TopDown Differential Privacy

Travis Near
Northwestern University
United States
travisnear2023@u.northwestern.edu

Abenezer Tamene
Northwestern University
United States
AbeniTamene2023@u.northwestern.edu

Dennis Wu
Northwestern University
United States
hibb@u.northwestern.edu

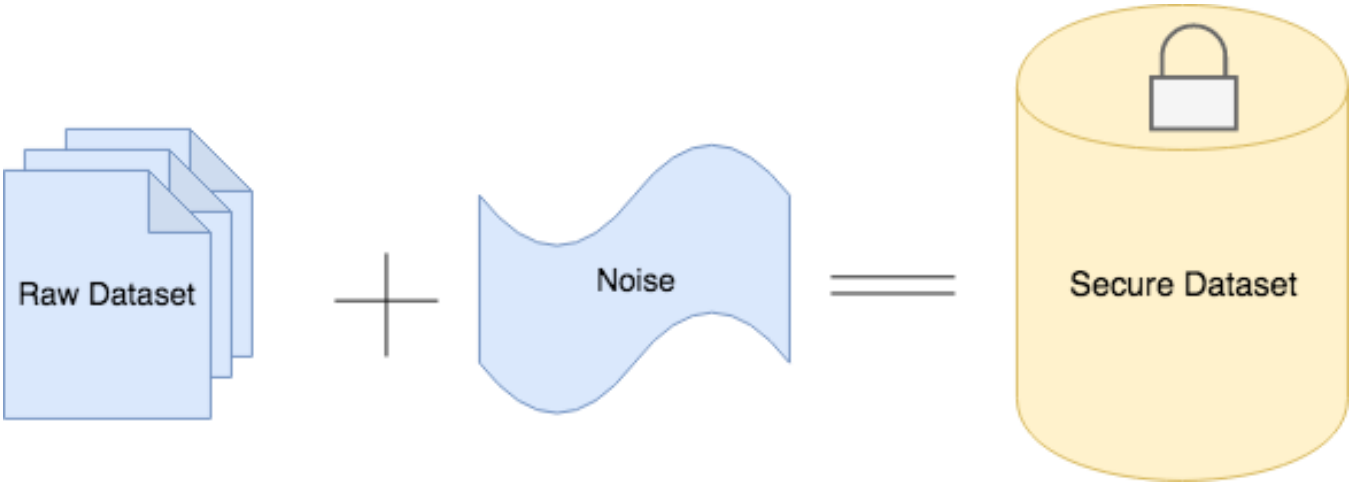

Figure 1: Differential privacy enables secure datasets. [9]

ABSTRACT

The TopDown Algorithm takes US census data as input and perform anonymization. Their approach transforms the data into a tree-like structure with different size/level of geographic constraints. However, graph data is more complex than trees due to its many-to-many relationships and therefore requires additional design. In our project, we implemented a simplified but applicable version of graph-based TopDown Algorithm. We used two types of post-processing to preserve the fairness and sanity. Also, we performed several simulations to observe the effectiveness of such algorithm on the Pokec Social Network dataset. [8]

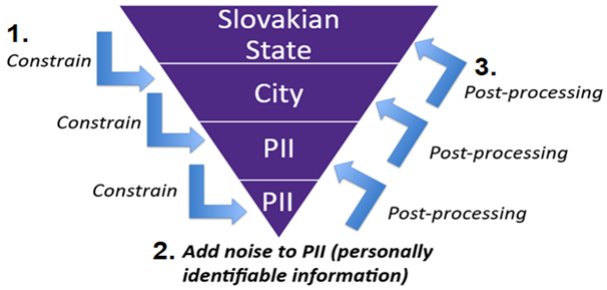

Figure 2: TopDown approach

1 INTRODUCTION

In 2020, the US Census introduced a new TopDown Algorithm (TDA) for adding noise to population counts to achieve differential privacy [1]. TDA has three high-level requirements: (1) breaking large data into smaller data ("Top-Down"), (2) adding noise to smaller data, and (3) post-processing to counteract any unfair imbalances ("Bottom-Up") [6]. The United States geography has rigid state, county, and tract boundaries which can be modeled as a tree with parent/child relationships. This rigid structure makes TopDown a natural approach for tree data.

Our challenge is to extend this TopDown concept to graph data. Our idea is to introduce constraints which end-users must use when querying our graph data. We developed three constraint hierarchies: (1) top, (2) medium, and (3) bottom. All three constraint hierarchies are required for our command-line interface.

When combined, these constraints filter the data in a way which clusters results. This behavior mimics zooming on the US Census's interactive tool [*link*] to see population counts. Figure 2 visually explains our approach.

By constraining the results into clusters, we reproduced the "Top-Down" aspect of scoping results to a region. Once we have clustered data, we add noise to the count. Lastly, to perform "Bottom-Up" privacy, we apply post-processing to the result for additional refinement. Finally, we return the noisy count to the client.

2 DATASET

We use the Pokec Social Network[8] graph dataset, containing the information about friendship relation between users on Pokec and their profile information. The attributes we used were listed below in Table 1. To limit the scope of our project, we sampled a subset in Pokec Social Network, which results in 20k nodes and 300k edges.

2.1 Data Cleaning Pipeline

Our data cleaning pipeline contains 4 steps:

(1) Translating data from Slovak into English
(2) Fill out missing/NAN data
(3) Simplifying user inputted attributes (hobbies, job types, etc.)
(4) Replacing incorrect data (900 cm height, 1 kg weight, etc.)

Attribute	Type	Example
User ID	numeric	0
Gender	discrete	0 or 1
Age	numeric	36 yr
Height	anonymized numeric	18X cm
Weight	anonymized numeric	7X kg
Region-Large	Text	Zilina Region
Region-Small	Text	Kyscuke Town
Hobby	Text	music

Table 1: Attributes we used in Pokec Dataset

We used the Google translation API for translation, which took about 47 hours to process all the data. Then, we replaced the missing/NAN data by randomly picking values from the same column. Next, we captured some of the most frequent keywords in each of the two attributes (hobby and job type), and then we used these keywords to replace the user-typed information. Such normalization helps us to obtain a better result for visualization and interpretation.

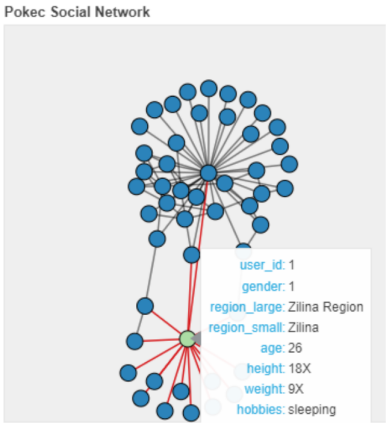

Figure 3: Visualization of Pokec Social Network Graph

3 BACKGROUND & RELATED WORK

3.1 The US Census

The 2020 US Census data collects the population in the US with different attributes, including their age, living area, race, etc. This dataset can be used for several purposes like voting age, scientific research, social science research and other related fields. It is collected so that it can provide an opportunity to identify the small area geography to conduct legislative redistricting [4]. The US census was also required to maintain the exact population count of each state. Therefore, the post-processing for certain accuracy is guaranteed in the 2020 US Census dataset.

3.2 The TopDown Algorithm

The TopDown Algorithm first transforms the US census data into histograms and then applies the top-down attribute to add noise on the lowest level of histogram. Then, it performs post-processing with a bottom-up approach with respect to the predefined hierarchy. They used concentrated DP to replace the standard DP to handle the loss of privacy with the increase of the number of queries on the same group of information [5]. In other words, once the user asked for a single query for multiple times, he/she can easily calculate the average of each query value and approximate the true value. The concentrated DP was introduced to handle such a problem. One major difference between the TDA and other database anonymization methods is they used a list of self-defined rules to preserve the logic and sanity of the data. For example, removing negative counts and preventing unrealistic attributes on certain types of people. Lastly, they predefined the privacy budget across different levels of attributes.

3.3 The Discrete Gaussian

Since both our dataset and the US Census were mostly handling "counts" of different groups of people, so we have to sample the noise from discrete distributions. In our project, we mostly following the equation proposed by [3], where they proved that the Discrete Gaussian provides the same concentrated DP guarantee as its continuous version.

4 METHODOLOGY

Our method can be seen as a simplified version of TDA [2]. Two major differences are (1) TDA predefined the distribution of the privacy budget, and ours divides the privacy budget by the query amount. (2) The US census contains many rule-based post-processing which are not revealed. Ours, however, uses a fairness balance and removes negative counts as post-processing.

For TDA, they divided the US census data into top-down levels with geographic attributes (states, county, cities, etc.). However, such a hierarchy was not applicable to graph data. In order to overcome such disadvantages, we take the user's input constraints and divide the graph into small clusters with the constraints.

The concentrated DP was defined as below, where $\frac{1}{2}\epsilon^2-$concentrated DP implies:

$$\frac{1}{2}\epsilon^2\text{CDP implies } (\frac{1}{2}\epsilon^2 + \epsilon \cdot \sqrt{2log(1/\delta)}, \delta) \text{ DP} \quad (1)$$

The workflow of our method is shown in Figure 4. After we accept the user's query and constraints, we then convert the DP parameter to its concentrated version in Equation 1. Then, we sample noise from the discrete Gaussian distribution [3]. Next, we start to apply post-processing. We first remove the rows with negative counts to control the sanity of noisy data, and then try to counterbalance unfair biases in the TopDown Algorithm.

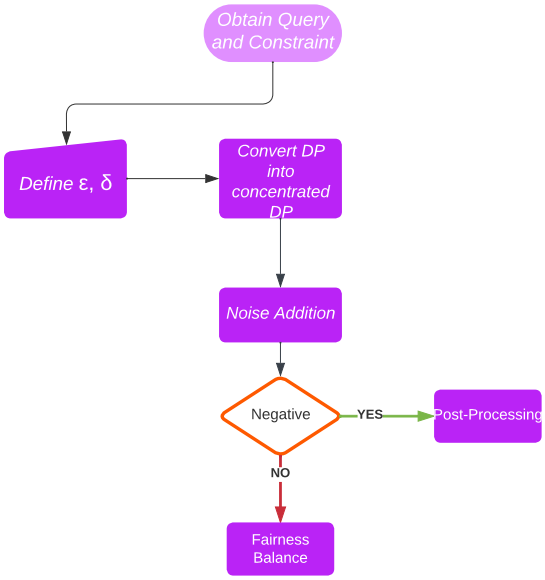

Figure 4: The workflow of our anonymization process

5 EXPERIMENT

Our experiment can be divided into two parts, graph related and traditional database DP.

We ran a simulation on the epsilon and utility tradeoff on discrete Gaussian distribution with delta = 0.1. It shows that we have the best tradeoff around 0.3-0.5 epsilon as shown in Figure 5.

Figure 5: Tradeoff between epsilon and utility

5.1 Degree distribution privacy

One property of the graph that we wanted to protect was the degree distribution. The more neighbors that a node has (i.e., more friends), the more the information that we can extract from it. Thus, as can be seen from Figure 6, we see that the higher the degree of a node, the higher the noise added. This, in turn, implies a lower utility.

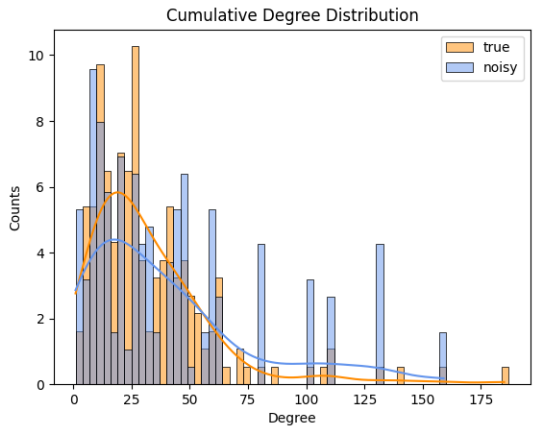

Figure 6: Cumulative degree distribution

5.2 Post-processing

The post-processing done by the US Census is a carefully orchestrated plan to re-balance inherent limitations with their TopDown approach. The vanilla TopDown algorithm unfairly inflates sparse data [1]. This means that their noisy measurements inadvertently "move" people from large cities to small towns despite $\mu = 0$. To counterbalance this unwanted bias, TDA assigns a privacy budget (denoted by ρ). This budget reflects how much privacy an entity is permitted to lose. Small clusters, for example, are granted a larger ρ value so that the net effect of post-processing is to "move" small population centers back to large. This post-processing step sometimes induces small towns to disappear. [7]

TDA has a sophisticated mechanism for the privacy budget which we simplified for our project. Our simplified implementation *[link]* assumes a *linear* relationship between (1) non-noisy count and (2) the amount of re-balancing required. For example, if the count includes 100% of people in a geographic region, then we know that this dense region will become sparser after adding noise. Therefore, to make it dense again, we add a *second* round of noise with $\mu > 0$. This μ value re-balances the shifted mean. Conversely, if the non-noisy count returns <1% of people in the region, we know our first addition of noise will inflate this number so our second round of noise needs to be deflated using $\mu < 0$.

6 DISCUSSION

6.1 Heat map

Figure 7 indicates the group differential privacy. When a user requests more queries at one time, we have to increase each individual query's privacy. For example, if we have k queries at one time, each single query must satisfy $\frac{\epsilon}{k}$ to preserve the overall guarantee. From the heat map, there is a trend we can see that with more query amount and lower ϵ value, the more noise we should expect to be added on each query.

The group differential privacy can be defined as the following: we say data $X = \{x_1, x_2, ...x_k\}$ containing k subgroups. We say an algorithm A satisfies $\epsilon-$differentially private on X if the following holds

$$\Pr[\mathbf{A}(x_i) \in Y] \leq \Pr[\mathbf{A}(x_i') \in Y]e^{\frac{\epsilon}{k}}. \quad (2)$$

where x_i and x_i' denotes two adjacent dataset and $i = 1 : k$. [5] From this inequality, we can see that to make the overall dataset X private to a certain degree, each of its subset has to satisfy a more strict privacy condition (lower ϵ).

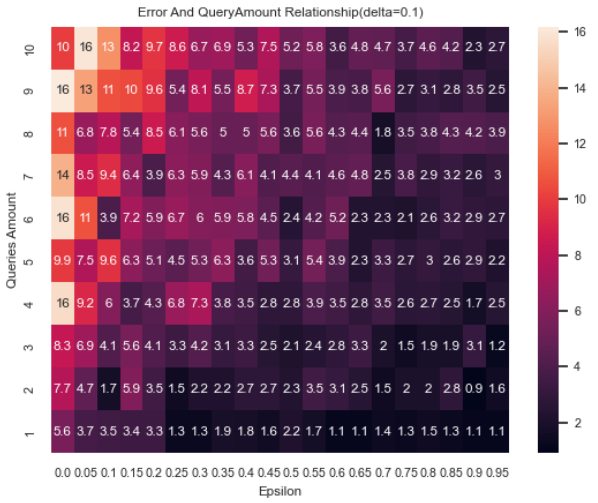

Figure 7: The impact on group DP on discrete Gaussian

6.2 Post-processing discussion

As detailed in the *Experiment* section, we performed two rounds of noise addition: (1) the initial noise to attain differential privacy, and (2) a second round of noise to counterbalance the tendency for large populations to "move" to smaller regions. We successfully achieved #1, however our implementation for #2 did not reproduce this trend. Figure 8 shows that, as we added weighted noise, we continued to lose accuracy. More importantly, accuracy was lost in a way which *changed* the overall trend. Our first noise addition (blue line) was impartial. Our second noise addition (orange line), however, tried to "counterbalance" a migration which did not exist in the graph. The net result is that big regions became bigger and small regions became smaller. From the perspective of a researcher, this would lead to misleading conclusions after querying our graph dataset. Based on this, we concluded that the second round of noise was unnecessary.

There are multiple ways to interpret our inability to reproduce this innate large-to-small population migration described by TDA:

- All datasets are unique. TDA is tailored for US Census data and was never designed for graphs. Our graph models social network connections and has little semblance to the type of data which the US Census analyzed.

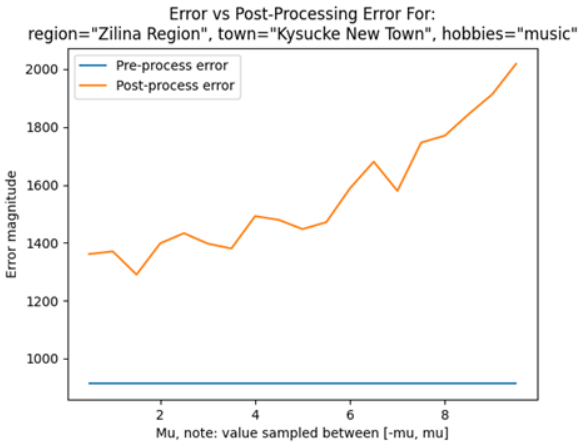

Figure 8: Error accumulation of noise

- We greatly simplified our implementation of the TDA algorithm. Any of our simplifications detailed in this report could have disturbed TDA's calibration.
- Lack of understanding of the TDA algorithm. We based our project on our understanding of the official research paper and corresponding video [6] rather than actual Census source code. Their paper and video did not convey every intricacy of the problems they faced.

7 LIMITATIONS

- One limitation we have in our implementation of the Top-Down approach is that unlike the US census bureau, we are not preserving some attribute value at some level (top, medium, bottom). The US Census, for example, preserves overall count at the state level.
- Our implementation is only restricted to concentrated DP. However, we might need a more relaxed guarantee for post-processing such as sanity preserving and removing negative counts.
- For graph differential privacy, we only consider node differential privacy. *Edge* differential privacy is out of scope for our project.

8 CONCLUSION

8.1 Challenges

The two main challenges we faced were reducing the scope of the project in a way that is applicable to graph data structures and cleaning the dataset. The original TopDown approach was implemented in the US census data which categorized geographical regions in a tree-like structure. We were able to mimic the original approach by enforcing top-level, medium-level and low-level constraints. The other challenge was cleaning the dataset. For this task, we had to translate the dataset from the Slovak language to English and then create realistic records by either filling out or changing null attribute values.

8.2 Implications and potential future work

The TopDown approach works well in hierarchical data structures. It allows us to preserve some attribute value at some level. In the US census data, population count was preserved at the state level [1]. Even though the algorithm has inherent limitations like "moving" people from large to smaller clusters, it can be corrected by using privacy budget which dictates how much privacy an entity is allowed to lose. Thus, based on the privacy budget chosen, we can tune the privacy vs. utility to desired levels. We experimented with a simpler mechanism to address the problem of "inflated" data by adding noise values to larger clusters since they become sparse when divided into smaller clusters. However, we were unsuccessful in increasing accuracy and additional work is needed to determine the culprit.

8.3 Work remaining

For analysis purposes, we tried to tune different variables to see their impact on privacy vs. utility. We analyzed the impact of the graph's degree distribution vs noise level, epsilon values vs error, the number of queries asked vs noise needed, and fairness adjustment vs error magnitude. However, we have not merged all these individual pieces to compute one noise value which we will use to return a computed function value to the user. Thus, finishing the tool end-to-end with testing is one extra step we need to take before it can be deployed.

9 SOURCE CODE

The source code can be found here on GitHub.

REFERENCES

[1] John M. Abowd. 2022. The 2020 Census Disclosure Avoidance System TopDown Algorithm. https://www2.census.gov/adrm/CED/Papers/CY22/2022-002-AbowdAshmeadCumingMenonGarfinkelEtal.pdf.

[2] John M. Abowd, Robert D Ashmead, Ryan Cumings-Menon, Daniel Kifer, Philip D. Leclerc, Jeffrey C. Ocker, Michael R. Ratcliffe, and Pavel I Zhuravlev. 2022. Geographic Spines in the 2020 Census Disclosure Avoidance System TopDown Algorithm. *ArXiv* abs/2203.16654 (2022).

[3] Clément L. Canonne, Gautam Kamath, and Thomas Steinke. 2020. The Discrete Gaussian for Differential Privacy. *ArXiv* abs/2004.00010 (2020).

[4] US Census. 2021. 2020 Census: Redistricting File (Public Law 94-171) Dataset. 2020Census:RedistrictingFile(PublicLaw94-171)Dataset.

[5] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation—TAMC* (theory and applications of models of computation—tamc ed.) *(Lecture Notes in Computer Science, Vol. 4978)*. Springer Verlag, 1–19.

[6] Michael Hawes. 2021. Differential Privacy 201 and the TopDown Algorithm. https://www.youtube.com/watch?v=bRIoE0rqwAw.

[7] Mike Schneider. 2021. People, homes vanish due to 2020 census' new privacy method. https://apnews.com/article/religion-wisconsin-new-york-tampa-florida-68c96e7eb701da74ae7c8df3c3476705.

[8] Lubos Takac. 2012. DATA ANALYSIS IN PUBLIC SOCIAL NETWORKS.

[9] Abhishek Tandon. 2019. Differential Privacy. https://medium.com/secure-and-private-ai-writing-challenge/differential-privacy-e5c7b933ef9e.