Modeling the Tradeoffs in Publishing Urban Travel Data Sets
Graduate Student Mathematical Modeling Camp 2022

Manuchehr Aminian[1], Darcy Brunk[2], Derek Drumm[3], Vivian J. Goshashy[4], Alanna Haslam[5], Katherine Johnston[6], Johnathan Makar[7], Hannah Scanlon[8], and Kriti Sehgal[9]

[1]Department of Mathematics and Statistics, Cal Poly Pomona
[2]Department of Mathematics, San Francisco State University
[3]Department of Mathematics, Worcester Polytechnic Institute
[4]Department of Mathematics, Eastern Washington University
[5]Department of Mathematics and Statistics, Boston University
[6]Department of Applied Mathematics, University of Washington
[7]Department of Mathematics, East Stroudsburg University
[8]Department of Mathematics, Duke University
[9]Department of Mathematics, The Ohio State University

June 2022

Abstract

Geospatial travel data is enormously valuable for local governments to make informed decisions on urban planning, zoning, fiscal policy, and so on. When governments collect such data using taxpayer money, there are expectations for (a) transparency in data and data collection, (b) privacy of participants involved in data collection, (c) utility of data to make informed decisions. These three priorities are often at odds with one another. The project sponsor wants us to mathematically model these trade offs, suggest approaches to privacy of individuals, and do a risk analysis of a hypothetical malicious actor seeking to discover details of individuals in the data. The sponsor would like these applied to data from a large travel data set from the Minneapolis/St. Paul metropolitan area during 2018-2019.

Contents

List of Figures

1 Introduction

1.1 Data Summary

The data used for this project was collected via surveys conducted by call centers, through smartphone applications or online, by the Minnesota Metropolitan Council. The purpose of the data collection was assisting "Twin Cities Metropolitan Planning Organization with transportation planning and project programming in the Twin Cities region" [7]. Smartphone participants submitted 7 days of travel information, while online and call center participants submitted 1 day of travel data. All participants were presented the same set of questions, which were available in English, Spanish, Karen, Oromo, Somali, and Hmong. The data was collected from October 1, 2018 to September 30, 2019. The data can be accessed via the Minnesota Geospacial Commons website, [7].

The data has been divided into categories and sorted into spreadsheets accordingly. The categories are: household, person, vehicle, day, and trip. The 'household' data includes the following information about each participating household: number of people, income, rent, home type, location, number of vehicles, duration in location, whether renter vs. homeowner, and the start and end dates of participation in the survey. The 'person' data consists of each participants age, education level, gender, whether licensed or not, ethnicity, mode of transportation, household, and a personal ID number. The vehicle information collected includes the make, year, fuel type, whether owned vs. leased, and whether a toll transponder is associated. The 'day' data details the number of trips each person took each day along with their delivery habits, sorted by household or individual. It also describes why a participant did not take a trip on a given day. The 'trip' data includes information on mode of transport, duration of trip, origination and destination, and the purpose of the trip.

1.2 Questions About Data Set

During the beginning stages of analyzing the data, several categories of questions were defined: first order, second order, and complex. A first order question can be answered by statistical analysis of one column of one category of data. A second order question can be answered by statistical analysis of one column of one category of data, stratified by a second column from the same category of data. A complex question requires combining more than one category of data and then performing the analysis.

After performing an exploratory analysis on all the spreadsheets both individually and in combination, it was clear that a variety of questions could be answered with the available data. After some discussion, the following list of questions was selected:

- First order statistics questions:
  - What is the household income distribution?
  - How many trips do people take per day?
- Second order statistics questions:
  - What is the distribution of the modes of transport used by the day of the week? For example, we expected that people would drive their own vehicle more on the weekends and take public transport during the week.
  - How does the working hours per week compare to education level?
- Complex questions:
  - What are the demographics of the people based on the mode of transport they use?
  - How does the household income compare to the number of vehicles owned by the household?

In order to choose these questions, a large list of questions was compiled and sorted by question type. We then selected questions which provided a variety of information from the data and required a variety of statistical analyses.

1.3 Utility vs Privacy

Following the questions created about the data sets, one may wonder how to handle utility versus privacy of the information. To begin to tackle this dilemma, utility and privacy must be defined. In this project, utility describes how consistently statistical analysis and conclusions of a data set align with the truth. Privacy describes how well participants' sensitive data and identifying information is obscured and protected from recreation. With utility and privacy defined, the goal of the project was to save as much possible utility of the data while also making the data private enough that those who chose to take the survey cannot be singled out and tracked. This was a challenge in and of itself as across the five data sets supplied, there were many ways to easily cross reference a person or household and find out more than enough information to allow for malice (assuming the data provided was not previously randomized). This concept of utility versus privacy served as a main motivator for this project.

1.4 Malicious Actors

The survey data provided has removed personally identifiable information (PII) from the original survey, such as names, phone numbers, addresses, etc. Despite this, the presentation of the data as given still allows for potential malicious activity. From the demographic analysis, we observed mode of transportation used by various gender, age, and ethnicity (Figures 6, 7, and **??** respectively). The data also contains information on how many trips participants made per day, whether the participants are employed, and how many hours each participant works per week (see Section 1.1). Along with the county and survey tract data, a malicious actor could use this information to target specific groups or individuals. The following paragraphs describe some methods by which malicious actors could make use of the current data to breach the privacy of the survey participants.

Consider an individual or group attempting to make scam phone calls. These sorts of groups tend to target the elderly, as they have a higher likelihood falling for phone scams. The phone scammers could make use of the county and survey tract data by determining which counties and survey blocks contain the highest proportion of elderly individuals. Knowing which counties contain the highest proportion of the elderly, the phone scammers can then restrict themselves to only calling phone numbers with the area code corresponding to these counties.

Consider another individual who wishes to locate the home of a survey participant based on this data. As with the phone scammers, this individual could directly locate the county and survey block from the data, but alone this information does not help the individual locate the corresponding home address. However, this individual could make use of the income bracket data, vehicle information, and average number of trips of the survey takers. With this data along with the county and survey block data, the malicious individual could make an accurate guess as to the home of the survey taker. For example, the data can be filtered to target elderly couples with an annual income greater than $250,000. As well, the malicious actor can use the FIPS code of the participant (a code similar to a ZIP code, which gives information on the state, county, and survey tract location) to determine a local region where the participant lives (See Figure 1) This information, along with the participant's number of vehicles, type of vehicles, type of home, and work week hours, could be used to determine to accurately determine where the participant lives.

These malicious acts show the importance of prioritizing privacy of survey participants, and motivate the development of methods to modify collected data in ways that removes the most sensitive data of the participants, while keeping important statistics the same.

2 Data Analysis

2.1 First Order Statistics

The first statistic of interest from the trip data set was the distribution of trips taken by the day of the week. This statistic would demonstrate any skew for more or less popular travel days. To compute this statistic, the trip data was subgrouped by the value of the "TRAVEL_DOW" column and the length of the data frame for each subgroup was recorded. After rearranging the data from alphabetical to the order of days of the week, the result was visualized in a histogram (Figure 2). This visualization shows no extreme preferences

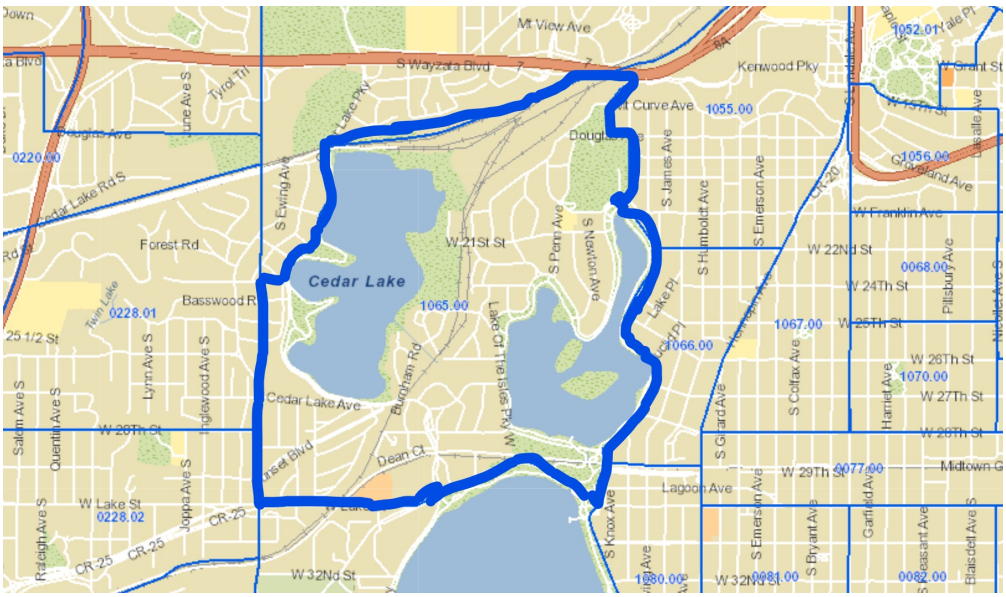


Figure 1: Map of region local to FIPS code 27053-106500-2. The geographic region has been manually highlighted for emphasis.

for any day of the week with Tuesday, Wednesday, Thursday and Friday being the most popular travel days while Monday and Sunday were the least popular.

Another first order statistic was the distribution of income of each household. A distribution plot is shown in Figure (3), where it can been seen that income has a uni modal peaked distribution with a significant number of households declining to answer.

2.2 Second Order Statistics

Building on the previous first order statistic of trips by day of week, the modes of transportation taken for each trip were examined. The hypothesis was that public transportation, such as buses, would constitute a larger portion of trips taken during the work week as compared to the trips take on the weekend. To compute this statistic, the trips per each day of the week were subgrouped by the data in "MODE_TYPE" and visualized it in a stacked histogram (Figure 4). Overall, this visualization highlighted that household vehicles were the most commonly used mode of transportation. Also, the hypothesis that fewer trips would be taken by public bus on the weekend as compared to during the work week can be seen by the smaller light pink section on the Saturday and Sunday bars in Figure 4.

For the working data from the persons data set, our exploratory analysis suggested correlations between the education status, working hours, number of jobs, etc. In figure 5, we see how the number of working hours vary across the education level. The education levels we considered whether they finished high school, have graduate or postgraduate degree. The category 'Other' refers to "Some college/ Associate degree/Vocational/technical training".

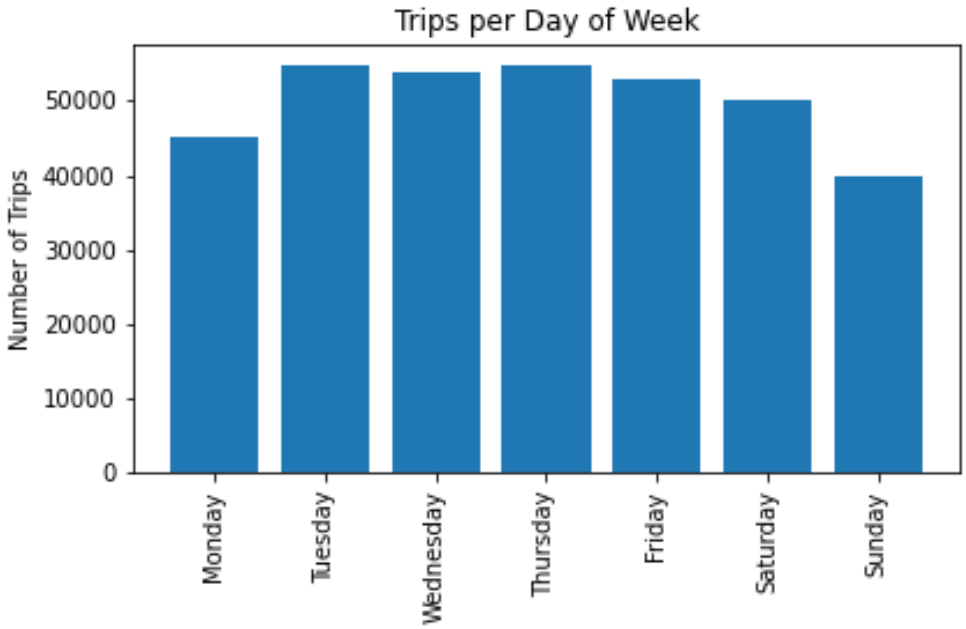


Figure 2: Number of trips recorded in the data set plotted against the day of the week the trip was taken

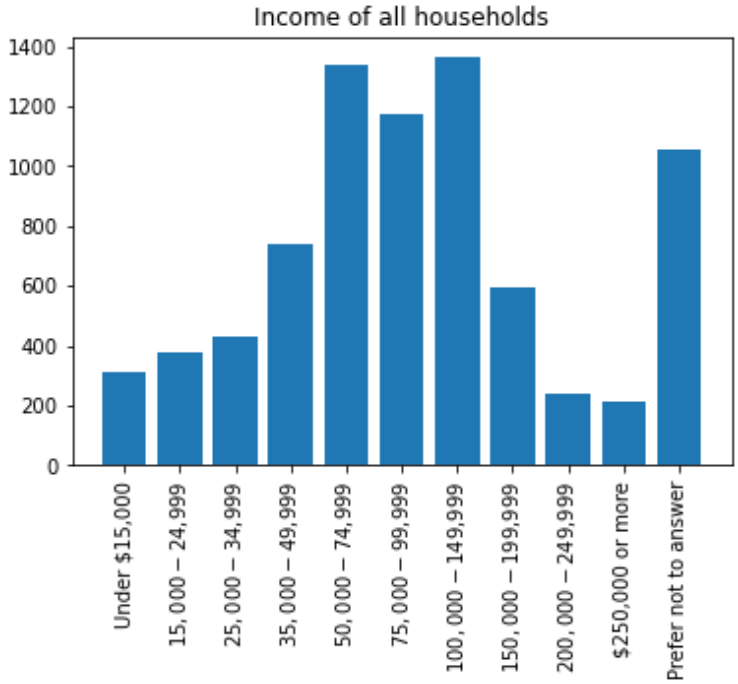


Figure 3: Income distributions of each household

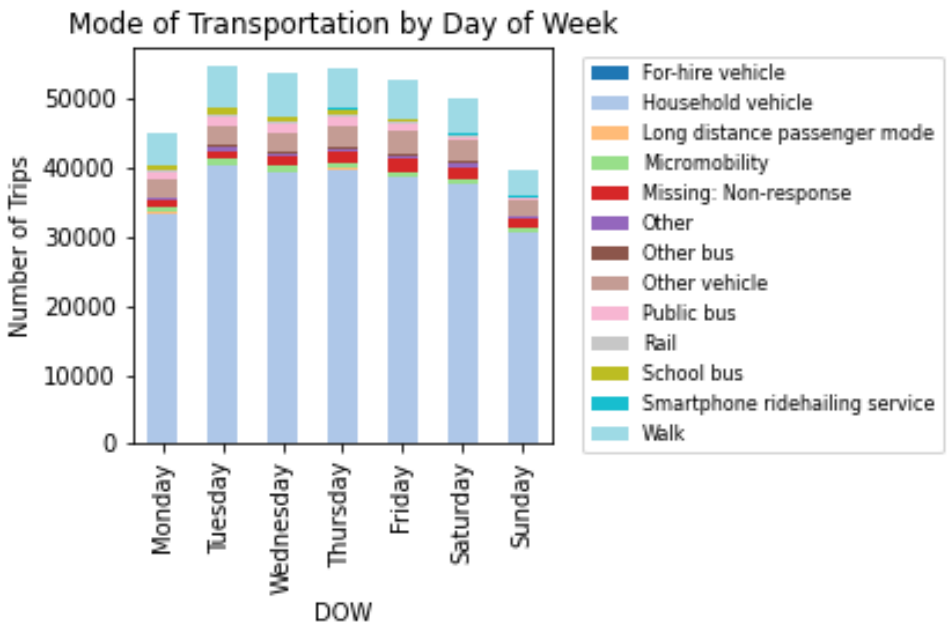


Figure 4: Number of trips recorded in data set plotted against day of the week and grouped by the type of transportation used

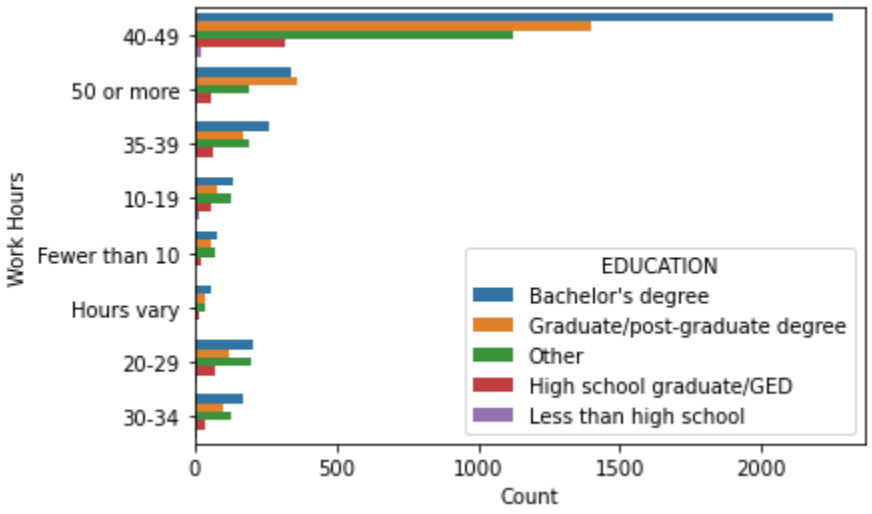


Figure 5: Number of working hours per week compared to the education level

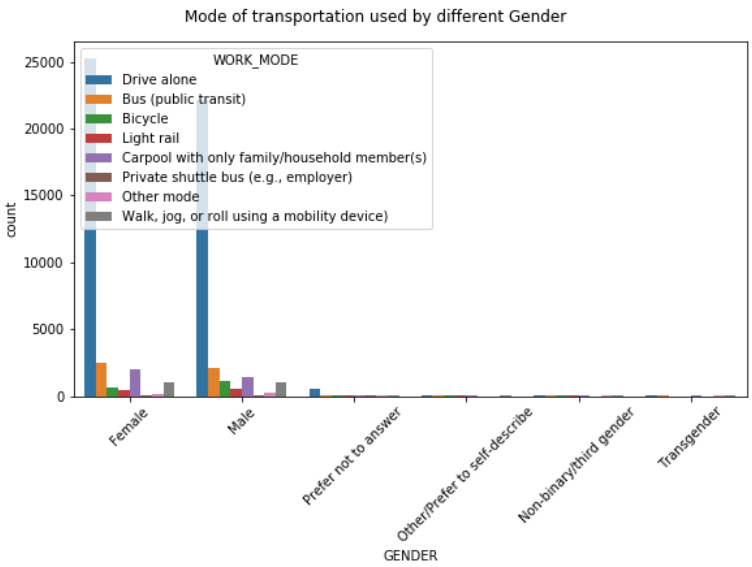


Figure 6: Mode of transportation used by different gender.

2.3 Complex Statistics

Finally, more complex statistics were considered using columns across data tables. For example, the demographic of people in relation to the mode of transportation have questions that span multiple tables. Demography includes gender, age, and ethnicity, and modes of transportation include driving alone, Bus (public transport), light rail, bicycle, and private shuttle. As one can observe in the provided figures, various modes of transportation are used by different demographics. For instance, in Figure (6), it is observed that most male and females prefers to drive alone and very few use the private shuttle bus, as compared with non-binary and other genders. Further, considering age group in Figure (7), it is observed that people between the age of 35 - 44 prefer to drive alone as compared with other age groups. Additionally, the age range 18 - 24 has less people who prefer to drive alone. The second preferred mode of transport of all ages is the public transit.

Another more complicated question was the distribution of cars by income. Because of the way the data tables are formatted, this is a complicated question because the household table is 1 row per household while the vehicle table is 1 row per vehicle. Thus, when the data is changed for privacy purposes, the results of this question may change in unexpected ways. The results of this analysis can be seen in Figure (8).

3 Randomization

3.1 Differential Privacy

Differential privacy is a mathematical framework to describe how well a randomization algorithm privatizes data. A randomization algorithm is a process by which randomization is introduced to a data set in order to provide security to participants. There is a process by which to calculate the degree to which a synthetic data set is differentialy private, however we did not have time to understand and implement this calculation for our data. For more information see [1].

A basic method of introducing randomized responses in the data has been constructed in the social sciences. This method, called *randomized response*, determines whether to collect the true answer or a random answer from a participant about having a property P. Participants are asked whether or not they have property P and report their answers according to the following process, as defined in [1] and shown in figure 9:

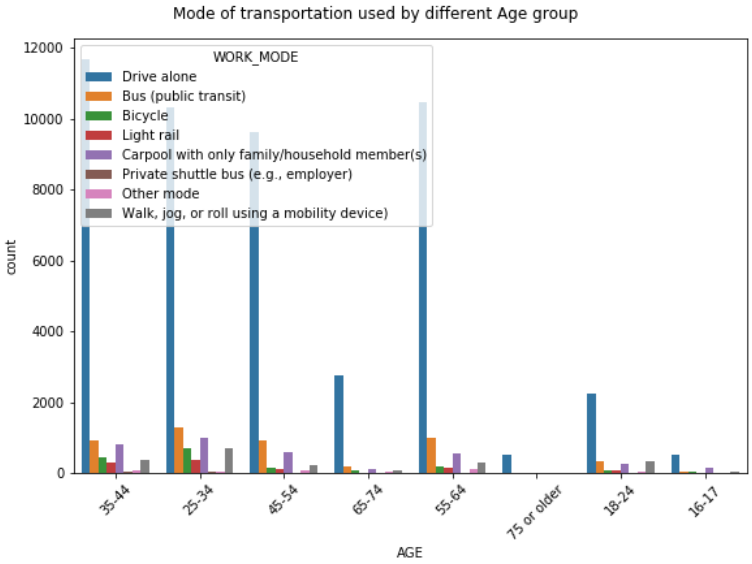


Figure 7: Mode of transportation used by different age.

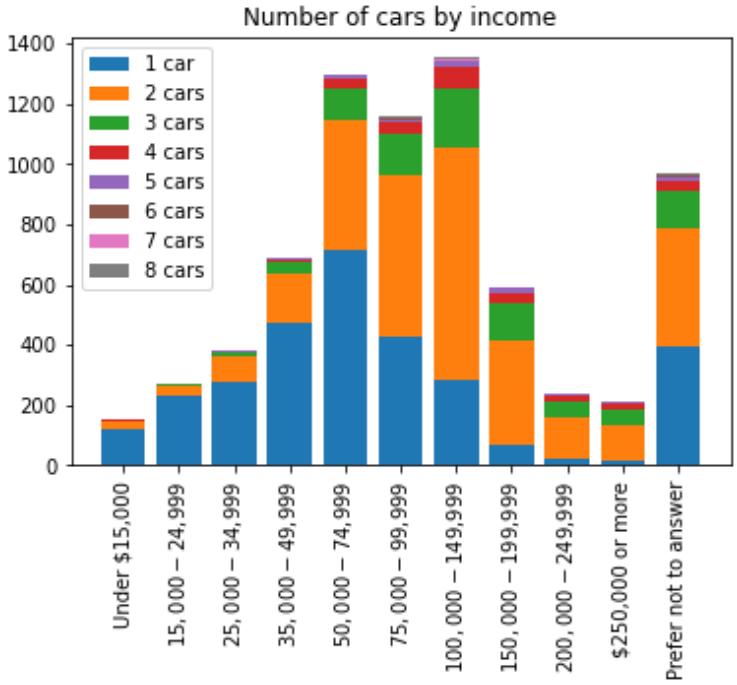


Figure 8: Income vs number of cars in household.

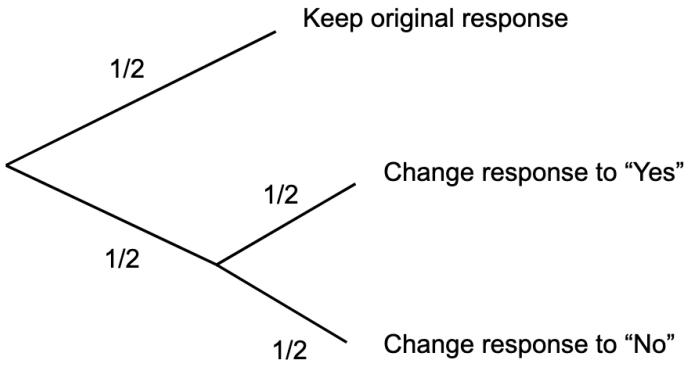


Figure 9: Process for randomizing responses

1. Flip a coin.
2. If **tails**, give a truthful answer (either “Yes” or “No”).
3. If **heads**, flip another coin. If heads, respond “Yes”; if tails respond “No”.

This method creates privacy by introducing a level of “plausible deniability.” That is, even if a participant’s answer is recorded as “Yes” for an illegal or embarrassing activity, they cannot be implicated because there is at least a $1/4$ probability of a “Yes” answer regardless of whether or not a participant actually has property P. Determining the level of accuracy preserved by this method requires an examination of the procedure by which noise, or randomized “Yes” and “No” answers, are generated. The expected number of “Yes” answers can be calculated by $1/4$ times the number of participants who do not have property P plus $3/4$ the number having property P [1]. The true portion of participants who have property P is defined as p. Then p is estimated by twice the fraction answering “Yes” minus $1/2$, that is, $2((1/4) + p/2) - 1/2$.

3.2 Variations

In practice, the above coin flip method is implemented as as way of randomizing the original data set; it generates a synthetic version of the data. Given a recorded answer in a column of the data, this coin flip method can be used to randomize the response, either keeping the same or changing it by sampling from a probability distribution of all possible responses. The procedure described in the previous section can be varied in three primary ways.

First, allow the first coin flip to be weighted. This means that the same recorded answer is kept with probability w (see Figure 10). In a sense, this parameter w can act as a proxy for privacy. When $w = 1$, the data set will be completely the same. When $w = 0$, every recorded answer will be replaced with a randomly sampled response. However, it can cannot be assumed that w gives the probability of a single data point being the same as the true recorded answer. The probability is instead a different value, dependent on w,

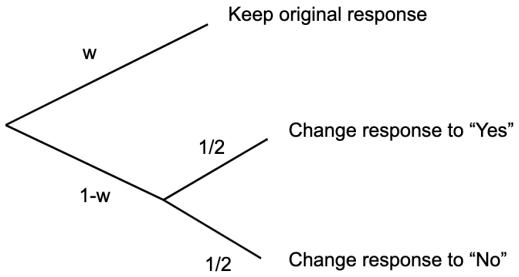


Figure 10: First variation of the coin flip method for data randomization. Weight w specifies the probability of leaving a data point unchanged.

that can be calculated:

$$
\begin{aligned}
&\mathbb{P}(\text{Synthetic data point matches original}) \\
&\qquad = \mathbb{P}(\text{Response unchanged}) + \mathbb{P}(\text{Response changed but matches original response}) \\
&\qquad = w + (1-w)\mathbb{P}(\text{New response matches old response}|\text{response changed}) \\
&\qquad = w + (1-w)\big(\mathbb{P}(\text{New response is “Yes” and old response was “Yes”}|\text{response changed}))+ \\
&\qquad\qquad \mathbb{P}(\text{New response is “No” and old response was “No”}|\text{response changed}))\big) \\
&\qquad = w + (1-w)\big(\mathbb{P}(\text{New response is “Yes”}|\text{response changed})\mathbb{P}(\text{Old response was “Yes”})+ \\
&\qquad\qquad \mathbb{P}(\text{New response is “No”}|\text{response changed})\mathbb{P}(\text{Old response was “No”})\big) \\
&\qquad = w + \left(\frac{1-w}{2}\right)\big(\mathbb{P}(\text{Old response was “Yes”}) + \mathbb{P}(\text{Old response was “No”})\big) \\
&\qquad = w + \frac{1-w}{2}
\end{aligned}
$$

Therefore if $w = 0.5$, there is a $\frac{3}{4}$ probability that any given data point in the synthetic data will be the same as the original data point.

In the second variation of the coin flip randomization method, more than one response category is allowed (see Figure 11). As a pure extension of the first variation, uniformly pick one of the n categories with probability $1 - w$. In this case, the above calculation becomes:

$$
\begin{aligned}
&\mathbb{P}(\text{Synthetic data point matches original}) \\
&\qquad = \mathbb{P}(\text{Response unchanged}) + \mathbb{P}(\text{Response changed but matches original response}) \\
&\qquad = w + (1-w)\mathbb{P}(\text{New response matches old response}|\text{response changed}) \\
&\qquad = w + (1-w)\Big(\sum_{k=1}^{n}\mathbb{P}(\text{New is k-th category}|\text{response changed})\mathbb{P}(\text{Old was k-th category})\Big) \\
&\qquad = w + \frac{1-w}{n}
\end{aligned}
$$

The third variation of the method, instead of sampling from a uniform random distribution of the possible response categories, sample from the empirical distribution. Therefore with probability $1 - w$, choose a random data point from the original data set and use it to replace the data point in question. In this case, the probability of any given data point in the synthesized data being identical to the original data point is

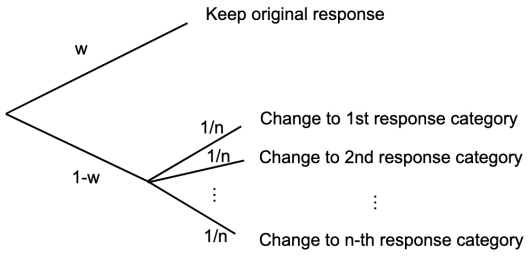


Figure 11: Second variation of the coin flip method for data randomization. Weight w still specifies the probability of leaving a data point unchanged. There are now n possible response categories. When a data point is changed it is replaced by a category that is selected uniformly at random.

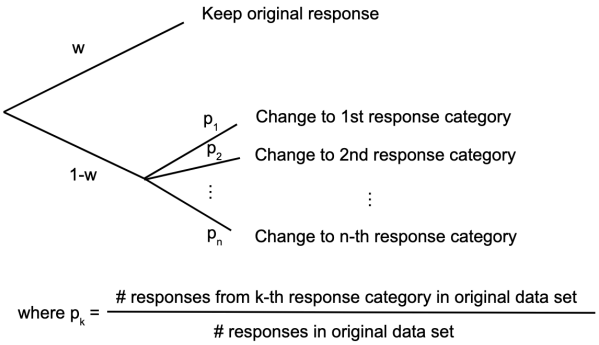


Figure 12: Third variation of the coin flip method for randomization. When a data point is changed it is replaced by a category that is selected from the empirical distribution.

given by:

$$\begin{aligned}
&\mathbb{P}(\text{Synthetic data point matches original}) \\
&\qquad = \mathbb{P}(\text{Response unchanged}) + \mathbb{P}(\text{Response changed but matches original response}) \\
&\qquad = w + (1-w)\Big(\sum_{k=1}^{n}\Big)\mathbb{P}(\text{New response matches old response}|\text{response changed}) \\
&\qquad = w + (1-w)\Big(\sum_{k=1}^{n}\mathbb{P}(\text{New is k-th category}|\text{response changed})\mathbb{P}(\text{Old was k-th category})\Big) \\
&\qquad = w + (1-w)\Big(\sum_{k=1}^{n}\mathbb{P}(\text{Old was k-th category})^2\Big)
\end{aligned}$$

Then observe that w can still act as a control parameter for the level of privacy, but the degree to which the pointwise data changes will depend on the empirical data. It will later be seen how well synthetic data obtained via empirical sampling preserves the distribution of the original data.

3.3 Other Methods

Differential Privacy is a simple and powerful tool for providing privacy to the data. However, there are limitations. For example, it is hard to tell what parameter w is needed to provide sufficient privacy of the data, and based on the parameters of the data, sometimes, w needs to be so large that the utility of the data is not useful. [2]

To this end, others have explored other methods of data privacy. Gambs et. al propose a method based on differential privacy that uses vine copulas to help preserve correlations between different columns of the data sets [3]. Jiang et. al. also propose a method to generate synthetic data sets based on a called multiple imputation (MI) framework [4]. MI is a strategy typically used to fill in missing data. Jiang and others used this strategy to completely general a new data set similar to the original data, and then added a masking step to further add privacy from the data. This method generates a completely new data set, so it preserves the privacy of the participants. Yet, this new data set is based off the data in a way that correlations between variables are preserved.

4 Utility vs Privacy

4.1 Metrics

Because we are interested in preserving the utility of a data set, we want a way of measuring how close the distribution of the synthesized data set is to the original distribution of the data. One common statistical distance that is used in information theory, machine learning, and Bayesian inference is the *Kullback-Leibler divergence*, referred to as the *KL-divergence* [5],[6]. Intuitively, it can be understood as the average amount of information that is lost by using the synthetic data rather than the authentic data.

For discrete probability distributions (which we restrict to), the KL-divergence is computed as follows:

$$D_{KL}(P_{\text{true}}, P_{\text{synth}}) = \sum_{k=1}^{n} P_{\text{true}}(k) \log \left(\frac{P_{\text{true}}(k)}{P_{\text{synth}}(k)} \right)$$

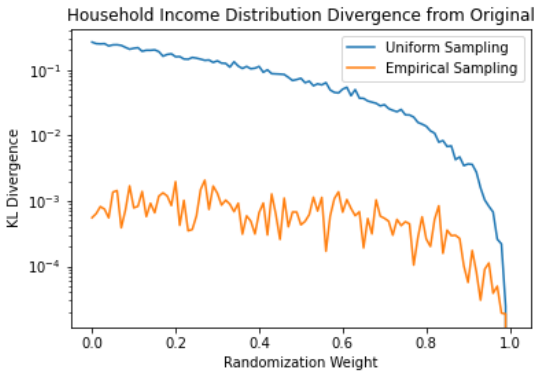
where $P_{\text{true}}(k)$ and $P_{\text{synth}}(k)$ are the probabilities of a given data point from the original and synthetic data sets, respectively, being in category k. Therefore KL-divergence the expected logarithmic difference between the probability distributions with respect to the original probability distribution.

For more information about this measure, see [5] and [6].

4.2 Results

In order to analyze the trade-off between privacy and utility, the original data for various statistics was compared to randomized data using a uniform sampling method as well as an empirical sampling method. First, the distribution of household income was analyzed. Figure 13 (a) plots the KL divergence of the distributions created by both sampling methods for a range of randomization weights. The results are plotted on a log scale for visualization purposes. Note that a randomization weight of 1 corresponds to maintaining all of the original data. Since the KL divergence measures the change in the distribution between two data sets, the KL value for data sets randomized with weight 1 (aka not randomized) has a KL Divergence of 0. From that baseline, a higher KL Divergence value corresponds to a greater difference in distribution from the original data set. Since the empirical sampling method selects randomized data points from the same distribution as the initial data set, the KL-divergence for these sets does not have a consistent trend and are much lower than those for the uniform sampling. For the uniform sampling method, the KL divergence values increase by orders of magnitude as the weight decreases indicating the new data set that poorly represents the original data distribution. With either sampling method, a high enough weight could protect the distribution of the original data set, but these results show the empirical sampling method could be used to largely maintain the original sampling distribution at any randomization weight. Figure 13 (b) plots the ratio of data entries which were unchanged from the original data set for both sampling methods. This figure is used as a check of the method. As expected, all of the original values are maintained when the randomization weight is 1. When the randomization weight is 0, a low proportion of randomized values will be equivalent to the original values by chance with that proportion being slightly large in the empirical sampling which matches the original distribution than the uniform sampling. Between these two extremes, the ratio maintained increases linearly as expected.

Since the KL divergence is a relatively abstract measurement, additional histograms were plotted in Figure 14 to visually compare the distributions of the original, uniformly randomized and empirically randomized data sets for several different weight values. For a weight value of 0.1 in Figure 14 (a), the empirical data



(a) KL Divergence values for House Hold Income Distributions with ranging randomization weight used in Uniform (blue) and Empirical (orange) sampling methods

(b) Ratio of Entries consistent with original data for ranging randomization weight with both Uniform(blue) and Empirical (orange) sampling methods. Dashed line plots line representing where weight would be equal to ratio maintained

Figure 13: Utility vs Privacy Comparisons for Household Income Data Randomization

set matches the true data set very closely. However, the uniformly sampled data set has an almost uniform distribution which is markedly different from the original. As the weight increases, the uniformly randomized set more closely matches the true data set. However, the empirical data set does not appear to improve in distribution accuracy with higher weight given that the method maintains the original distribution with all randomized values. This suggests that a high weight may be required for the uniform randomization method to maintain the utility of the data set as measured by the distribution of this variable while a low weight could be used with the empirical randomization to increase the privacy of the information without disrupting the targeted data distribution.

While the above analysis demonstrates the effectiveness of the empirical method at protecting both privacy and utility for the distribution of a single variable, it is important to consider higher order statistics. While the empirical method can maintain the distribution of the variable randomized, it may skew the distribution of a second order statistic that depends on the randomized variable and another variable. To understand the utility in this context, we considered the distribution of trips taken over days of the week and subcategorizatized by mode of transportation after randomizing the day of week. Figure 15 (a) plots the KL divergence against randomization weight for the trip by day of week distribution (solid) and the trip by mode of transportation per day of week distribution (dashed). The divergence is computed for both the uniform randomization method (blue) and the empirical randomization method (orange). Since the day of week distribution is a first order statistic, the KL divergence plots for these statistics closely resemble that of the household income distribution shown in Figure 13 (a). However, an important difference to note is the scale of the value for KL divergence values. The maximum value for the household income KL divergence was 10 times larger than that of the day of week KL divergence. This is likely because the trip per day of week data is close to uniformly distributed while the household income distribution is much less consistent. This observation demonstrates that the utility of the data set after randomization may depend on the distribution of the original data. Figure 16 plots several histograms of the data distributions for the true, uniform, and empirical data sets for various weights. The same patterns as seen in the histograms for the household income are seen here, although the difference in the uniform distribution to the true data is less extreme for low weight values since the true data is more uniform in distribution. Figure 15 (b) again shows the ratio of values maintained compared to the randomization weight. The only notable difference in this plot compared to that from the household income distribution is that the values for the uniform and empirical data sets are roughly equal. This is also due to the fact that the original data is almost uniformly distributed so the uniform method keeps roughly the same number of values equal to the original. However, the KL divergence still demonstrates the relative utility of the empirical method as compared to the uniform

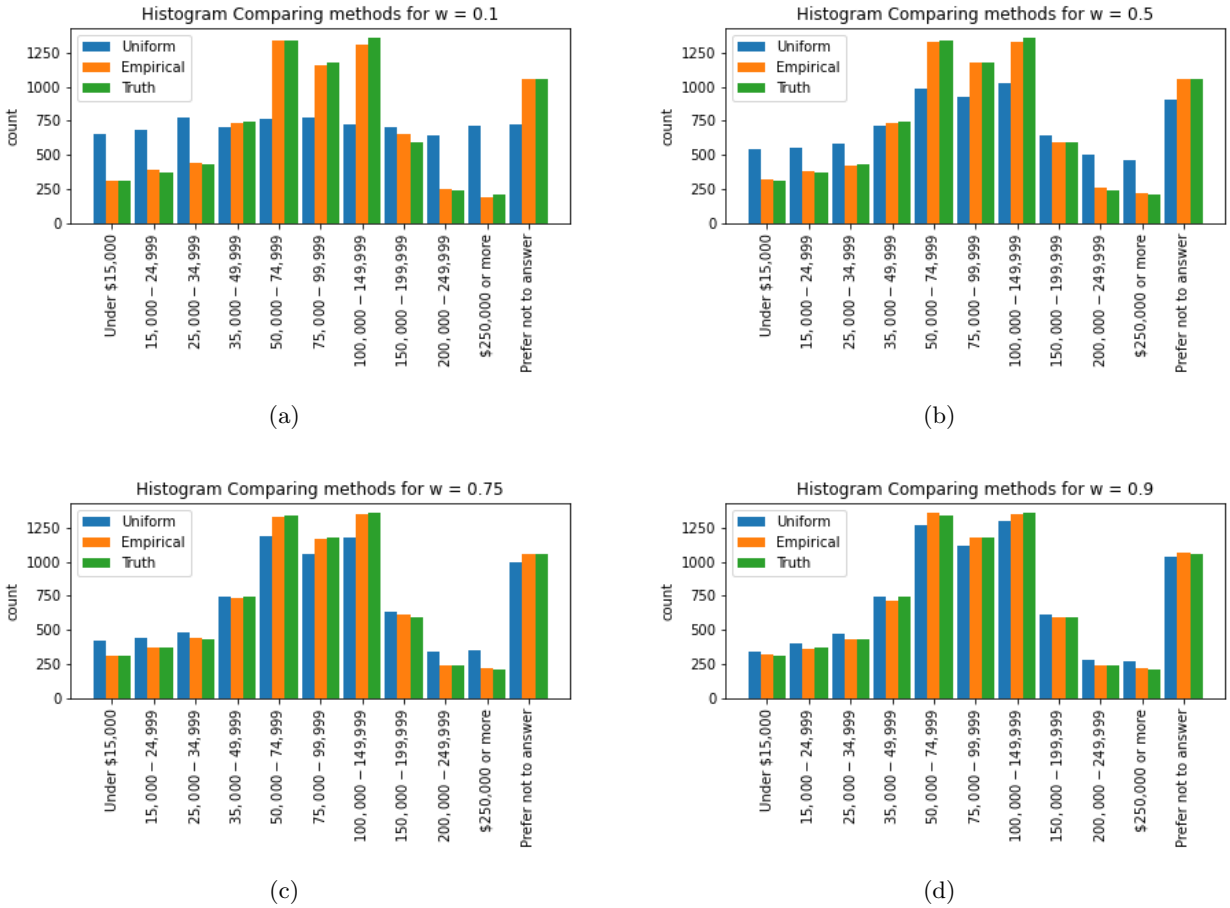
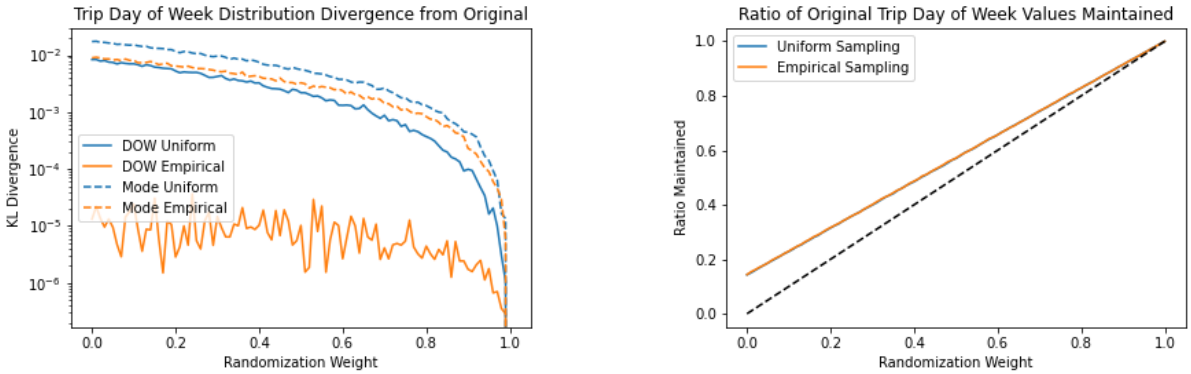


Figure 14: Household Income Distribution for Uniform (blue) and Empirical (orange) data randomization techniques compared to the Truth(Green) from the original data for randomization weight values of (a) 0.1, (b) 0.5, (c) 0.75, (d) 0.9



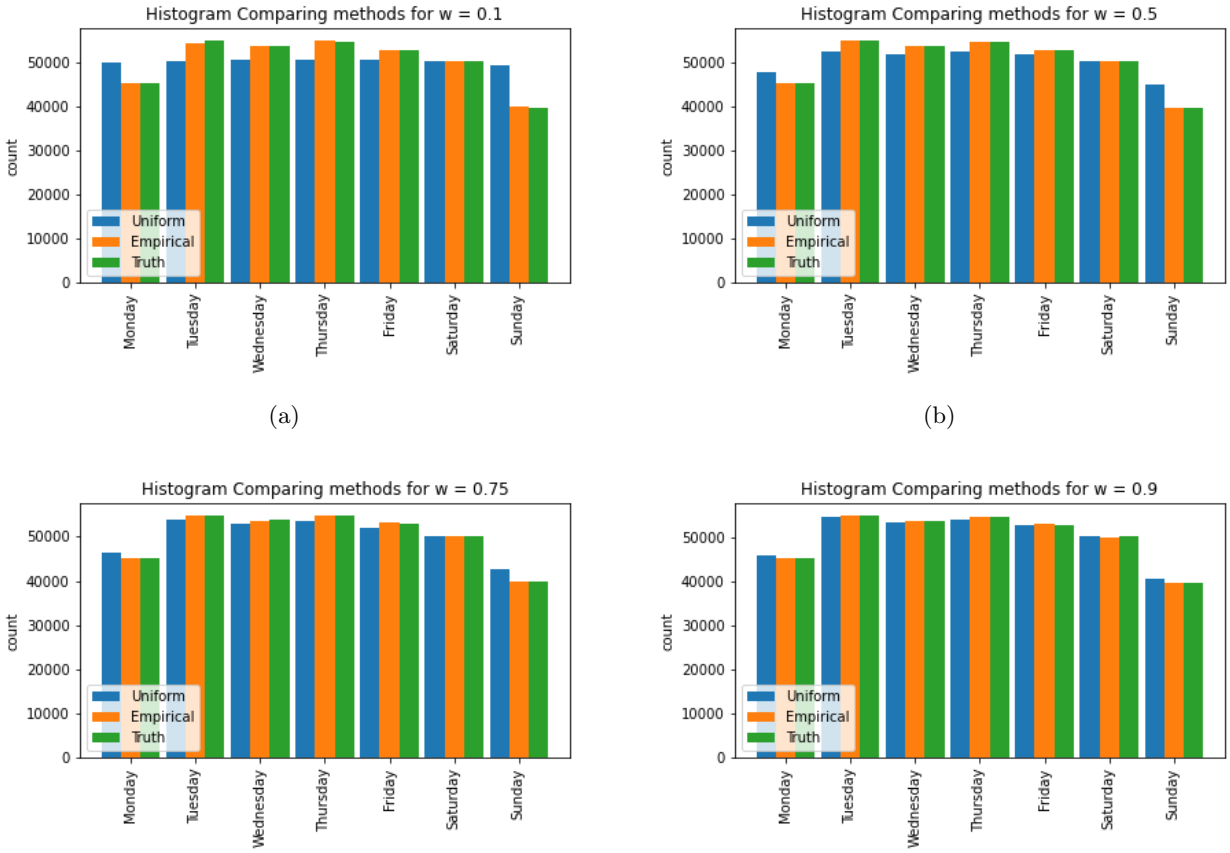
(a) (b)

Figure 15: Utility vs Privacy Comparisons for Trips by Day of Week and by Mode of Transportation Data Randomization

method since it is a more holistic comparison of distributions.

For the mode of transportation per day of week distribution, the KL divergence computed is similar for both empirical and uniform randomization methods Figure 15). This is the first time the empirical method has been seen to depend significantly on weight and not be markedly better than the uniform method. This is a result of the mode of transportation distribution being a second order statistic that depends first on the day of week of the trip and then the mode of transportation used. These results are consistent with what one might expect since the randomization method was only implemented on a single variable. Additionally, implementing the same randomization method on both variables in question would not protect the distribution of the second order statistic since the distributions are dependent on one another. A randomization technique that takes into account the connects between all variables of interest would be necessary to protect the utility of this data for lower randomization weights. It may also be possible to choose a higher randomization weight to protect the utility of the data while introducing some privacy, but the appropriate cut-off for this value that will balance both desired attributes is unclear.

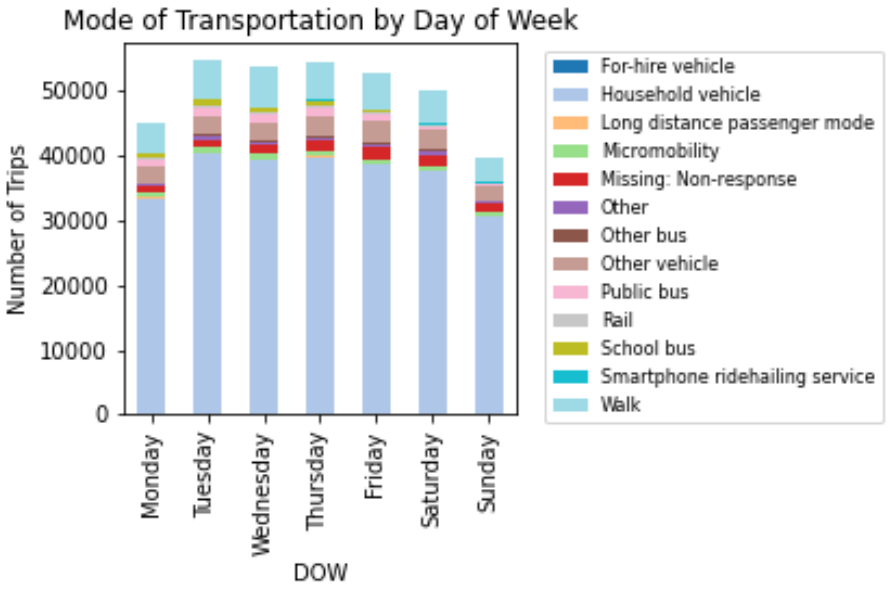
More detailed visualization of the changes in this second order statistic is challenging given the number of categories involved. However, Figure 17 plots an example of each version of this statistic. Figure 17 (a) shows the original data while plot (b) plots the uniform randomization method and (c) plots the empirical randomization method for a weight of 0.5. As expected, the height of each bar is relatively consistent between the empirical method and the original data while the heights of the bars are closer to uniform than the original for the uniform method. However, within each bar, the distribution of each mode of transportation has changed. If one were to plot just the distribution of modes of transportation not dependent on the day of the week, they would be the same for each method since these values were not randomized. However, the distributions relative to the day of the week were not maintained given the randomization done on the day of week variable. This analysis shows decrease in utility of second order statistics when the randomization technique is applied to a single statistic. The same analysis could be generalized to more complex statistics as well.



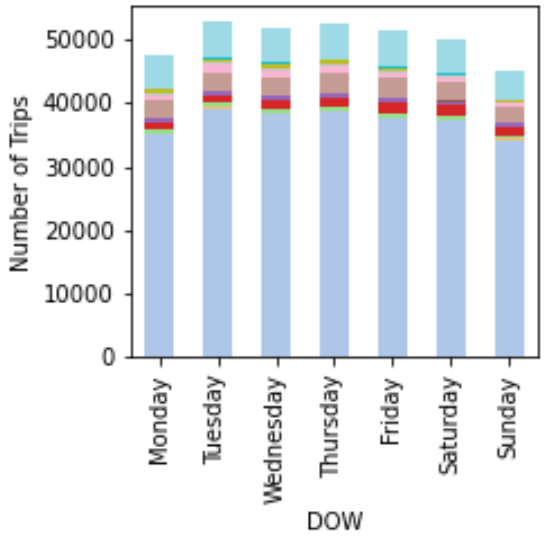
(a) (b)

(c) (d)

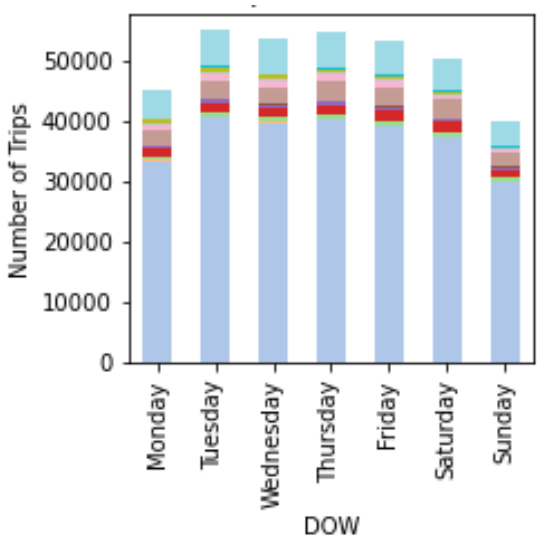
Figure 16: Trips by Day of Week Distribution for Uniform (blue) and Empirical (orange) data randomization techniques compared to the Truth(Green) from the original data for randomization weight values of (a) 0.1, (b) 0.5, (c) 0.75, (d) 0.9



(a) Original Data



(b) Uniform Randomization Method



(c) Empirical Randomization Method

Figure 17: Trips by Day of Week and Mode of Transportation Distribution for original data (a) compared to Uniform (b) and Empirical (c) randomization sampling for a weight of 0.5

5 Conclusions

5.1 Summary

Governments often contract out data collection efforts to better understand the needs of the public and make informed policy decisions. For transparency and the nature of government funding through taxpayer money, such data collection comes with the condition that the data and analysis gets released publicly. We were motivated by this situation for a Travel Behavior Inventory in the Minneapolis/St. Paul metropolitan area, which went beyond collecting travel information such as mode of transportation, vehicle types, among others, and also collected detailed socioeconomic information of participants.

Such detailed data is valuable to answer a broad number of research questions. However, releasing this data publicly can be a major violation of privacy for participants. Motivated by this, our broad goals were to mathematically model the trade-off between privatization methods and maintaining properties of aggregate information such that individuals included in the data cannot be confidently identified by a malicious actor.

We explored some ideas in how a malicious actor may abuse the information to target individual(s) even without precise information about where they live. In particular, FIPS codes can identify households in the data to approximately a region of a square mile, if not less. Since this data ties information across the datasets by unique identifiers (for research purposes), it means a malicious actor could also identify the historical patterns of trips made by those in the household. Especially for those in minority groups (whether by age, gender, ethnicity) we anticipate there is already enough in this data set to de-anonymize individuals and households for a motivated actor.

We studied this with a few working hypotheses:

- Maintaining utility for a research question may be more difficult for more complex research questions than simpler questions. For instance, asking "what is a statistic of the values in a column" (that we call a first-order statistic) may be easier to preserve utility than a more complex question which requires subsetting and/or cross-referencing (that we loosely call a second-order or complex question).

- Randomized-response is first introduced based on flipping a fair coin. A more useful version may use a biased coin, whose bias is a tuneable parameter which controls the probability of the true response being reported (or sampled at random from a distribution).

- The probability distribution which is sampled in this coin-flip generalizes to a discrete uniform distribution. Research questions may be better served if the replacement follows an empirical distribution observed in the ground truth.

Finally, we collected a variety of potential research questions one may ask of this data set, of varying complexity. Once methods to compute these answers from the data were established, we explored the numerical tradeoffs with the parameter w (Randomization weight) in section 4. Qualitatively the limits $w \to 0$ (full randomization) and $w \to 1$ (ground truth) behaved as expected, when measured in terms of the Kullback-Leibler (KL) divergence (to measure the how the probability densities varied) and the fraction of individual values which maintained their original value after applying the Randomized-Response algorithm (to measure a level of individual privacy). Surprisingly, while the KL divergence reflected the fact that the "replaced" dataset's distribution closely agrees with the ground truth distribution, we observed a general trend that the ratio of values which maintained their true value showed less difference than expected when sampling from a uniform distribution or the empirical distribution on that data – even when the empirical distribution was fairly non-uniform. However, we observed that similar methods applied to a single column which was then used in calculating second-order statistics almost completely removed the utility which could be maintained for first-order statistics – qualitatively performing no better than uniform sampling with respect to randomization weight.

5.2 Future Work

Here's a list of ideas.

1. Exploring some of the more sophisticated methods preserving privacy, such as "vine copula" and "multiple imputation," and seeing how they compare to the more simple differential privacy.

   (a) Explore other methods for privacy and compare how utility responds with other methods.

2. Exploring approaches to randomized replacement to preserve the utility in calculating second-order statistics.

3. Exploring what is the minimum amount of data that needs to be changed to preserve participant privacy.

4. Interpretability of KL-divergence in the context of the data, and in a toy example of overlapping Gaussian distributions.

5. Exploring different definitions of privacy and utility and the applicability of these methods to those definitions.

References

[1] Cynthia Dwork and Aaron Roth. "Formalizing Differential Privacy". In: *The algorithmic foundations of Differential Privacy*. now Publishers Inc, 2014, pp. 15–26.

[2] Matthew R. Francis. "Protecting Privacy with Synthetic Data". In: *SIAM NEWS* (May 2022).

[3] Sébastien Gambs et al. "Growing synthetic data through differentially-private vine copulas". In: *Proceedings on Privacy Enhancing Technologies* 2021.3 (2021), pp. 122–141. DOI: 10.2478/popets-2021-0040.

[4] Bei Jiang et al. "Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation". In: *Journal of the American Statistical Association* 117.537 (2021), pp. 52–66. DOI: 10.1080/01621459.2021.1909597.

[5] *Kullback–Leibler divergence*. June 2022. URL: https://en.wikipedia.org/wiki/Kullback%5C%E2%5C%80%5C%93Leibler_divergence#Motivation.

[6] *Relative entropy*. URL: https://www.sciencedirect.com/topics/engineering/relative-entropy.

[7] *Travel behavior inventory (TBI) 2018-2019 household interview survey*. URL: https://gisdata.mn.gov/dataset/us-mn-state-metc-society-tbi-home-interview2019.