as would be expected from a metropolitan area. (Also see left of Fig. 12.)

Hence we generalize our previous model by allowing the grid spacing in each direction to be nonuniform. This still produces a regular grid of rectangles, but with different dimensions and areas. In particular,

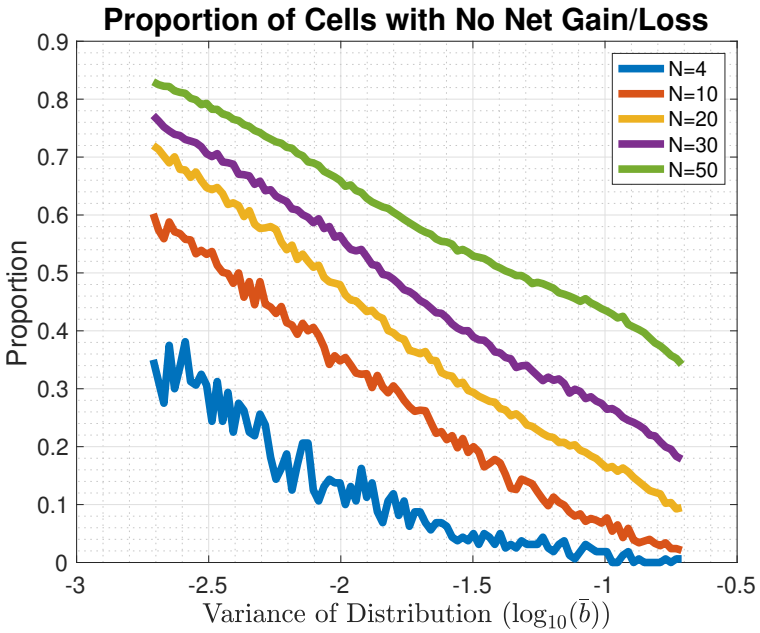

FIGURE 16. Comparison on proportion of blocks with no net change after perturbation.

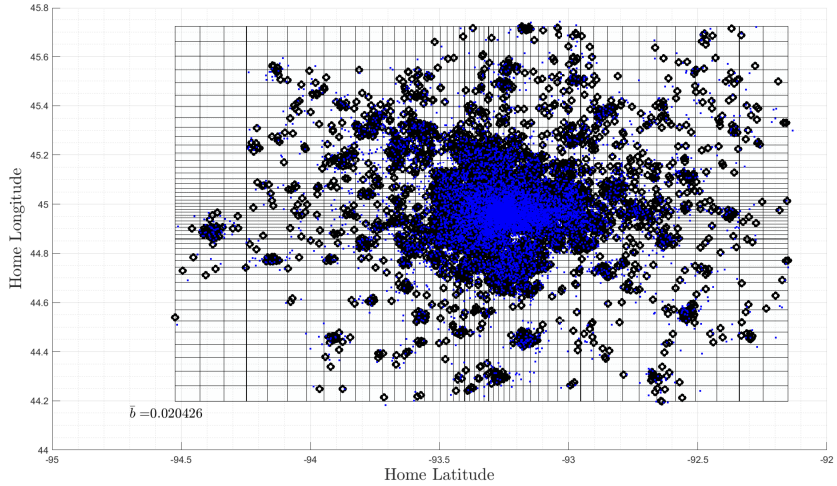

FIGURE 17. Examples of household location data overlaid on nonuniform grid. Parameters are as in Fig. 13.

we took an irregular spacing of x and y coordinates using a Gaussian distribution to describe the spacing between each point, which produced smaller regions near the densely populated center; see Fig. 17. We would expect that if we generated a histogram of the number of blocks by household population, the nonuniform grid would produce a much more uniform distribution. However, the workshop ended before we had time to investigate this more fully.

We repeated the analysis shown in Fig. 14 and obtained similar results, which are shown in Fig. 18. A comparison of the uniform and non-uniform grids is shown in Fig. 19. As expected, since for fixed N the nonuniform grid contains more blocks of a smaller size, it is easier to displace a household from its original block, the curve for the nonuniform grid lies above the uniform case.

We also repeated the analysis shown in Fig. 16 and obtained similar results, which are shown in Fig. 20.

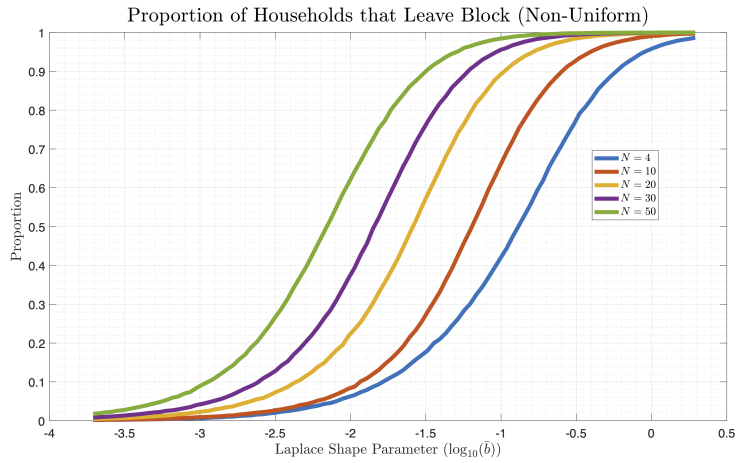

FIGURE 18. Comparison on proportion of locations perturbed outside the original block group.

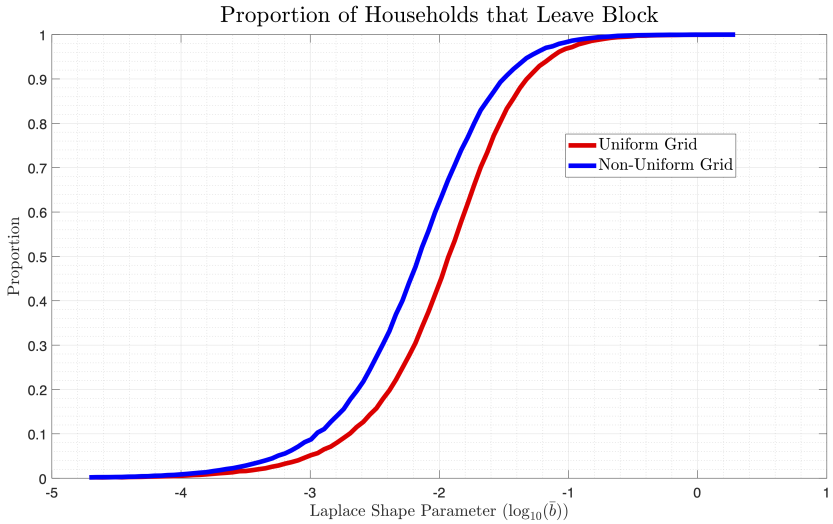

FIGURE 19. Comparison between uniform and nonuniform grids for $N = 50$.

Using the Matlab command `inpolygon()`, one can determine whether a given point is inside a given polygon. Using the command we modeled the proportion of points that escaped a given polygon after applying Gaussian noise. For demonstration purposes we studied this model with a pentagon inscribed in the unit square but we must note that this model can be easily adapted to any polygon of any size, and any probability distribution for the noise. In Figure 21 we model the relationship between the proportion of escaped points and the variance of the perturbation. Note this simulation was run with 100 uniformly distributed points inside a pentagon. We studied the case of 10 points and and 50 points and the result was near identical. For each variance we output the average over 100 computations of the proportion of escaped points.

We can also output the variance among these 100 computations to study how stable this procedure is. Stability of this procedure of really important because it gives us probabilitic guarantees that our computations are consistent.

So far we have discussed the relationship between the proportion of escaped points and variance in a given polygon. This is a local analysis and we were also interested in studying the accuracy of the data after perturbation over all census blocks. We note the following. If we have N census blocks each with population

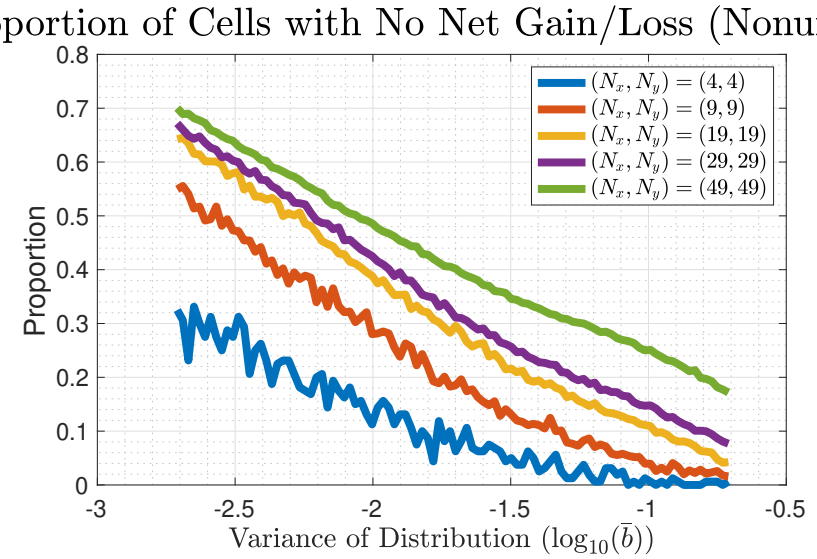

FIGURE 20. Comparison on proportion of blocks with no net change after perturbation, nonuniform case.

P_i and the proportion of escaped points in census block i is ρ_i, then the proportion of escaped points over the entire city population is

$$\frac{\sum_i P_i \rho_i}{\sum_i P_i} \tag{15}$$

Therefore, since each census block has roughly the same population, we think of each P_i as roughly some quantity P. If we have that each ρ_i is roughly some quantity ρ, then the proportion of escaped points over the entire city population is roughly ρ. Having all the ρ_i roughly equal to some ρ is a very natural condition since one would like that the accuracy is preserved uniformly in each census block.

This argument reduces the analysis over preserving accuracy globally across the city into an analysis of the accuracy over each single census block or polygon.

In conclusion, with this approach one can determine the variance that provides a desired proportion of escaped points (accuracy of data after perturbation); and with plots like Fig. 21 one can study this relationship and empirically tune the noise that would provide the desired accuracy and privacy. However, we have not established if, or how at what level, this approach provides differential privacy.

6. Recommendations and Future Work

We have explored various approaches extending the basic Random Response algorithm associated with differential privacy working with categorical (as opposed to spatial) data. Here we provide recommendations and detail potential algorithms which may be of use in the future.

6.1. **Recommendations relating to categorical data.** Below are some suggestions, primarily relating to other algorithms published in the literature, for publishing trip data that may provide better differential privacy guarantees.

- Publish the 10 most common destination blocks after addition of noise. This data set can be made ϵ-DP by taking the following steps:
  (1) Compute N_i, $i = 1, \ldots, 2500$, the frequencies/counts of trips that end in each of the 2500 blocks.
  (2) Add Laplace$(1/\epsilon)$ noise to each count:
  $$\tilde{N}_i = N_i + \eta_i, \quad i = 1, 2, \ldots, 2500,$$
  where $\eta_i \sim Lap(1/\epsilon)$.
  (3) Publish the 10 largest values from $\{\tilde{N}_1, \ldots, \tilde{N}_{2500}\}$.

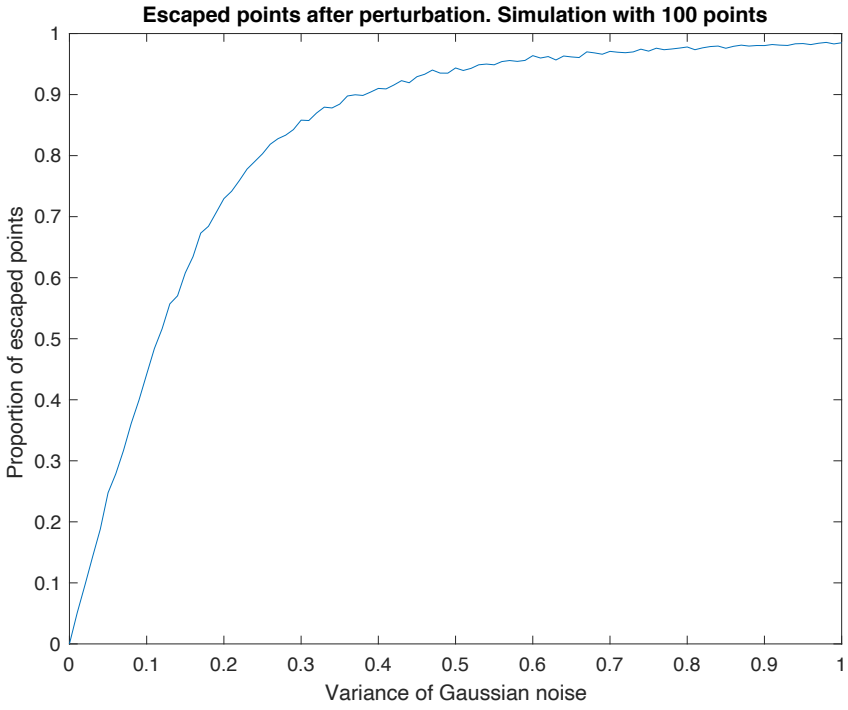

FIGURE 21. Relationship between proportion of escaped points and variance of perturbation

It has been shown [DR14] that the Laplace mechanism in steps 1 and 2 are ϵ-DP. The magnitude of noise added depends on ϵ which may have to be large to keep the noise small. This will prevent counts from becoming negative and to maintain accuracy.

- Publish randomized start at the county level for each trip (there are 19 of them – see Fig. 12). Using randomized response with $p = 1/2$ and $t = 19^2$, we calculate $\epsilon \approx 6$ for two-stop trips, so this method gives a better bound than applying similar methods at the county-block level.
- Adapt the SSD ("Sampling Distance and Direction") method from [JSB+13] which would process longitude/latitude measurements (x_i, y_i) from the original data. The method would allow the publication of a *single*, ϵ-differentially private trip of the form

$$\{(x_0, y_0), (\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_n, \tilde{y}_n), (x_{n+1}, y_{n+1})\}. \tag{16}$$

Note that the start and end positions are not perturbed in this version of SSD. Assuming that the distance between any two consecutive stops does not exceed M, SSD randomizes the trip in the following way:

(1) Let $\tilde{x}_0 = x_0$, $\tilde{y}_0 = y_0$, $\tilde{x}_{n+1} = x_{n+1}$, $\tilde{y}_{n+1} = y_{n+1}$ (start and end positions are not perturbed)
(2) For $i = 1, 2, \ldots, n$:
(3) Let $\mathbf{v}_i = (x_i - \tilde{x}_{i-1}, y_i - \tilde{y}_{i-1})$.
(4) Compute $r_i = ||\mathbf{v}_i||$ and $\theta_i = \arg(\mathbf{v}_i)$ (angle of $\mathbf{v}_i$).
(5) While $||(\tilde{x}_i, \tilde{y}_i) - (x_{n+1}, y_{n+1})|| \geq (n + 1 - i)M$:
  (a) Sample $\rho_i \in [0, M]$ and $\alpha_i \in [0, 2\pi]$ with

$$\begin{aligned} \text{Prob}(\rho_i) &\sim \exp\left(-\epsilon|\rho_i - r_i|/(8M)\right), \\ \text{Prob}(\alpha_i) &\sim \exp\left(-\epsilon|\alpha_i - \theta_i|/(8\pi)\right). \end{aligned}$$

  (b) Let $\tilde{x}_{i+1} = \tilde{x}_i + \rho_i \cos\alpha_i$ and $\tilde{y}_{i+1} = \tilde{y}_i + \rho_i \cos\alpha_i$.

The value of ϵ in the algorithm can be tuned: smaller ϵ corresponds to more noise, more privacy but less accuracy. This trade-off between accuracy and privacy is typical in most problems.

6.2. **Live travel perturbation.** When considering spatial data, we have primarily focused on the beginning and end of a trip – but it may be important to also know precise routes. This information was included in the data provided, but we concluded after much discussion that this was too challenging to handle during the time span of the workshop. Figure 22 illustrates a first attempt at such perturbation. However, investigating notions of differential privacy at such fine granularity, as is producing realistic, synthetic data (which is a persistent problem underlying the Random Replacement algorithm for complex data) and could be the subject of its own project.

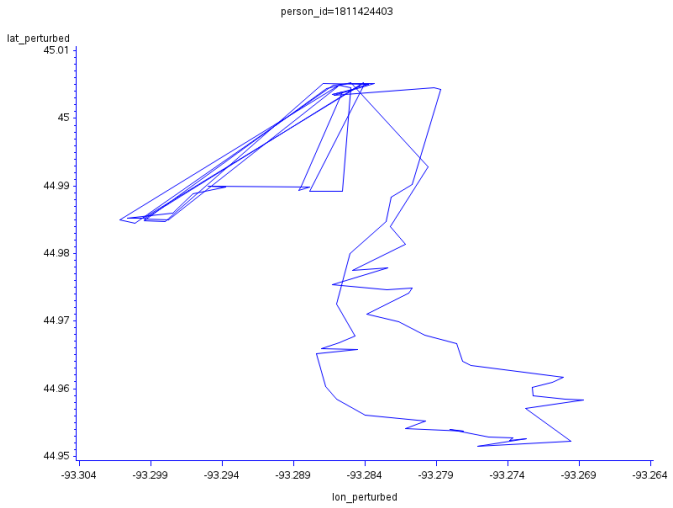

FIGURE 22. Trips latitude(lat) perturbed versus longitude (lon) perturbed.

6.3. **A bare-bones example illustrating a randomized selection of "to" and "from" locations.** Here we describe the details of a potential location perturbation algorithm which may be applied to the "Trips" data set.

6.3.1. *The simplest case: "to" and "from" locations are independent.* We *assume* that:

(1) The company (RSG) has true statistics (histograms) of how many households (HHs) travel from and to specific locations.
(2) The locations are coded as the census blocks (in what follows, just "Blocks") where they lie in. Therefore, the true histograms

$$\text{ht}_{\text{from}}(i) \qquad \text{and} \qquad \text{ht}_{\text{to}}(i), \qquad 1 \le i \le N, \tag{17}$$

are assumed to be known; here i is the index of the Block and N is the total number of Blocks. The notations 'ht_{from}' and 'ht_{to}' stand for 'histogram–true–from' and 'histogram–true–to', respectively.
(3) The company *needs to report* trips in a way that (approximately) preserves the individual histograms in (17). For example, it will report that HH α has taken a trip from Block i to Block j in such a way that the *reported* histograms

$$\text{hr}_{\text{from}}(i) \qquad \text{and} \qquad \text{hr}_{\text{to}}(j) \qquad 1 \le i, j \le N, \tag{18}$$

are (approximately) the same as the true histograms ht_{from} and ht_{to}. Here the 'r' in the notations hr_{from} and hr_{to} stands for 'reported'. Also, we will enumerate Blocks with Latin letters and HHs with Greek letters.
(4) In this subsection, we assume that the histograms for trips 'from' and 'to' Blocks are independent. I.e., one is interested only in these two statistics and is *not* interested in any statistics derived from them, e.g., in the number of miles traveled. The more complicated case when one is interested also in a derived statistics (e.g., that of total miles traveled), is considered in the next subsection.

Then the proposed randomized process of the location (i.e., Block) selection is as follows.

Step 1 For the first trip (during the day, or in a given report), and for each HH_α with $1 \le \alpha \le M$ (where M is the number of HHs), do:
- With probability p, report the true "from" Block;
- With probability $(1-p)$, report a Block whose number is taken according to the probability distribution dictated by the histogram $\mathrm{ht}_{\mathrm{from}}$.

Example: Suppose that $N = 5$ (five Blocks) and that

$$\mathrm{ht}_{\mathrm{from}} = [15, 10, 25, 5, 20]. \tag{19}$$

Then (in this sub-step of Step 1) the company will report that HH_α has taken its trip from:
- ∗ Block 1 — with probability $15/(15+10+25+5+20) = 15/75$;
- ∗ Block 2 — with probability $10/75$;
- ∗ Block 3 — with probability $25/75$;
- ∗ Block 4 — with probability $5/75$;
- ∗ Block 5 — with probability $20/75$.

Note 1: The are standard built-in algorithms of how one can draw entry i from a given probability distribution.

Note 2: There is no need to reconcile the "from" locations at this Step with anything from the previous trips, since this is assumed to be the first trip in a given report. Step 3 below will handle the issue of connection between two "adjacent" trips.

Note 3: It may be a good idea to update the histogram used for each α (the index of the HH) based on the drawings of the "from" Blocks by the HHs before the HH_α. Here is an Example. Suppose that for HH_1, the *reported* "from" Block is Block 2. (It does *not* matter whether this reported Block is the true or made-up one.) Now, for HH_2, the histogram to use will be:

$$\mathrm{ht}_{\mathrm{from}} = [15, \underline{9}, 25, 5, 20], \tag{20}$$

where the change compared to (19) is underlined. Then, if for HH_2 the reported Block ends up being Block 5, then for HH_3 one would use the histogram

$$\mathrm{ht}_{\mathrm{from}} = [15, 9, 25, 5, \underline{19}], \tag{21}$$

and so on.

Note 4: As was mentioned during the group discussions and also during the final presentation, it makes sense to use p close to 1 (i.e., near-true reporting) for Blocks that contain many HHs, but use a smaller p (i.e., more distorted reporting) for Blocks that are sparsely represented in the survey.

Step 2 Following exactly the same procedure, but using $\mathrm{ht}_{\mathrm{to}}$ instead of $\mathrm{ht}_{\mathrm{from}}$, report the Block where HH_α traveled *to* in their trip 1.

Step 3 For trip 2, it is assumed that its origin must coincide with the destination of trip 1 for each HH. Thus, there is no selection needed at this Step.

Note: If the above assumption is not made, then Step 3 repeats Step 1, where the histogram needs to be that updated after all 'trip 1's by all HHs, as illustrated in the Example in Note 3 of Step 1.

Step 4 Repeat Step 2 using the updated histogram $\mathrm{ht}_{\mathrm{to}}$, similarly to what is said in the Note for Step 3.

6.3.2. *Two more complex variations of the the simplest case.*

Reporting the joint histogram of Blocks traveled 'from' and 'to'. Suppose that the company needs its report to satisfy not only the two individual histograms as in the Simplest Case, but also satisfy (approximately, of course) the *joint* histogram

$$\mathrm{ht\_from-to}(i, j), \tag{22}$$

where 'i' and 'j' are the Block's numbers that HHs travel 'from' and 'to', respectively. An example of such a joint histogram would be a matrix:

$$\mathrm{ht}_{\mathrm{from-to}} = \begin{pmatrix} 5 & 2 & 4 & 3 & 1 \\ 1 & 2 & 2 & 1 & 3 \\ 5 & 6 & 2 & 8 & 4 \\ 0 & 1 & 1 & 1 & 2 \\ 4 & 3 & 4 & 2 & 7 \end{pmatrix}. \tag{23}$$

Here the entry (i, j) is the number of trips taken from Block i to Block j by all HHs combined. For example, there is a total of 8 trips from Block 3 to Block 4. Note that the numbers in the columns for each row add up to the numbers one finds in (19). Then the process of selecting Blocks to report uses the same conceptual Steps as in Section 6.3.1, but with different histograms in some of the Steps.

Specifically:

- The histogram in Step 1 remains unchanged. Note that, as mentioned a few lines above, $\mathrm{ht}_{\mathrm{from}}(i) = \sum_{j=1}^{N} \mathrm{ht}_{\mathrm{from-to}}(i, j)$.
- In Step 2, suppose that HH_α was reported to start from Block s. Then one needs to use

$$\mathrm{ht}_{\mathrm{to}}(j) = \mathrm{ht}_{\mathrm{from-to}}(s, j) \tag{24}$$

  For example, if $s = 4$, then for the example given by (23), one has $\mathrm{ht}_{\mathrm{to}} = [0, 1, 1, 1, 2]$.
- Step 3 is the same as that in Section 6.3.1.
- For Step 4, proceed using the updated histogram $\mathrm{ht}_{\mathrm{to}}$ computed from $\mathrm{ht}_{\mathrm{from-to}}$ as shown in (24).

Reporting the individual "from" and "to" histograms and the histogram for traveled distances. This is actually a special case of that considered in the previous subsection. Indeed, if one preserves the histogram of the relation how often someone travels from Block i to Block j, then one *automatically* preserves the histogram of traveled distances.

6.3.3. *What needs to be investigated further in this approach.* The first two of the suggestions listed below and addressing the question in the title of this subsection are fairly obvious; the third one is probably less so.

- How does parameter p (in Step 1 in Section 6.3.1) affect both privacy and utility of the data reported according to the approach proposed above? How does this depend on the size of the survey?
- Is there a (probably empirical) rule on how p should depend on the size of the Block to which it applies? (This may (?) also depend on the relation between the sizes of the "from" and "to" Blocks for which a given trip is reported.)
- The company may also want to be concerned about the *plausibility* of the reported trips, where the term 'plausibility' is explained by the following

  Example: Suppose that the adversary has a means to detect, and then filter out, outliers in the reported statistics. Such outliers may well come from the made-up trip data. If the adversary ignores them, then he/she/they can obtain a more truthful statistics and from it can infer the target.

  The ways to counteract this are two:
  - Make the made-up trips appear more like true ones (for example, in terms of being a closer match of the trip destinations);
    and
  - Create more made-up trips in such a way that it would make outliers no longer look like outliers and hence will make it harder for the adversary to detect. (This latter method is, of course, well-known from the "Ali Baba and the Forty Thieves" folktale.)

  The question of how these methods could be implemented remains open, to the knowledge of the author of these notes.

7. Author Contributions

- MAA advised the various sub-groups during the workshop and produced visualizations of the MSP area.
- PWF wrote the section on DP Analysis for Trip Data.
- DAE assisted BE with the code used in the rectangular-block work of the Home Location Perturbation section. He also wrote the algorithm that computed the proportion of blocks with no net gain/loss.
- LDG wrote a portion of the Data Collection section and Differential Privacy section, and worked on and wrote parts of the modified coin flip algorithm and the generalized probability algorithm.
- SKD wrote the subsection Graphical use of sharing differentially private information, under the section, Data Analysis. Helped create using SAS the Figure 9.

- GK contributed to the edits and literature review of the Background section and discussions on perturbation and aggregation simulations.
- ASC suggested some initial readings on the topic of differential privacy, helped with the literature review, and offered advice on implementation and limits of differential privacy throughout the workshop, but did not contribute to the report.
- YS contributed to the edits and took part in discussion of the basics of differential privacy and coin-flip examples.
- ZS, SW and LY wrote code to visualise: trip information such as: frequency, trajectories and destination distribution. Also took part in discussion of making categorical data differentially private.
- TAD produced data summaries, participated in discussions about and explorations of the implications of DP for location data, and assisted with the use of the coin-flip algorithm on car makes/models.
- CRM developed model for studying effects of perturbation before aggregation of data points in general polygons.
- EG wrote code in R to implement the modified coin flip algorithm to perturb vehicle makes and models, and assisted in writing and editing section 4.1.
- TL provided details of a potential location replacement algorithm in the Future Work.

References

[Buk19] Preston Bukaty. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, Ely, 2019.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 2014.

[Dun15] Ted Dunning. *Sharing big data safely : managing data security*. O'Reilly, Sebastopol, CA, first edition. edition, 2015.

[GBM15] Jeffrey Gonder, Evan Burton, and Elaine Murakami. Archiving data from new survey technologies: Enabling research with high-precision data while preserving participant privacy. *Transportation Research Procedia*, 11:85–97, 2015. Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia.

[Glo21] GlobeNewsWire. Electric vehicles market share projected to reach usd 700 billion by 2026: Facts and factors. *Facts and Factors*, 2021.

[gov21] Privacy rights and data collection in a digital economy : hearing before the committee on banking, housing, and urban affairs, united states senate, one hundred sixteenth congress, first session, on evaluating current approaches to data privacy regulation, including the european union's general data protection regulation, and its application to financial institutions, may 7, 2019., 2021.

[GV16] W Gregory Voss. European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. *The Business lawyer*, 72(1):221–234, 2016.

[HG05] Baik Hoh and M. Gruteser. Protecting location privacy through path confusion. In *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*, pages 194–205, 2005.

[Hol16] Naoise Holohan. *Mathematical Foundations of differential privacy*. PhD thesis, School of Computer Science and Statistics, Trinity College, The University of Dublin, 9 2016.

[JSB+13] Kaifeng Jiang, Dongxu Shao, Stéphane Bressan, Thomas Kister, and Kian-Lee Tan. Publishing trajectories with differential privacy guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, pages 1–12, 2013.

[Kur21] Christoph Kurz. Understanding differential privacy. *Significance*, 18(3):24–27, 2021.

[LDGE19] 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.

[McG18] Nancy A. McGuckin. Summary of travel trends : 2017 national household travel survey, 2018.

[McN11] Brian McNeil. Report: Netflix sued over possible breach of federal privacy law. *SNL Kagan Media & Communications Report*, 2011.

[mtc12] The metropolitan transportation commission backs california traveler survey. *Travel & Leisure Close-Up*, 2012.

[NWB+14] Philippe Nitsche, Peter Widhalm, Simon Breuss, Norbert Brändle, and Peter Maurer. Supporting large-scale travel surveys with smartphones – a practical approach. *Transportation Research Part C: Emerging Technologies*, 43:212–221, 2014. Special Issue with Selected Papers from Transport Research Arena.

[Tea17] ITGP Privacy Team. *EU General Data Protection Regulation (GDPR) : an implementation and compliance guide*. IT Governance Publishing, Ely, 2nd ed edition, 2017.

[VS15] Akshay Vij and K. Shankari. When is big data big enough? implications of using gps-based surveys for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, 56:446–462, 2015.