VISUALIZATION AND DIFFERENTIAL PRIVACY

BY

HYUN BIN LEE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Professor Carl A. Gunter

ABSTRACT

Privacy-preserving statistical databases are designed to provide information about a population while preventing end-users from learning about an individual. Meanwhile, scholars [1, 2] have shown that a sophisticated adversary can break such assumption against primitive privacy protections. Differential Privacy (DP) measures how likely an adversary learns about an individual from statistical database queries. Recent state-of-the-art Privacy-Enhancing Technologies (PETs) often implement noise injection based mechanisms in order to satisfy a strong DP protection level. While these privacy protection guidelines minimize risks of private information disclosure, many people have raised concerns on impracticality of the implementation. Based on statistical figures and quantitative experiment results, much literature formalized the utility-privacy tradeoff caused by the noise injection.

In contrast, this work describes a qualitative analysis of the Laplacian noise mechanism, one of the most prevalently used DP mechanisms, with regards to the utility-privacy tradeoff on various types of visualization products. The dataset used for the analysis is a time series meter readings from smart grid electricity consumption of selected households from the Republic of Ireland. We examined how five types of visualization products, bar graphs, pie charts, heatmaps, linear plots and scatterplots, present information from statistical database queries. Visualization products showed seasonal, daily and weekly periodic consumption patterns such that power utilities can make a qualitative analysis of consumption profiles. We call these patterns as "visual cues." After applying the Laplacian noise mechanism on these visualization products, we made qualitative observations on the privacy-preserved figures and looked for notable changes.

The project provides graphic findings of a relationship among the composability of queries, the number of queries, and the scale of the Laplacian noise. We observed that visualization products which required less than ten queries

from the dataset suffered minimal information loss. However, we spotted a high degradation of visual cues when we implemented the noise mechanism to heatmaps with up to 25,200 composable queries. These visualizations no longer conveyed most key information that used to be present on their unprotected counterparts. To best of our knowledge, no state-of-the-art existing pre/post-processing techniques significantly recovered most visual cues. Finally, we found that some visualizations belonged to neither of the first cases. privacy-preserving linear plots (based on 336 composable queries) and scatterplots (based on 3,639 pairs of parallel queries) inherited some visual cues after executing privacy-preserving procedures. We further discovered some pre/post-processing mechanisms that recovered visual cues.

To my parents, for their love and support.

ACKNOWLEDGMENTS

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000780[1] and the Maryland Procurement Office under Contract No. H98230-14-C-0141[2]. The smart meter data used is provided by the Commission for Energy Regulation, and accessed via the Irish Social Science Data Archive.

[1]This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof.

[2]The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Maryland Procurement Office.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Queries from privacy-preserving statistical databases should only provide information of a group. However, we observed how these queries could be exploited from famous de-anonymization examples of the Netflix Prize dataset [1] and of the Massachusetts Group Insurance Commission medical encounter database [2]. These attacks raised awareness of statistical database privacy, and led to a formalization of private information disclosure risk. DP measures a risk of exposing any information of an individual included in the database. [3] Since its introduction, DP is often utilized as a measure of privacy protection in numerous PET-related papers.

Many DP mechanisms use randomized noise to hide trace of each individual. One of such mechanism is called the Laplacian noise mechanism which adds a set of noise distributed on a Laplace distribution to the query values. While the Laplace mechanism offers randomness that minimizes danger of identifying samples in the database, users reported its adverse effects on accuracy of the privacy-protected queries. Numerous authors formalized this utility-privacy tradeoff based on quantitative and statistical figures like mean-square error. Meanwhile, most of these scholars overlooked a possibility of analyzing this tradeoff from a different viewpoint.

1.1 Visualization as a Data Analysis Technique

Data analysis techniques summarize general trend of large data and illustrate important implied information. Technology advancements enabled computation and creation of large data, so the importance of analysis techniques has increased. Statistical databases output queries as data analysis sources. Data visualization, one of the most common data analysis techniques, assist our visual system's "highly tuned ability" to see patterns, identify outliers,

and enhance our understanding of data. [4]

Although the main purpose of visualization products is to visually illustrate overall trend of given data, the possibility of leaking PII from published visualization products must not be underestimated. For instance, an adversary can map visual signals (numbers, colors, length of objects, and etc.) to numerical query values. Depending on how sophisticated the mapping technique is, the adversary can obtain actual query data from visualizations without obtaining its numerical counterpart. The uncertainty caused by the translation process may fool us to conclude that the risk is negligible. Meanwhile, any knowledge of an individual gained from the visualization can be combined with exterior knowledge of the adversary. As a result, individual's sensitive information can be leaked from unprotected visualization products.

1.2 Contributions

Only a few number of research groups focused on implementing DP mechanism to visualization products. Even fewer groups attempted to measure a utility-privacy tradeoff when injecting Laplacian noise to queries that constitute visualizations. This project serves to provide a qualitative analysis of implementing the Laplacian noise mechanism to 2D plots and charts. The project largely focuses on degradation of utility caused by the randomness of noises. The project also provide countermeasures against potential adversaries who attempt to exploit vulnerabilities from unprotected data visualization products.

We simulated a statistical database that stores smartgrid meter readings of Irish households. Based on the statistical queries, we created visualizations of five different types such as bar graphs, pie charts, heatmaps, linear plots and scatterplots. We observed these visualizations and recorded qualitative observations of information contained in each visualization. We also simulated a DP counterpart database that outputs DP queries. We also created visualizations from these queries as well. Information from these visualizations was compared with those provided by unprotected products to make a utility-privacy tradeoff analysis.

The main contribution of our thesis includes:

- We examined how different visualization products conveyed informa-

tion to end-users. We made qualitative observations and found key patterns that were not captured when examining raw, unsorted data. These "visual cues" aided viewers to efficiently collect implied yet crucial information.

- We demonstrated a case study of utility-privacy tradeoff analysis for DP visualization products. We observed how the Laplacian noise mechanism degraded information contained in visualizations and found that different visualizations showed different degree of degradations. We concluded that degradation of visual cues worsened as the number of dependent queries has also increased.

- We provided guidelines for data pre/post-processing of data when implementing the Laplacian noise mechanism to different visualization methods. Some visualizations discussed in the thesis require data processing procedures. We delineated the procedures such that readers could utilize our findings on their works in the future.

1.3 Thesis Structure

The thesis is outlined as following. Chapter 2 serves to introduce the definition of DP and to explain existing DP mechanisms. The chapter also delineates current state-of-the-art privacy-preserving visualization techniques. Then on Chapter 3, we explain our experiment details such that readers can understand our contributions. We discuss our experiment results in three chapters, Chapter 4, Chapter 5 and Chapter 6. Details for each experiment will be explained in respective chapters. We examine limitations and future works in Chapter 7. Finally, we conclude our thesis in Chapter 8.

CHAPTER 2

BACKGROUND

Let x and y be databases of samples collected from a universe χ. Then, x and y can be redefined as histograms of χ such that $x, y \in \mathbb{N}^{|\chi|}$ where each $x_i \in x$ and $y_i \in y$ represents the number of sample $i \in \chi$ in the databases. Given these notations, Dwork and Roth [5] define DP as following.

Definition 1. A randomized algorithm M with domain $\mathbb{N}^{|\chi|}$ is (ϵ, δ)-differentially private if for all $S \subseteq Range(M)$ and for all $x, y \in \mathbb{N}^{|\chi|}$ such that $\|x - y\|_1 \leq 1$:

$$Pr[M(x) \in S] \leq exp(\epsilon) Pr[M(y) \in S] + \delta$$

Set S is a subset of the range of M. Inequality $\|x - y\|_1 \leq 1$ represents a distance between two databases, and the distance between the two is at most one. In other words, the database y is equivalent to the database x with or without an individual. Thus for small ϵ and δ, the definition above ensures a strong statistical guarantee that a presence or an absence of individual will unlikely affect the output of the algorithm M. In other words, two parameters δ and ϵ define a risk of privacy disclosure. Dwork and Roth recommended δ to be less than the inverse of any polynomial in the size of the database. Most scholars set ϵ to be less than or equal to one when they discuss a strong privacy guarantee.

Authors emphasize that DP is "immune" to post-processing techniques. That is, once an end-user receives output of DP algorithm M, he or she cannot change the privacy protection level of the output given that the user does not have additional knowledge about the database. This is extremely useful as we introduce several post-processing mechanisms that improve utility of visualizations throughout this paper.

2.1 The Laplacian Noise Mechanism

DP is a definition not a protection mechanism. In order to satisfy the definition, constructing a randomized algorithm M is crucial. In this section, we introduce the Laplacian noise mechanism as a randomized DP mechanism to construct M. This mechanism adds Laplacian noise, an independent and identically distributed (i.i.d.) random variable drawn from $Lap(\Delta f/\epsilon)$ distribution, to each query. Note that the distribution has a parameter $\Delta f/\epsilon$. The symbol ϵ simply represents the level of ϵ-DP to be accomplished. The remaining Δf is called the sensitivity of function f. Dwork and Roth [5] define sensitivity as below.

Definition 2. Sensitivity of a function $f : \mathbb{N}^{|\chi|} \rightarrow \mathbb{R}^k$ is:

$$\Delta f = max\|f(x) - f(y)\|_1$$

Sensitivity determines a range of output from DP algorithm M due to an inclusion of a sample individual. It represents a bound of randomness for each DP query. In case of histograms, function f would refer to frequency of each bin. Given that x and y differs by one sample, the maximum difference of frequency for each bin would be 1. Thus, $\Delta f = 1$ for a histogram.

Dwork and Roth [5] have shown that algorithm with ϵ-Laplacian noise mechanism is $(\epsilon, 0)$-DP. We used this mechanism to implement DP mechanism when constructing DP visualizations.

2.2 Composition

Another important characteristics of any DP mechanism is the concept of composition. Consider a statistical database which outputs aggregate electricity consumption of consumers during Christmas Eve and Christmas Day. One query would represent aggregate consumption during 24th and another one would represent the consumption during 25th. An adversary may gain information of an individual from both queries since the individual's consumption data are aggregated in both queries. We define two queries to be correlated or dependent when the adversary can compose two queries to gain information of an individual.

On the other hand, consider another statistical database which outputs average heights of women and men in Champaign, Illinois. One query represents average height of men in Champaign, and another query represents that of women. Then, adversary can only use one query to learn about an individual since any individual would belong to only one of two genders. We define two queries to be uncorrelated or independent when adversary cannot compose two queries to gain information of an individual.

Based on this notion of correlation or dependence, there exist two different types of composition, sequential composition and parallel composition. [6]

- When ϵ-DP mechanism is queried t times where each query is correlated with another query, then overall query is thought to be $(\epsilon \times t)$-differentially private. This concept is called the *sequential composition* of differentially private queries.

- When ϵ-DP mechanism is queried t times where each query is not correlated with another, then overall query is thought to be ϵ_{max}-DP where ϵ_{max} represents the largest ϵ_{max} among t DP queries. This concept is called the *parallel composition* of differentially private queries.

While each query is ϵ-DP, a composition of t queries make overall output $(\epsilon \times t)$-differentially private. Therefore, to ensure this group of queries to be ϵ-DP, one must ensure each query to be ϵ/t-DP. In the case of sequential composition, the number of queries and the size of the Laplacian noise are proportional to each other.

Composition and sensitivity determine the scale of noise for DP mechanisms. Therefore, one must consider these two aspects when constructing DP visualizations. In case of histograms, computing sensitivity would not require much effort. However, how would one compute sensitivity of a scatterplot? How would one quantify presence and absence of a point? Furthermore, naively implementing a DP mechanism often result in poor privacy-utility tradeoff. When creating any visualization product, each respective algorithm usually repeats a set of noise-injection operations over each visualization component. For instance, the scatterplot algorithm needs to run a dot-drawing procedure for each sample and the histogram algorithm needs to run a bar-drawing procedure for each frequency bin. Since each query of visualization mechanism is often correlated with one another, the noise may grow too large such that utility of the output is greatly degraded.

2.3 Advanced Composition

Dwork and Roth [5] introduce theory of advanced composition which reduces the scale of noise growth and ensures the size of noise to be $O(\sqrt{k})$. This theorem can be applied in case where repeated use of differentially private algorithms on the same database occurs. Constructing visualizations requires collecting iterative queries from the same database as each query represents an attribute or a reading of visualization. The theorem of advanced composition is shown below.

Theorem 1. *(Advanced Composition) For all $\epsilon, \delta, \delta' \geq 0$, k-repetitive or k-fold adaptive composition of (ϵ, δ)-differentially private mechanisms satisfies $(\epsilon', k\delta + \delta')$-differential privacy for:*

$$\epsilon' = \sqrt{2kln(1/\delta')\epsilon} + k\epsilon(e^{\epsilon} - 1).$$

Note that if $0 < \epsilon' < 1$, $\epsilon = \frac{\epsilon'}{2\sqrt{2kln(1/\delta')}}$.

The size of noise is $O(\sqrt{k})$ with respect to k composable queries. Nonetheless, with large k, the noise can grow significantly large.

2.4 Related Work

To the best of our knowledge, only a few group of researchers published literature regarding the topic of privacy-preserving visualization techniques. Dasgupta et al. implemented k-anonymity and l-diversity on parallel coordinates. [7] The same research group implemented their aforementioned technique on scatterplots and introduced various metrics which measure privacy-utility tradeoffs. [8] Unfortunately, the research group did not use DP to measure privacy protection of visualizations. It is known that for a high-dimensional dataset, k-anonymity cannot provide sufficient privacy protection. [9]

Xu et al. implemented differential privacy on histograms. [10] The authors showed that the DP-protected histogram with finer bins led to lower accuracy than a coarse one. Authors also discussed some cases when histogram structure itself leaked sensitive information. They introduced two algorithms, Noise First and Structure First, for computing DP-compliant histograms. Eibl et al. implemented DP to dataset of smart-grid consumers'

aggregate electricity consumption. [11] The authors implemented DP on aggregate meter readings and plotted linear plots to visualize the results. They applied a smoothing algorithm as a post-processing mechanism to improve utility of output.

CHAPTER 3

EXPERIMENT INFORMATION

Before we explain our experiment results, we would first like to describe our smartgrid electricity dataset for the experiment. The Commission of Energy Regulation (CER) and the Irish Social Science Data Archive (ISSDA) provided electricity consumption profiles of around 5,000 Irish homes and businesses [12]. The dataset was built using smart meter measurements that took place from July 2009 to December 2010. The meters installed in each household made measurements once per 30 minutes, and the consumption was measured in kWh. The dataset was anonymized by removing any PII and each household was identified by a unique id number. The ISSDA commented that this anonymization would not provide sufficient privacy protection, so we presented only visualizations of the dataset as an abstraction of the dataset. If the adversary had the original dataset, he or she could compromise privacy of individuals.

3.1 Data Analysis

Based on a survey from ISSDA, each household belonged to one of three categories, Residential, Small and Medium Enterprises (SME) or Others. We created heatmaps of aggregate electricity consumption for each category to see consumption patterns for each category. Figure 3.1 represents heatmap of residential households aggregated electricity consumption and Figure 3.2 represents that of SME.

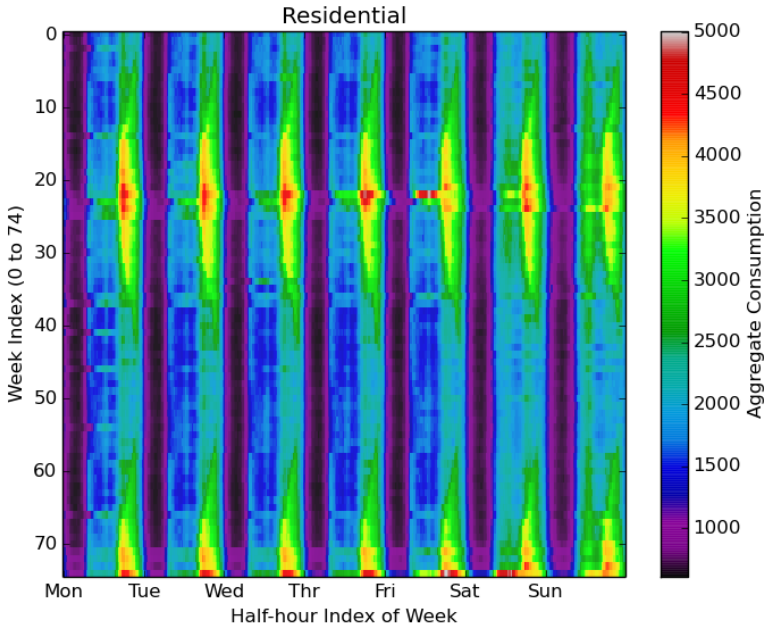

Figure 3.1: Heatmap of Residential Aggregated Electricity Consumption

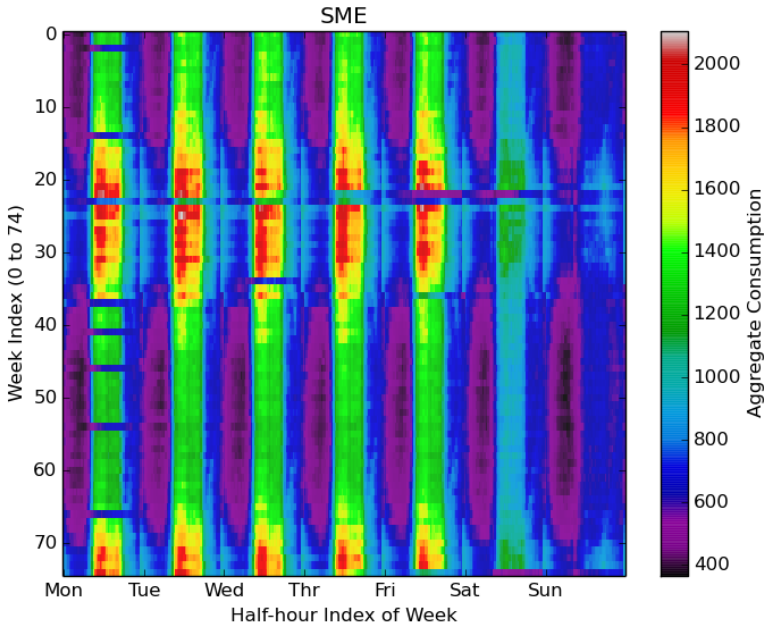

Figure 3.2: Heatmap of SME Aggregated Electricity Consumption

For both Figures, x-axis represents 48 half-hour indices of each week where index 0 represents 12:00 AM Monday and index 47 represents 11:30 PM

Sunday. The y-axis represents 75 weeks from July 2009 to December 2010. The colormap on the right represents a mapping from a color to aggregate electricity consumption measurement. These two heatmaps provided rich information of electricity consumption patterns. For instance, in Figure 3.1 we observed that Irish people tend to have highest electricity consumption during evening time from dinner time to bed time. Then, the consumption dipped at dawn when people went to sleep. Also the figure showed that people spent more electricity during winter than during summer. Ireland has relatively cool weather during summer (average high temperature of 19 Celsius during July and August) such that people probably did not use air conditioning much during summer. Meanwhile, many households use electrical heating in winter, so electricity consumption largely increased during winter. Furthermore, we observed that Friday of week 22 for Figure 3.1 showed a unique consumption pattern where the peak consumption happened during day time. Furthermore, the Thursday and Friday of week 22 both showed relatively higher electricity consumptions than rest of the week. We concluded that this meter data showed a special holiday pattern, because Thursday and Friday of week 22 are Christmas Eve and Christmas Day in year 2009. Note that Friday and Saturday of week 74 also showed a similar consumption pattern. We confirmed that those two days also represented Christmas Eve and Christmas Day in year 2010.

The Christmas holiday weeks consumption pattern was also clearly displayed in Figure 3.2 for SMEs aggregate electricity consumption heatmap. Week 22 had very low consumptions compared to those of adjacent weeks such that a horizontal line was created between each day's band. We inferred that many SMEs closed their businesses during Christmas holiday week such that the electricity consumption dropped largely during that week. We observed similar horizontal band patterns for Mondays, and we concluded that these bands represent Irish bank holidays. Also, in Figure 3.2 we found out that week day electricity consumption was relatively higher than that of weekend electricity consumption since many SMEs did not operate on weekends.

We decided that the residential dataset and the SME dataset clearly showed many different consumption characteristics, so we used the residential dataset for the rest of the experiment. We conducted a series of experiments on a variety of widely used visualization methods to observe effects of differential

privacy.

3.2 Experiment Design

Different visualization products convey different information to end-users. Among those products, we have chosen five widely used techniques. Those include bar graphs, pie charts, heatmaps, linear plots and scatterplots. We plotted each visualization with different aspects of the dataset to distinguish each product. Then, we examined these visualization products and analyzed information implied in the visualizations. The information aided end-users to answer questions and make decisions such that the viewers had more understanding of the dataset after viewing each visualization. For most visualization types, we attempted to find if it is possible to recognize effects of the Christmas holidays in visualizations of calendar data.

Then, we applied DP to each visualization product. In this experiment, we implemented the Laplacian noise mechanism. In most cases, each data query was related with another query such that adversary potentially learns more about an individual by composing information of two queries. The DP mechanism counters this additional information gain by increasing the scale of Laplacian noise directly proportional to that of the number of related queries n. We used two different kinds of composition, naive sequential composition and the advanced composition [5]. Advanced composition, given that each query is similar to another, reduces the noise scalability by $\mathcal{O}(\sqrt{n})$, so we implemented advanced composition. We only used sequential composition when the size of the noise injected to each data query was too minuscule such that the end products did not vary much. We set default epsilon to 1, which was the largest epsilon with sufficient privacy protection.

After producing DP visualizations, we once again made the same observations as we did with its unprotected counterparts. We asked a following question. "After adding noise for privacy protection is it possible to recognize key effects in visual data representations?" In other words, did visual cues which existed in original visualizations disappear after applying DP mechanisms? The experiment section is divided into three large categories. The first part of experiment on bar graphs and pie charts depicted straight-forward implementation of DP without much challenge. We called this experiment "low

degradation" experiment. Then the next experiment on heatmaps showed that the DP implementation substantially degraded original products such that to best of our knowledge, no state-of-the-art technique can be implemented to preserve utility and protect privacy. We called this experiment "high degradation" experiment. Finally, two experiments on linear plots and scatterplots are called "medium degradation" experiments. After implementing the Laplacian noise mechanism on these two visualizations, we observed that the privacy-protected products contained some information. Then, we made some post-processing tricks onto these products which recovered lost information.

Due to limitation of time and resources, author of the thesis was the only participant of the experiment. Therefore, we are fully aware that qualitative observations made in this thesis can be biased. We would like to incorporate crowd-sourcing based surveys to complement this research in future works.

CHAPTER 4

LOW DEGRADATION OF VISUAL CUES

Our experiment first focused on visualizations with a small number of attributes. These products usually conveyed simple, coarse-grained information. For instance, plotting daily average consumption during the week of Christmas requires seven queries from statistical database, one for each day of week. Directly displaying numerical values in tables would also display similar information if the number of attributes is small. Thus, we also wanted to observe if providing numerical values along with visualizations provide different information than visualizations without numbers.

For this section, we focused on two visualization types, bar graphs and pie charts, because we believed these two visualization methods were most suitable to deliver coarse-grained information to viewers. Furthermore, we evaluated that these two products conveyed analogous information to end users such that using the same data queries on each visualization would result in similar results. When observing these products, we asked following questions.

- Do bar graphs and pie charts with the same data give the same degree of information gain? In other words, can pie charts substitute bar graphs without sacrificing any utility?
- Do numerical values assist readers to understand the visualizations?
- In bar graphs and pie charts of Christmas week calendar data, is it possible to recognize effects of the Christmas holidays?

Since applying Laplacian noise mechanism on small number of attributes won't degrade data much, we tagged this experiment to be an easy experiment. For the experiment, we first used daily aggregation consumption during week of Christmas for residential households. The visualization products without privacy protection are shown at Figure 4.1 and 4.2.

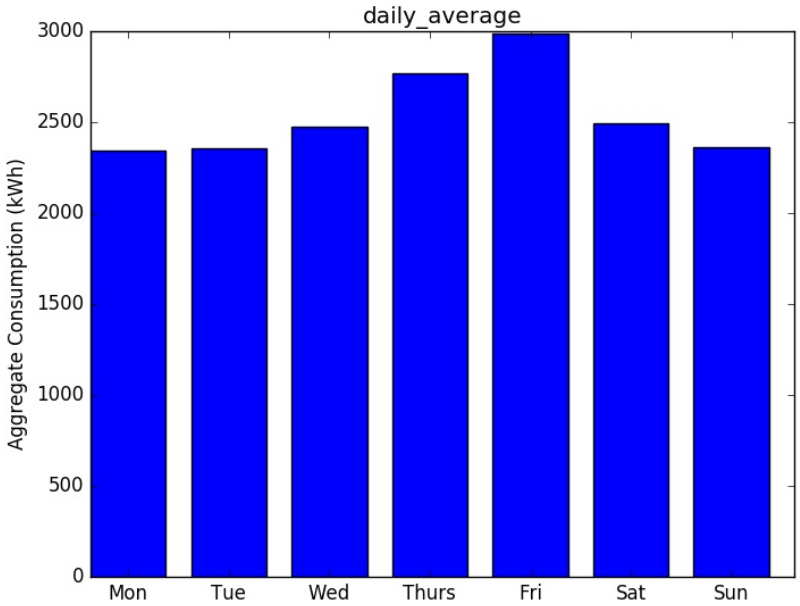

Figure 4.1: Bar Graph of Daily Aggregate Consumption during Week of Christmas

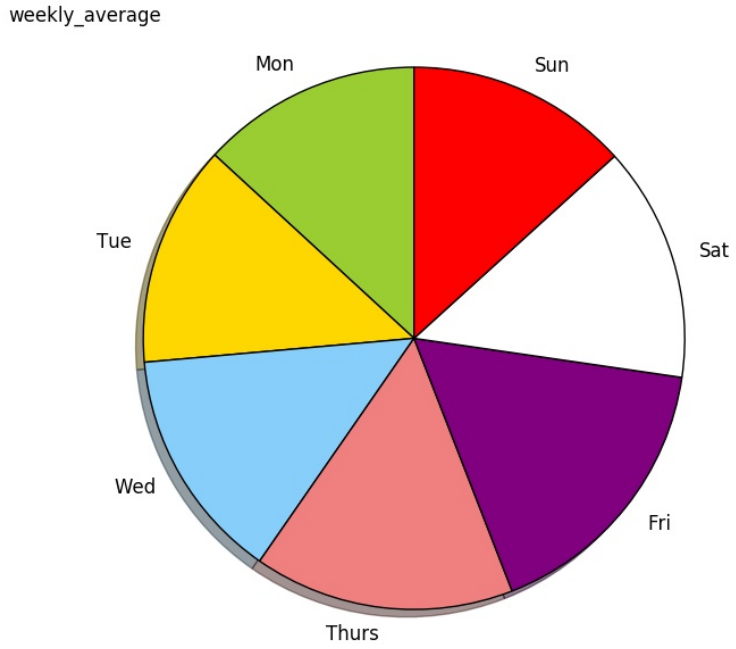

Figure 4.2: Pie Chart of Daily Aggregate Consumption during Week of Christmas

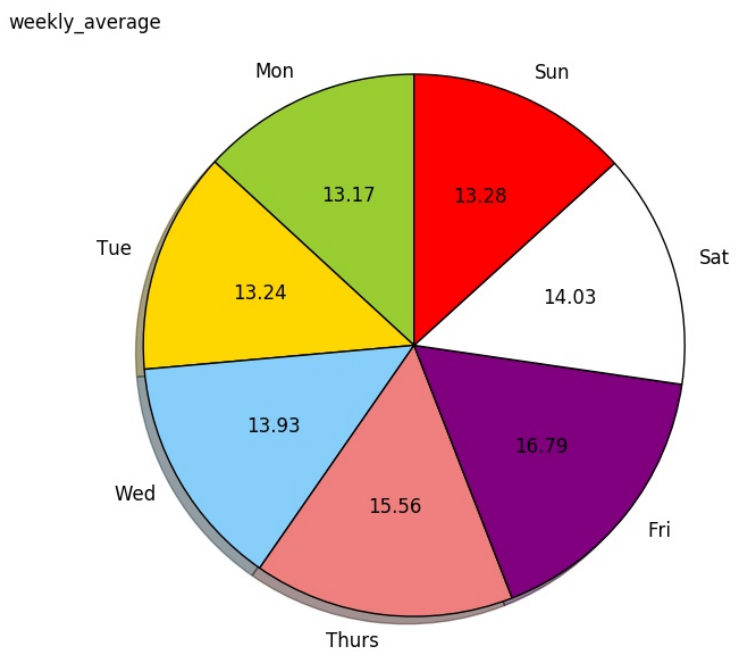

Figure 4.3: Pie Chart of Daily Aggregate Consumption with Percentage

Without knowing prior knowledge of what day Christmas was in year 2009, the end-users are most likely to guess which day Christmas was with $\frac{1}{7}$ probability of being correct. Then the user sees the bar graph of daily aggregate consumption from Figure 4.1. Given that viewer knows that Christmas Eve and Christmas Day have more electricity consumption than other days of week in residential households, viewers must be able to distinguish a pair of days with consecutive high electricity consumption from Figure 4.1's bar graph. We concluded that bar plots effectively showed those days in Figure 4.1 such that the end-user has gained information regarding when the Christmas holiday was.

On the other hand, we could not gain such information from pie charts in Figure 4.2. For the bar graph, each bar was sided along with each other so that length comparison tasks among different bars were trivial. Meanwhile, comparing areas of each sector for pie charts was not an easy task. Figure 4.3 attempted to add a numerical percentage for each day's sector to assist the readers. After adding the numerical percentages along with the visualization, then viewers could distinguish that Thursday and Friday were the Christmas Eve and Christmas Day pair. Meanwhile, in such case viewers gained information from numerical values not from the pie chart itself. We

concluded that bar graphs provided more information than pie charts for the data set of daily electricity consumption during the week of Christmas. After making qualitative observations from unprotected visualizations, we applied the Laplacian noise mechanism to both visualizations. Figure 4.4 and Figure 4.5 are the products.

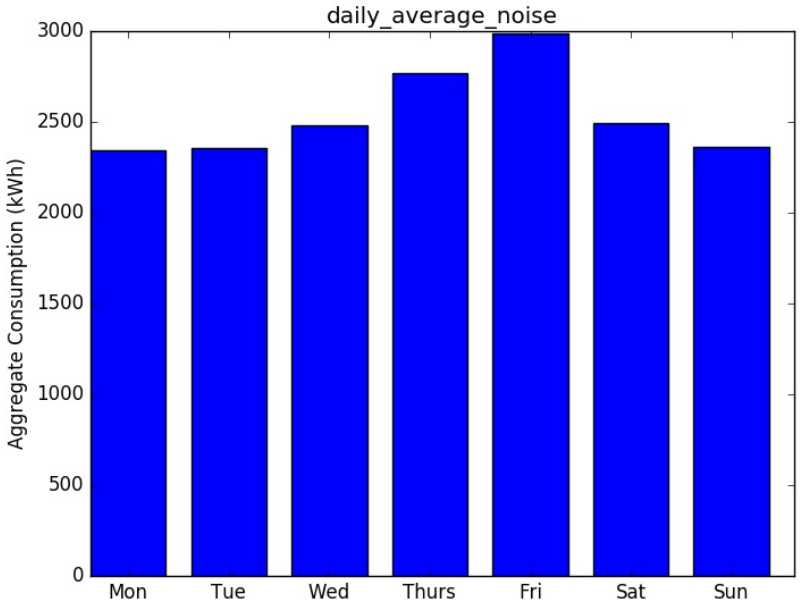

Figure 4.4: (1,∞)-DP Bar Graph

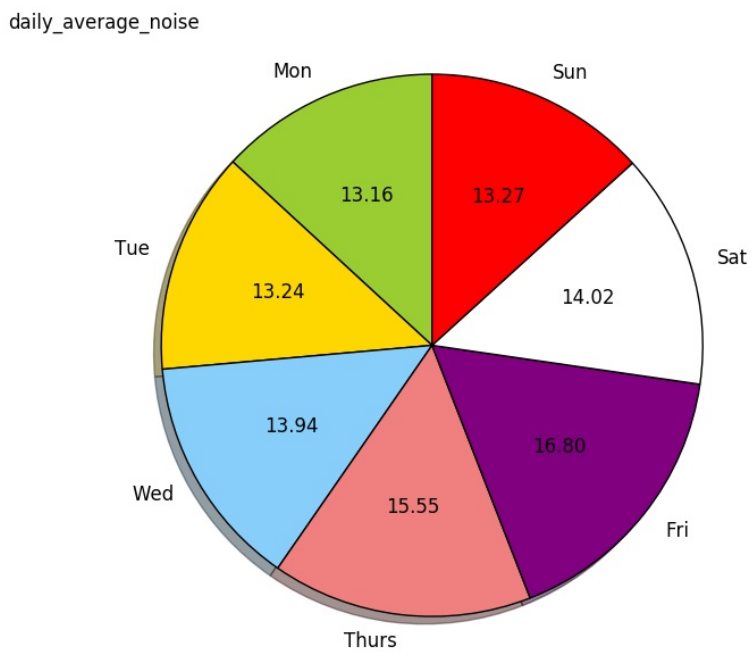

Figure 4.5: (1,∞)-DP Pie Chart

Figure 4.4 and 4.5 are (1,∞)-DP counterparts of Figure 4.1 and 4.2. There was no visible difference after injecting Laplacian noises. Viewers would receive the same information from (1,∞)-DP counterparts as they did before, so the DP mechanism would not impact end-users' decision making process. This was expected since visualizations required only seven queries such that only minuscule noise was injected to each attribute value.

We attempted to create visualizations with another set of data queries to confirm if pie charts can completely substitute bar graphs or not. From the same dataset, we plotted average consumption during 4-hour period.

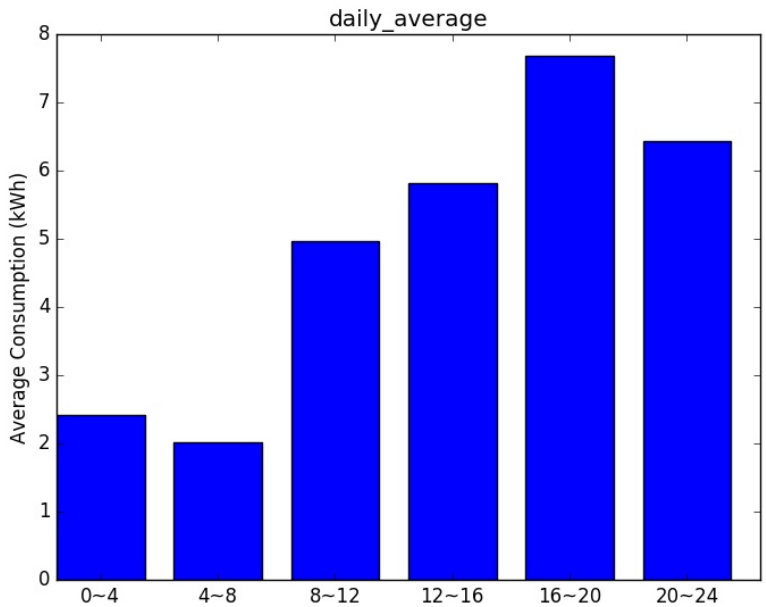

Figure 4.6: Bar Graph of 4-Hour Aggregate Consumption during Week of Christmas

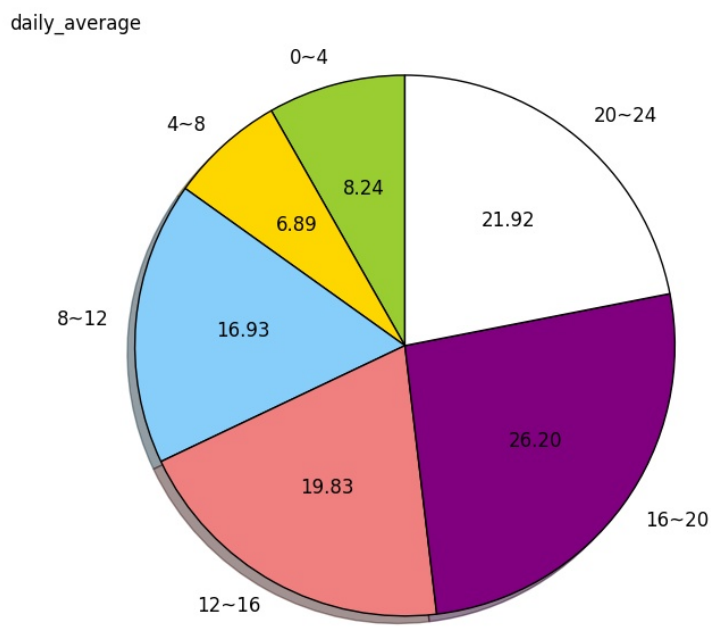

Figure 4.7: Pie Chart of 4-Hour Aggregate Consumption during Week of Christmas

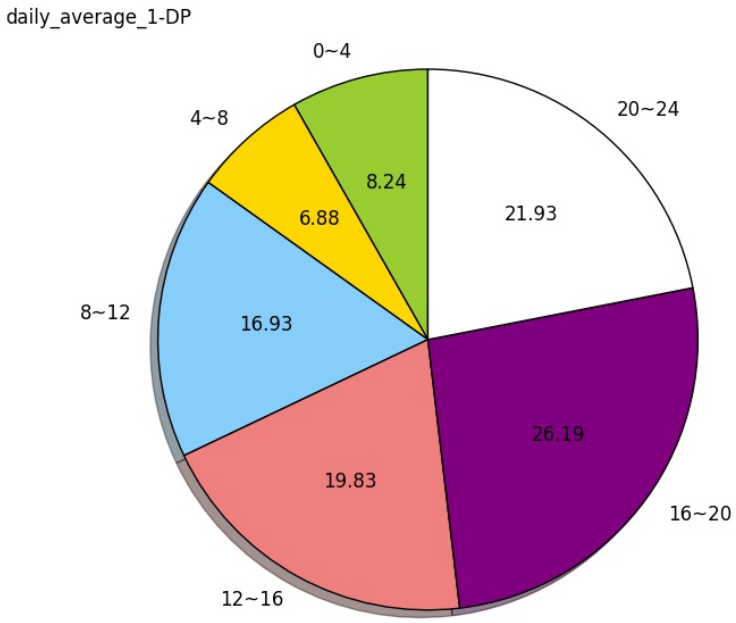

Figure 4.8: (1,∞)-DP Private Pie Chart

In Figure 4.6, consumption decreases as time elapses from first bin (0 to 4 AM) to second bin (4 to 8 AM) period, because most people went to sleep until around 7 to 8 AM. Then the consumption drastically increased during 8 to 12 PM period. The measurements were only made once in 4 hours, so the information provided were too coarse-grained to make any conclusion, but viewers could nonetheless see some chronological pattern. For its pie chart counterpart, Figure 4.7, we now could compare area of each sector since the difference among each sector was large enough to be observed by human's naked eye. We observed the same drastic increase from 4 to 8 AM period to 8 to 12 PM period, and we concluded that 4 to 8 PM period had highest electricity spending period for both Figure 4.6 and 4.7. Thus, pie charts can convey similar information as bar graphs do if a value of each attribute differs significantly. The Figure 4.8 is the DP counterpart of 4.7. As we have observed from previous visualizations, the number of queries was too small such that there was no visible change after applying the Laplacian noise mechanism.

CHAPTER 5

HIGH DEGRADATION OF VISUAL CUES

While bar graphs and pie charts displayed useful summarized information of the dataset, the queries only displayed the week of Christmas and one aggregated reading per day. In contrast, the heatmaps we saw in Figure 3.1 and 3.2 showed aggregate consumption information of all 75 weeks with 48 readings per day. We would like to ask if we can protect privacy of each consumer when we publish this heatmap to the public. Can a DP heatmap created with the Laplacian noise mechanism still display information that we saw from Figure 3.1 and 3.2? For the DP heatmaps, we would like ask a following question. In the heatmap of smartgrid calendar data, is it possible to recognize effects of

- Christmas holidays
- seasonal changes
- weekends and weekdays
- day and night time

Figure 5.1 and 5.2 are heatmaps created with the sequential composition, which is the same technique we used when we created DP bar graphs and pie charts.

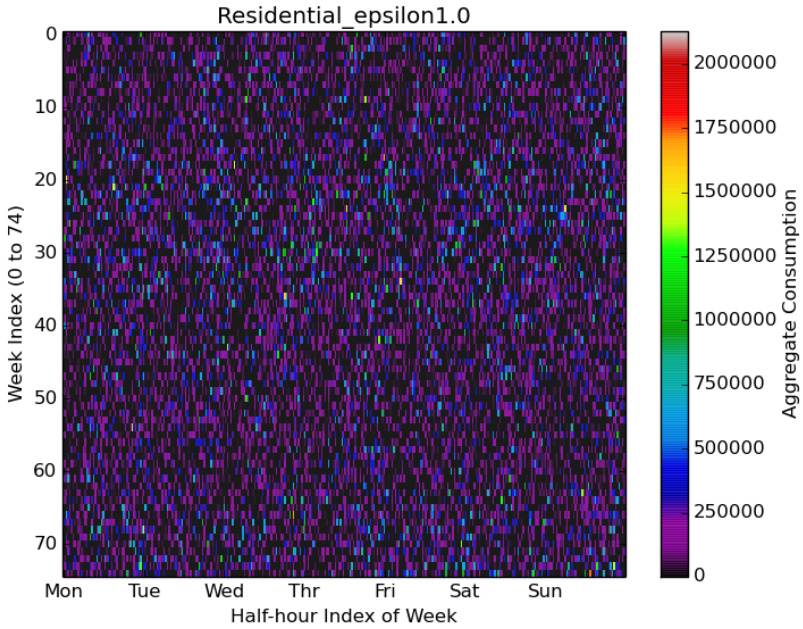

Figure 5.1: (1,∞)-DP Residential Heatmap with Sequential Noise Composition

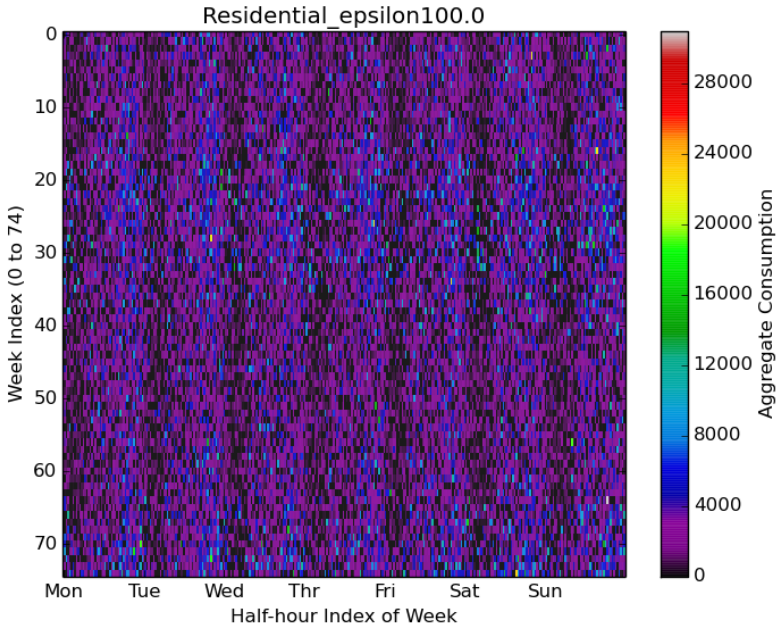

Figure 5.2: (100,∞)-DP Residential Heatmap with Sequential Noise Composition

In Figure 5.1, the heatmap no longer shows any useful information. We

could no longer see holiday patterns, seasonal patterns or daily patterns that were present in its unprotected counterpart from Figure 3.1. As we increased the epsilon to 100 in Figure 5.2, we could barely distinguish night and day time with black vertical bands. Also, we could observe that winter weeks have many sky blue and dark blue pixels while summer weeks mostly have purple pixels. Thus, we could infer that electricity consumption is higher in winter than in summer. Nonetheless, most of the information provided by Figure 3.1 was no longer available. Furthermore, (100,∞)-DP is not considered to be sufficient protection measure against any adversary.

Sequential composition required very conservative standards when composing queries. We used the same noise injection mechanism on the same database throughout the algorithm. Advanced composition [5] allows noise to grow slower when repeated use of differentially private mechanism on the same database occurs. Thus, we believed use of this composition could greatly improve the quality of visualization products. The DP visualizations created with advanced composition mechanism are shown at Figure 5.3 and 5.4.

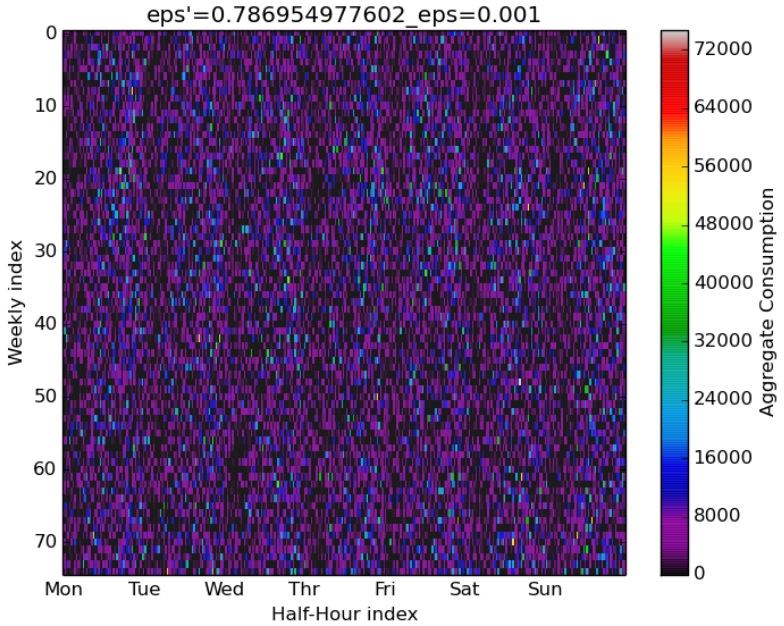

Figure 5.3: (0.79,10^{-5})-DP Residential Heatmap with Advanced Noise Composition

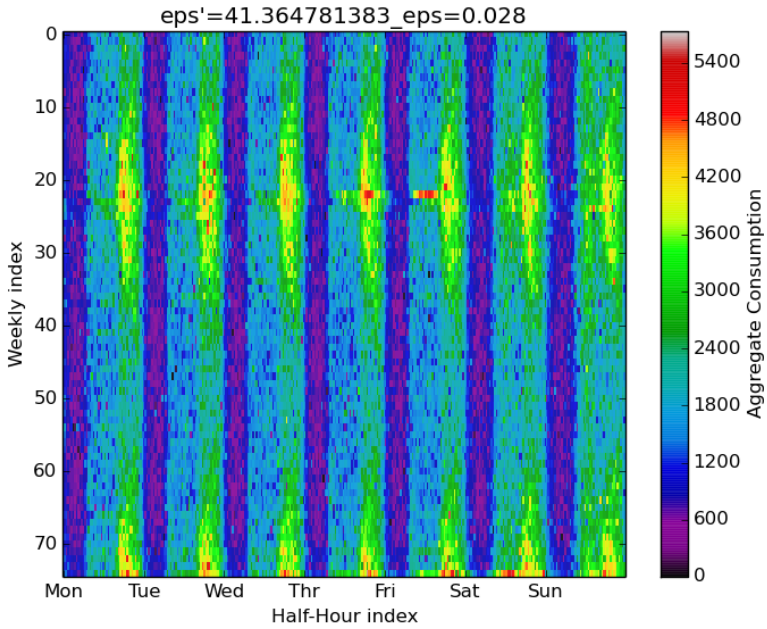

Figure 5.4: (41,10^{-5})-DP Private Residential Heatmap with Advanced Noise Composition

Figure 5.3 is the (0.79,10^{-5})-DP counterpart of Figure 3.1. Unfortunately we still could not see most information that used to be present on its unprotected version. We can claim some findings from this visualization, but we concluded that we had to increase epsilon in order to recover original information from Figure 3.1. We kept increasing the epsilon, and found that by the epsilon of 41, we finally were able to reproduce most information from original dataset. The resulting visualization product is shown at Figure 5.4. We can spot Christmas Eve and Christmas Day based on electricity spending peaks, and we could distinguish summer and winter based on average spendings. Also, day time and night time can be clearly distinguished as well. However, epsilon of 41 is too large to be considered as sufficient privacy protection.

We applied many state-of-the-art post-processing mechanisms to improve our results, but none of them yielded a satisfactory result. We would like to present our final post-processing effort with "divide and aggregate" algorithm. This algorithm attempted to aggregate readings that are adjacent and similar to each other. This could reduce the number of queries and the size of the Laplacian noise. The pseudo-code of the algorithm is the following.

```
inputs:
M, aggregate smartgrid meter readings matrix
h, threshold

divagg(M,h) {
    avg = mean(M)
    flag = true
    for each m in M:
        if abs(m - avg) < h:
            flag = false
            break
    if flag is true:
        return M' which is a new matrix filled with avg
    else:
        divide M into four sub-matrices, M1, M2, M3, and M4
        lt = divagg(M1,h)
        rt = divagg(M2,h)
        lb = divagg(M3,h)
        rb = divagg(M4,h)

        merge lt, rt, lb, and rb into a new matrix M'
        return M'
}
```

Note that adjacent pixels of the heatmap often had a similar color. Thus, we concluded that we could combine entries with similar values into one representative query rather than returning every query. Our algorithm recursively finds a group of adjacent meter readings that have similar values within a threshold h and unifies all readings to one average value. Figure 5.5 is the result of the algorithm with threshold h of 800 kWh.

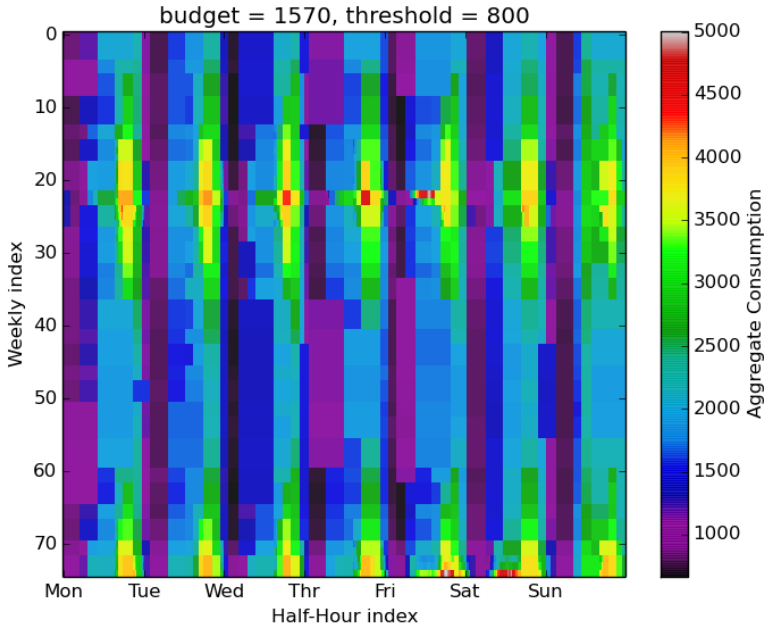

Figure 5.5: Residential Heatmap Created by Divide and Aggregate Algorithm

The total number of queries reduced from 25,200 to 1,570. However, we evaluated that the visual cues still remained in this processed heatmap. The heatmap still showed a unique consumption pattern during the week of Christmas, and it showed seasonal consumption patterns. Furthermore, the resulting heatmap did not display any false information that was not present in the real heatmap. Nonetheless, we evaluated that the number of queries (1,570) was too large such that applying the DP mechanism would greatly obscure the visual cues. Thus, we decided to reduce measurements from 75 weeks to 52 weeks, because 52 weeks constituted annual data such that extra 23 weeks of data repeated first 23 weeks. We reduced the total number of queries to 952 after making this decision. We calculated new sensitivity for each query and applied the advanced composition with epsilon of around 1.2 on the Figure 5.5. Figure 5.6 is the product of the implementation.

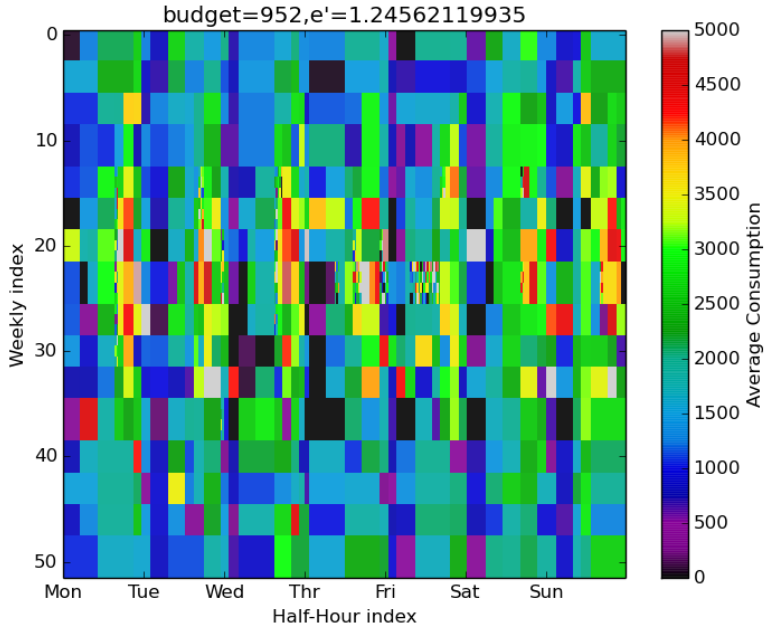

Figure 5.6: $(1.25,10^{-5})$-DP Residential Heatmap Created by Divide and Aggregate Algorithm

Although we applied pre-processing mechanisms that greatly reduced our privacy budget to 952 queries, the resulting product was too obscured. We could observe some purple or blue vertical bands that represented dawn period with low electricity consumption between each day, but we could not confidently claim that we observed the same daily consumption pattern as we did when we read Figure 5.5. The DP heatmap did show that consumption was higher during winter period, but the heatmap failed to show some key patterns that were seen in Figure 5.5. For instance, the private counterpart no longer showed an increase of electricity consumption during the week of Christmas in comparison to adjacent weeks. Furthermore, we could no longer spot a unique consumption pattern that happened during Christmas. Overall, we concluded that the visualization did not provide enough information to conjecture consumption patterns of Irish consumers. Therefore, we categorized this experiment to the experiment of high degradation of visual cues.

CHAPTER 6

MEDIUM DEGRADATION OF VISUAL CUES

So far, we have seen two categories of experiments. The first one was a "low" degradation experiment where the number of queries was less than 10 such that noise injections did not have any noticeable impact on visualization products. Meanwhile, the second experiment was a "high" degradation experiment where the number of queries was too large (25,200) such that after applying the Laplacian noise mechanism, the degree of data degradation was too large such that no state-of-the-art post-processing effort helped us to recover information displayed from its unprotected counterpart. This time, we conducted experiments for "medium" degradation of visual cues.

For first part of experiment, we used original 30-min period measurements like we did for heatmap experiment, but limited the range of data only to the week of Christmas not 75 weeks. In such case, the number of queries was greatly reduced from 25,200 to 336. We determined that a linear plot was the best visualization method to show weekly periodic patterns, so we created linear plots and made qualitative observations with the data. For the second part of the experiment, we collected 3,639 pairs of parallel queries from each sample such that each sample provided two queries. we used parallel composition when implementing the Laplacian noise mechanism, and plotted a scatterplot to visualize this collection of queries.

6.1 Linear Plots

For this experiment, we wanted to ask following questions when reading the plots.

In linear plot of Christmas weeks calendar data, is it possible to recognize:

- Effects of Christmas holidays

- Consumption peaks and troughs
- Characteristics of daily consumption pattern

Note that we successfully answered all three questions when we observed heatmaps. The heatmap has a higher dimension than linear plots do, because linear plots only show the data in x and y axis while heatmaps incorporate color to show information in addition to the two axes. Thus, the heatmap was suitable for visualizing the whole data set which contained weekly periodic information along y-axis and daily periodic information along x-axis. Meanwhile, linear plots only displayed information in two dimensions, so we concluded on plotting a week worth of data for daily pattern analysis.

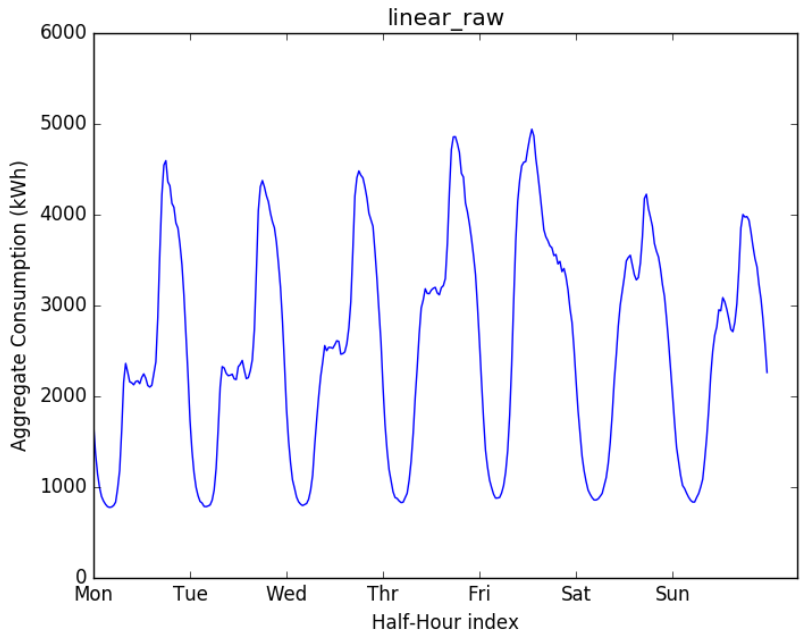

Figure 6.1: Linear Plot of Aggregation Consumption During Week of Christmas

Just like previous set of experiments, viewers with limited prior information gain information after reading the linear plot because they could infer that Thursday and Friday had highest consecutive electricity consumptions. This information gain was crucial for guessing the Christmas Eve and Christmas Day. Furthermore, we could distinguish each day since a consumption trough exists between two days when Irish consumers went to sleep. In addition, we

also observed that most days had peak consumption during night time while Christmas had its peak time during day time.

Figure 6.1 provided a fine-grained consumption information for weekly aggregate consumption data. We gained information regarding daily and weekly periodic consumption pattern. Furthermore, we easily identified Christmas Day by distinguishing the highest consumption peak. We experienced the same degree of information gain from Figure 3.1 when we conducted the heatmap experiment, but we lost all information when we injected Laplacian noise. Figure 6.2 shows $(1,10^{-5})$-DP Linear plot of aggregate consumption during the week of Christmas.

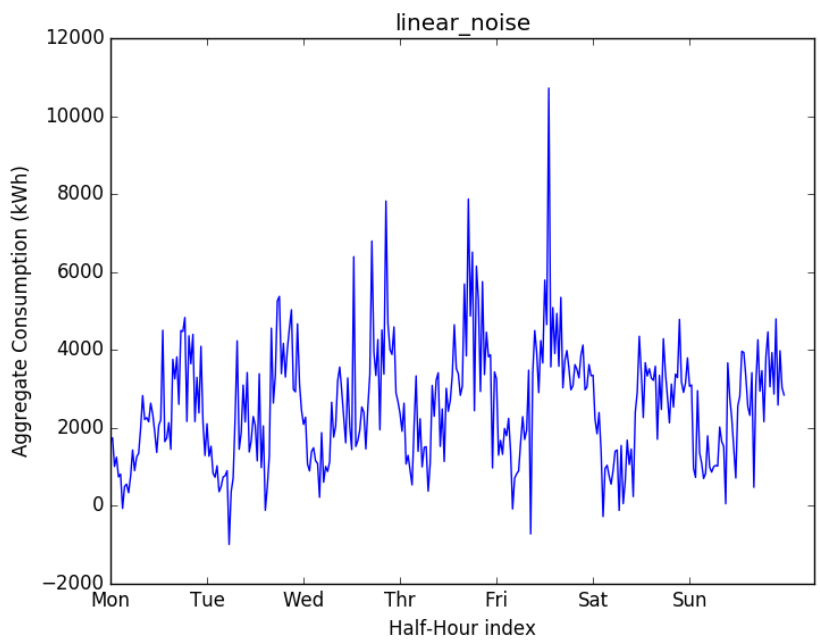

Figure 6.2: $(1,10^{-5})$-DP Linear Plot of Aggregation Consumption During Week of Christmas

We recovered some information that existed before noise injections. For instance, one could distinguish one day from another based on placements of consumption troughs. However, the plot contained too much noise for each measurement such that it lost many interesting features that we observed in Figure 6.1. Meanwhile, we could once again post-process the data without sacrificing privacy protection. For noise filtering purposes, the Savizky-Golay filter [13] is one of the most widely used filters, so we applied this filter to our visualization.

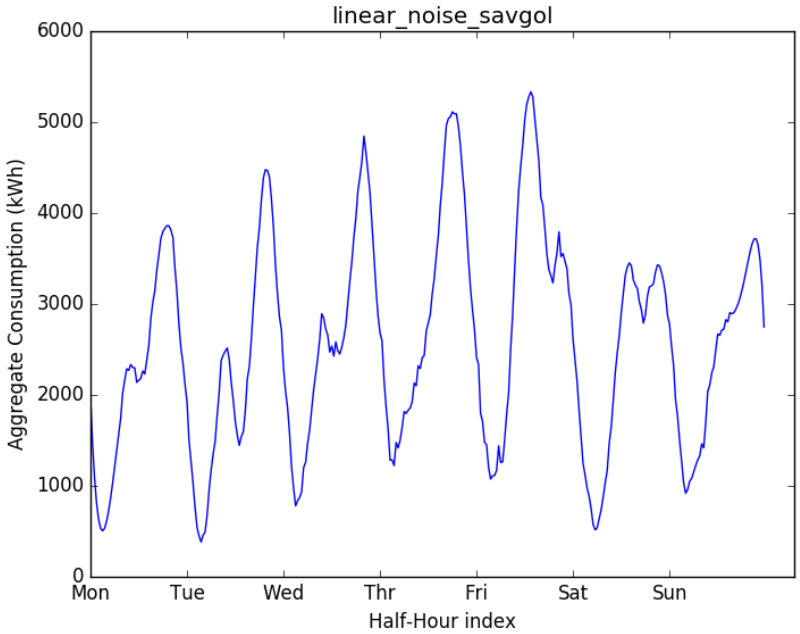

Figure 6.3: Applying SAVGOL filter to Noisy Linear Plot

After applying the filter in Figure 6.3, the visualization product was still (1,10^{-5})-DP Private since DP mechanisms are immune to post-processing. Nonetheless, viewers definitely gained more information from visualization after applying the filter. Now, end-users could confidently claim that there existed a small consumption spike during morning time before large spike in night time. Nonetheless, note that there existed some false information which did not exist in original raw product but existed in the filtered product. For instance, when one makes an observation for consumption during Saturday, he would conclude that the aggregate consumption during morning and night time are relatively similar to each other. Meanwhile, that was definitely false when we observed raw visualization product from Figure 6.1. For every day, peak consumption during night time was higher than that of day time.

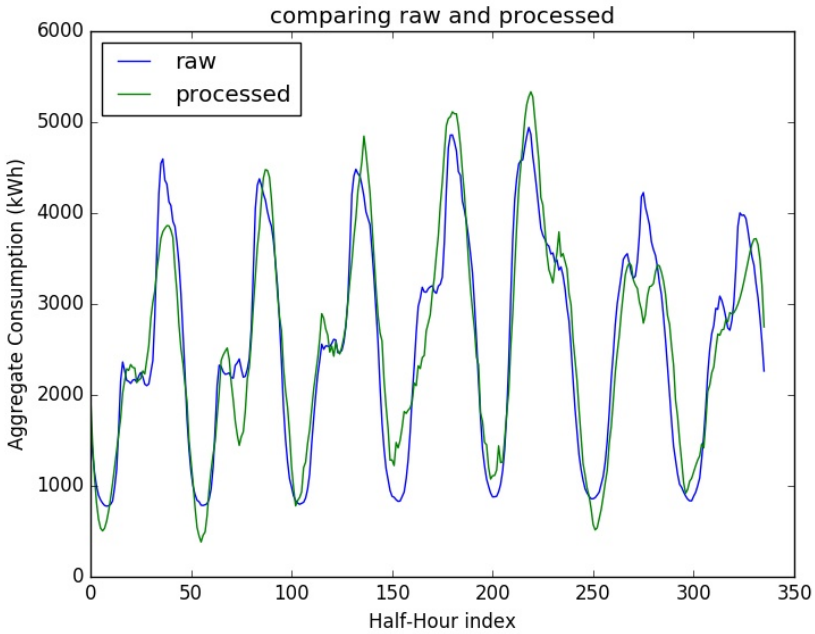

Figure 6.4: Comparing Raw and Processed Linear Plots

This figure illustrated degradation of information caused by noise injection and filtering. Although some macro-level information like daily peak consumption and daily trough consumption do not differ much, there existed significant difference in micro-level information where the exact shape of each day's consumption curve has changed after applying privacy protection mechanisms.

6.2 Scatterplots

So far, our visualizations plotted periodic data such that x axis represented time and day of measurement. Meanwhile, a scatterplot cannot plot such stream data, because a 2D scatterplot only allows two attributes per sample. Therefore, we would like to make different assumptions for this section of the experiment. For other visualization experiments we assumed that the producer of visualizations had no prior knowledge of the dataset other than its structure. For this experiment, we assume that producer of the scatterplot already read a heatmap of the dataset, so the producer already knows that the day of Christmas has higher electricity consumption than other Fridays or other days during the week of Christmas.

We created two scatterplots for this experiment. First, we plotted total electricity consumption from 1:00 to 1:30 PM period for December 18th (x axis) and December 25th (y axis). We knew that Irish people usually consumed less electricity during 1300 to 1330 period, but spent more electricity on 25th at the same time period. Next a scatterplot compared average electricity consumption of each consumer during December 18th and December 25th. We knew that people spent more electricity during Christmas than during other Fridays, so we drew a Line of Best Fit (LOBF) to see if the slope of the line was higher than one. Figure 6.5 and Figure 6.6 are the products.

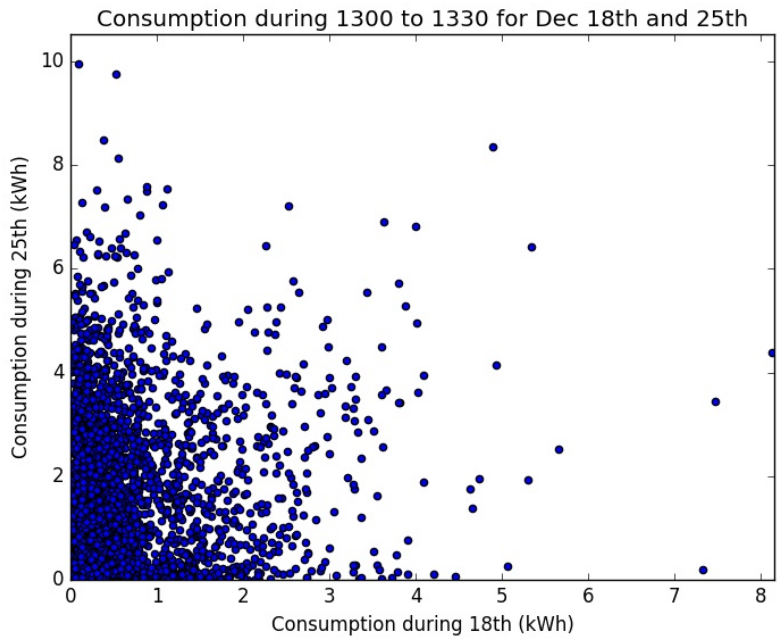

Figure 6.5: Scatterplot of Electricity Consumption during 1300 to 1330 for December 18th and December 25th

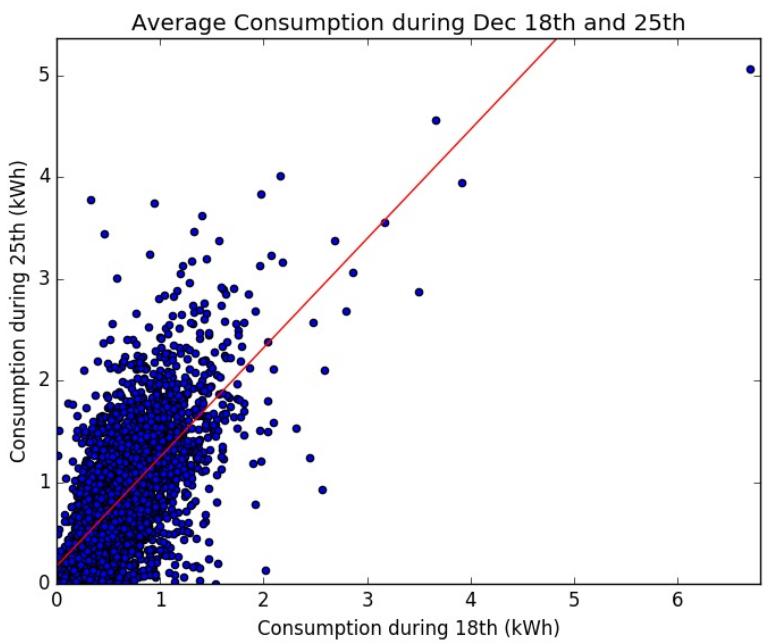

Figure 6.6: Scatterplot of Average Consumption for December 18th and December 25th

LOBF: $y = 1.07x + 0.17$

Figure 6.5 shows three groups of consumers. A group of consumers constituted the majority of samples where they spent more electricity on 25th than on 18th during 1300 to 1330 period. This was expected since 25th's aggregate consumption during 1300 to 1330 period was higher than that of 18th. Next group of consumers had very low electricity consumption for Christmas but had much higher consumption on 18th. We concluded that these consumers left their households for holiday trips during Christmas such that only minimal electricity for refrigerators or other home appliances was required. Finally we spotted some consumers who spent about the same electricity for both days. On the other hand, Figure 6.6 showed a linear pattern such that most consumers' average consumption fitted to the LOBF drawn as a red line. The slope of the LOBF was 1.07, indicating that consumers on average spent 7% more on Christmas than on other Fridays. Note that this scatterplot did not remove outliers yet. We will analyze how outliers impacted the visualization afterwards.

Implementing the DP mechanism on scatterplot required pre-processing procedures, because we could not directly measure sensitivity of each sample. Thus we decided to convert our scatterplot to a heatmap that was equivalent to a 2D histogram. Then we applied Laplacian noise to the heatmap via parallel composition. Finally we create scatterplot from the histogram by randomly drawing each bin's samples within the bin's area.

We implemented the Uniform Grid Algorithm [14] when converting the scatterplot to a 2D histogram. The algorithm partitions the space into $m \times m$ grid cells and adds Laplacian noise to each cell. Literature suggests $m = \sqrt{\frac{N\epsilon}{10}}$ provided the least amount of error, so we used following formula to divide scatterplots. Since this is equivalent to a histogram, the sensitivity of each bin is one and each consumer only contributes to a count of one bin. Thus, we used parallel composition for this experiment. Figure 6.7, 6.8 and 6.9 show process of creating $(1,\infty)$-DP counterpart of Figure 6.9.

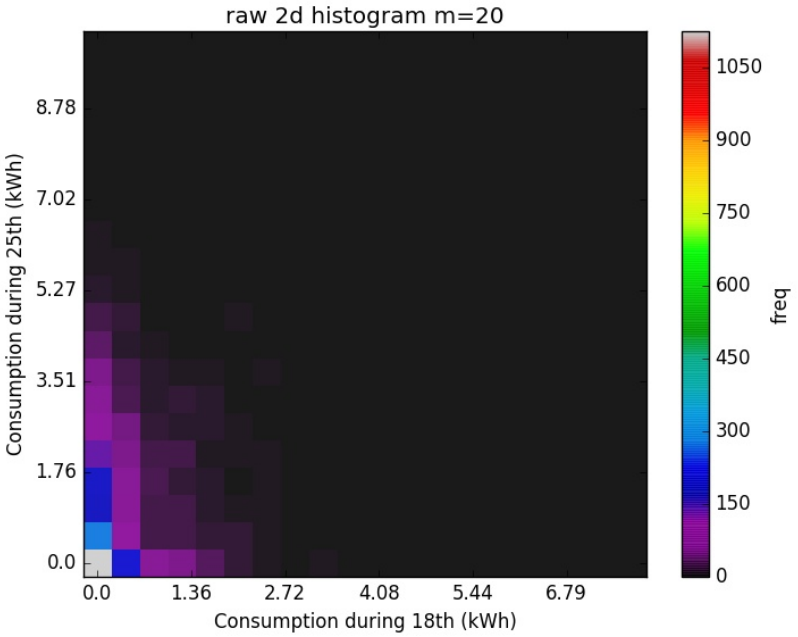

Figure 6.7: 2D Histogram Based on the Scatterplot from Figure 6.5

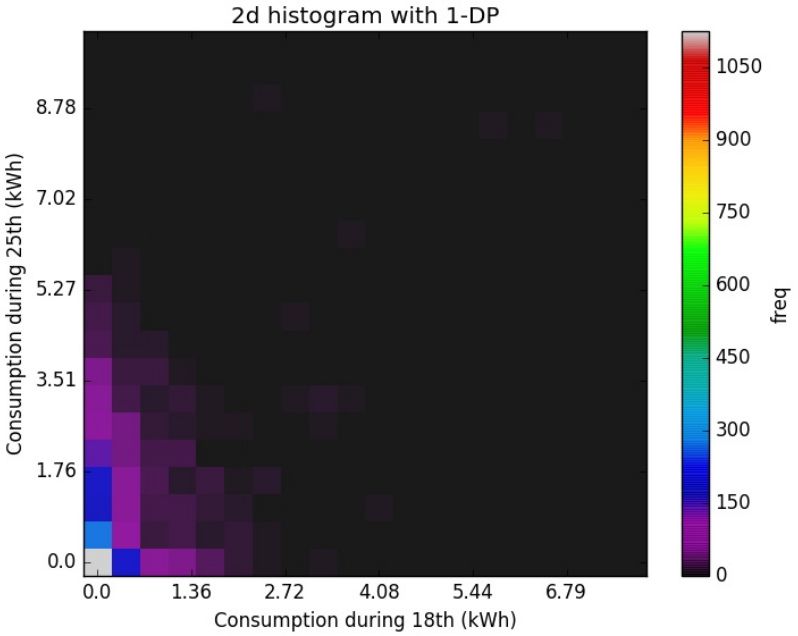

Figure 6.8: (1,∞)-DP Histogram of Figure 6.5

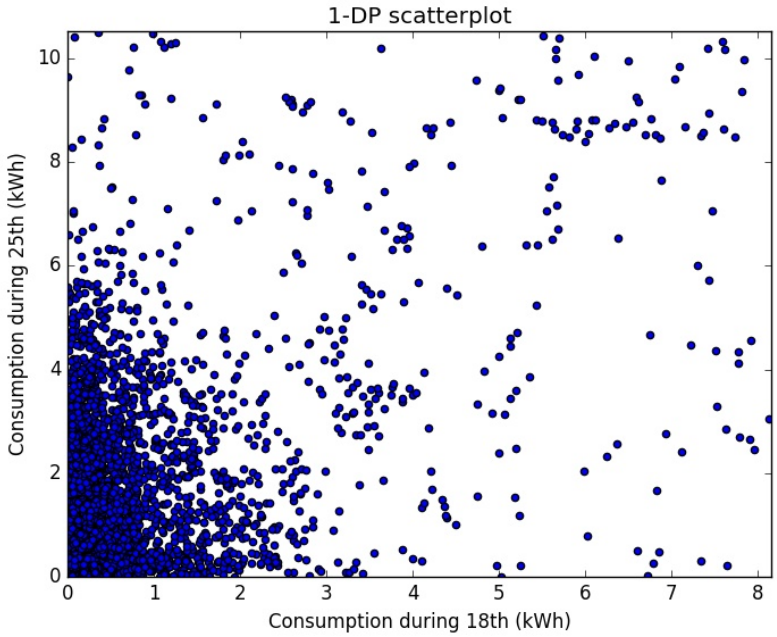

Figure 6.9: (1,∞)-DP Scatterplot of Figure 6.5

The protected scatterplot still shows a group of people who spent more electricity during Christmas and a group of consumers spent few electricity during Christmas. However, we found that the scatterplot showed a large number of outliers that were not present on its unprotected counterpart. Differential privacy aims to hide an individual in the dataset by adding an appropriate amount of noise. Meanwhile, outlier analysis aims to pinpoint individuals that do not fit to the dataset's normal trend. When converting unprotected scatterplot to 2D histogram, there existed many empty/low frequency bins since samples were grouped close to x or y axis. However when adding noise to the histogram, we added noise to empty bins such that we created artificial samples across an empty region to hide outliers. To best of our knowledge, there is no technique that can preserve outliers and satisfy (1,∞)-DP requirements. Thus, anomaly analysis must require unprotected visualizations or use synthesized databases.

Using the same algorithm, Figure 6.10 is the (1,∞)-DP Counterpart of Figure 6.6.

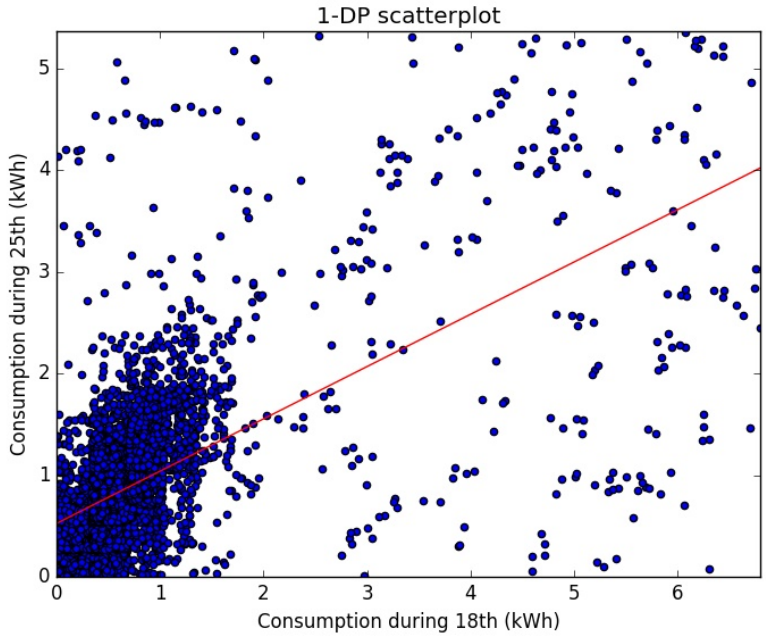

Figure 6.10: (1,∞)-DP Private Counterpart of Figure 6.6

While Figure 6.9 successfully showed major consumption trends of Figure 6.5, Figure 6.10's Line of Best Fit was tilted to the right after injecting artificial outliers. In this case, outliers scattered across the scatterplot affected the slope of the LOBF. Therefore, applying the same noise mechanism across all cells did not benefit our visualization. We either had to remove outliers when computing LOBF or use the other noise injection mechanism. Removing outliers for LOBF computation seemed to be a hard task, since it was hard to define outliers without viewing the original scatterplot. We could remove all samples that were not positioned near the crowd of samples at the left bottom corner, but there existed a large number of samples across the scatterplot. The number of spreaded points was too large to claim that all of those points were outliers. Thus, we incorporated the CBP mechanism [15] for this scatterplot. When we injected noise to the 2D histogram, we only added noises to bins with less than ten samples. For bins with less than ten samples, we suppress the bin count to zero. The mechanism removes outliers while ensures privacy of all consumers. The resulting scatterplot is shown at Figure 6.11 which is $(10, 1)$ Crowd-Blending Private.

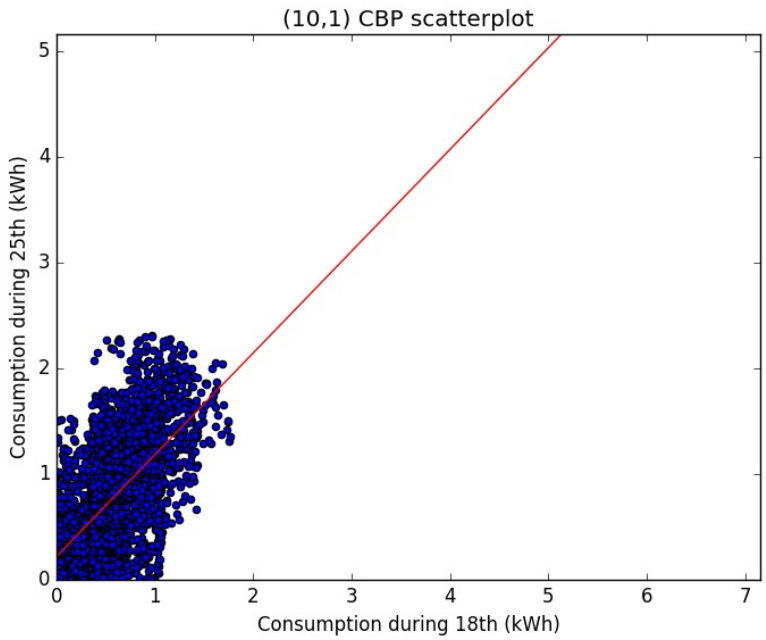

Figure 6.11: (10,1) Crowd-Blending Private Counterpart of Figure 6.6

LOBF: $y = 0.96x + 0.22$

The slope of LOBF decreased from 1.07 to 0.96. the CBP mechanism erased a number of samples that spent more electricity during Christmas than during the 18th of December, because the algorithm defined those samples as outliers. While aggregate electricity consumption during Christmas was larger than that of December 18th, majority of people spent higher electricity during 18th than during 25th. The mean consumption during 25th is larger than the median consumption. We thought a number of factors would have contributed to this phenomenon. For instance, two or three families could gather at their relative's place to hold a Christmas party. Then, there would exist two to three empty houses and one household with higher electricity consumption.

CHAPTER 7

LIMITATIONS AND FUTURE WORKS

Our project illustrated obscuring effects of visual cues caused by the Laplacian noise. However, we admit that our results would have strengthened if our experiments did not contain several limitations. In this section, we would like to discuss those limitations of the experiment.

7.1 Subjective Experimentation Methodology

Qualitative measurements are subjective as observations rely on observers' judgments. Observations and decisions are likely to be biased. Thus, studies based on qualitative measurements often require crowd-sourcing based user study for validity of results. Unfortunately, author was the only observer for this experiment. Our observations focused on evident visual cues such that we believe most readers agree with our qualitative observations. Nonetheless, lack of participants diminish the strength of our academic arguments. Thus, conducting a supplementary user study to confirm our findings is necessary. We would like to conduct a similar observation experiments to a large group of diverse crowds from crowd-sourced survey services like the Amazon Mechanical Turk (AMT) [16].

7.2 Decision Making based on Confidence Level and Costs

On the other hand, decision making process depends on multiple factors. Confidence level and cost-benefit analysis largely impact decision makers. While our preliminary questions introduced in the beginning of each experiment motivated readers to find and determine the visual cues, we did not

consider cost model and confidence level of observers. We would like to examine how people make different decisions on presence or absence of visual patterns depending on different environments which accompany different confidence levels and cost models. For example, visualizations involving stock market predictions and plots depicting temperature predictions would result in different degree of pressure and motivation for end-users to find the visual cues.

7.3 Complexity of Visualization

Modern visualization products often display multiple, diverse plots to efficiently deliver information to viewers. Multi-layer visualizations constitute intertwined information flow such that each layer is dependent on other plots. For these visualizations, directly implementing the Laplacian noise mechanism to numerical data queries would overlook a possibility of composing different layers to gain database sample's membership information. However, our study only focused on implementing DP noise injection mechanism to independent plots.

We only discussed non-interactive visualizations for the experiment. Our static productions did not receive or respond to user inquires. Interactive visualizations may require extra attention to privacy protection since this experiment did not account for uncertainties created by user inputs. We believe this subject is out of scope from this study, but we would like to prepare a different research experiment on interactive visualizations in the future.

7.4 Dataset Variety

Our study used one dataset provided by ISSDA. While this a year-long time series dataset provided motivating visualizations, our study was largely limited by the dataset. For instance, the dataset contained only 5,000 samples while each sample had 25,200 meter readings. If the dataset had larger sample size, we would not have encountered degradation problems throughout the experiment. Nonetheless, limitation of sample size also served us to present

novel pre and post processing algorithms for the project. We would like to conduct a separate study on a new time series dataset with large sample size to strengthen our findings. Much literature [17, 18] discussed implementation of DP mechanisms to protect patients' genomic data, and these authors encountered similar concerns that we observed. We would like to conduct a parallel study on those type of database to contribute our findings to their studies.

Progress can be made on comparing different visualization products. While our low degradation experiment compared visual cues of bar graphs and pie charts, our set of experiments is mostly disjoint. We believe that using a different dataset can mitigate this issue as each visualization technique is suitable for plotting different kinds of data. We are especially interested in comparing different visualizations to judge whether one type of visualization inherently provides more privacy protection than other types.

CHAPTER 8

CONCLUSION

Our work presented qualitative observations regarding the impacts of the Laplacian noise mechanism, as a DP mechanism, against a smart-grid dataset. We recognize limitations of our project, and we look forward to conduct a complementary user study which supplement our findings. We discussed how different visualization products resulted in various degrees of information loss when the Laplacian noise was injected to each visualization. We discussed three different levels of information loss with five different visualization products.

We discussed low degradation of visual cues with pie charts and bar graphs, and this experiment showed that bar graphs were more suitable for delivering information from the Irish dataset. Then we observed high degradation of visual cues with heatmaps. We concluded that no state-of-the-art DP mechanism could satisfy sufficient privacy protection and maintain key information for dataset with insufficient number of samples. Finally we discussed some visualizations with medium degradation of information. These visualizations required pre/post-processing efforts to maximize utility of visualizations. privacy-preserving linear plots maintained some visual cues from its unprotected counterpart, but some information was washed out after injecting noise and applying the noise-filtering algorithm. privacy-preserving scatterplot contained artificial outliers that obscured information. CBP scatterplot cleared these outliers and showed new information to end-users. While most visualization products only required noise injection to numerical data, some visualizations required pre/post-processing mechanisms to enhance utility of the products.

REFERENCES

[1] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[2] D. C. Barth-Jones, "The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now," 2012.

[3] C. Dwork, "Differential privacy," in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, vol. 4052. Venice, Italy: Springer Verlag, July 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/differential-privacy/ pp. 1–12.

[4] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo. retrieved may 17, 2013," 2010.

[5] C. Dwork, A. Roth et al., "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[6] F. McSherry, "Privacy integrated queries," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, Inc., June 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/privacy-integrated-queries/

[7] A. Dasgupta and R. Kosara, "Adaptive privacy-preserving visualization using parallel coordinates," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2241–2248, 2011.

[8] A. Dasgupta, M. Chen, and R. Kosara, "Measuring privacy and utility in privacy-preserving visualization," in *Computer Graphics Forum*, vol. 32, no. 8. Wiley Online Library, 2013, pp. 35–47.

[9] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 2005, pp. 901–909.

[10] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB Journal*, vol. 22, no. 6, pp. 797–822, 2013.

[11] G. Eibl and D. Engel, "Differential privacy for real smart metering data," *Computer Science-Research and Development*, pp. 1–10, 2016.

[12] "The ISSDA website." 2017. [Online]. Available: https://www.ucd.ie/issda

[13] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[14] N. Li, W. Yang, and W. Qardaji, "Differentially private grids for geospatial data," in *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013)*, ser. ICDE '13. Washington, DC, USA: IEEE Computer Society, 2013. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2013.6544872 pp. 757–768.

[15] J. Gehrke, M. Hay, E. Lui, and R. Pass, "Crowd-blending privacy," *Advances in Cryptology–CRYPTO 2012*, pp. 479–496, 2012.

[16] "The AMT website." 2017. [Online]. Available: https://www.mturk.com/mturk/

[17] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, 2014.

[18] M. Akgün, A. O. Bayrak, B. Ozer, and M. Ş. Sağıroğlu, "Privacy preserving processing of genomic data: A survey," *Journal of biomedical informatics*, vol. 56, pp. 103–111, 2015.