Bilkent University

Department of Computer Engineering

CS475 - Data Privacy

2021-2022 Fall

Final Report

Group Members

Osman Buğra Aydın	21704100
Ege Hakan Karaağaç	21702767
Mustafa Tuna Acar	21703639
Mehmet Alperen Yalçın	21502273
Ahmet Furkan Ahi	21501903
Fırat Yönak	21601931
İlhan Koç	21603429

Instructor: Erman Ayday

Final Report

January 9, 2022

This report is submitted to the Department of Computer Engineering of Bilkent University in partial fulfillment of the implementation of the Data Privacy course CS 475.

Github Repository: https://github.com/kiraci/CS475-DataPrivacy

Table of Contents

1 Problem statement and literature search

1.1 Problem Statement

Privacy is the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively. Privacy is a broad term that covers many fields. Location privacy is a privacy area which is the right of individuals to decide how, when, and for which purposes their location information could be released to other parties. It concerns location based activities of people and therefore life. Companies and enterprises need location information to make business decisions and improve their business strategies. It can also be misused by other parties.There are different ways of collecting location data such as using the network layer and the wireless networking to identify the user, using a person's mobile service provider and using mobile services from applications downloaded with permission of people. The data gathered using these ways can be shared with other adversaries. The possibility of misuse of this data brings the question of how to prevent location privacy issues from being a threat to our daily life. Our project aims to provide a solution for this problem by approaching location privacy by applying k-anonymity and addition of controlled noise to the datasets that include sensitive data that may harm a person's location privacy. Literature search will be completed before moving on the project implementation part to benefit from the other related works. Each one of these techniques used in the solution process will have different consequences, that’s why; after applying k-anonymity and addition of controlled noise, their result will be evaluated using the location privacy and utility metrics that are provided according to their utility and privacy performance.

1.2 Literature Search

1.2.1 K-Anonymity Related Search

K-Anonymity is based on the idea that by combining data with similar attributes, identifying information about each one of the individuals contributing to that data can be obscured. It is referred to as the power of ‘hiding in the crowd.’ Data of any person is put in a larger group, meaning information in the group could correspond to any single member, with that privacy of individuals are protected [1]. Applying k-anonymity to datasets in order to preserve a person's location privacy has been the subject of searches and different methods were used to apply this approach.

1.2.1.1 Anonymity Based Schemes

Clustering algorithm based k-anonymity:

Clustering is to partition the tuples with many clusters in which the features described the points are more similar to each other than points which are in different clusters. In order to increase privacy while providing a query service, a clustering k-anonymity algorithm is applied to the users location information and the outliers of a person location is eliminated [2].

Location based service architecture that supports k-anonymity :

Services may have an in-depth knowledge of the mobile users' whereabouts; by using this knowledge, these services may breach the privacy of the users; that’s why, concrete approaches should be provided to preserve the anonymity of the mobile users while LBS is requested. One of these approaches is k-anonymity. A framework with the functionality of providing the level of k-anonymity that the client specifies was developed using an architecture that has a message perturbation engine and an anonymity server [3].

1.2.1.2 Mobile Based Schemes

Mobile architecture based solution

With the proliferation of mobile devices such as smartphones and tablets, location-based services are becoming increasingly popular and due to the increasing use of social media, mobile services that take advantage of location privacy have also increased.[3] In order to provide a solution to the mobile location privacy, a mobile based architecture was introduced and compared with server based architecture using the privacy and utility metrics [4].

1.2.2 Differential Privacy Related Search

Geo-indistinguishability differential privacy for location based systems:

In order to protect the user's location by increasing the privacy level inside the user's location radius of r is achieved by adding controlled random noise to this r radius while keeping the utility aspect of the query result using a mechanism, differential privacy is utilised and privacy is ensured [5].

2 Planned methods

2.1 Location Perturbation

Location perturbation provides privacy to individuals with the false reported location. In this approach, the user location is represented with a wrong value. The accuracy and the amount of privacy mainly depends on how far the reported location from the exact location. This method is chosen because it will enable the project to prove the difference of just adding noise and adding noise in control. So it will be used for testing purposes in the implementation and the results of it in comparison with the controlled noise will be done with privacy and utility aspects.

2.2 Laplace Noise Addition

This is a more advanced approach to provide controlled noise addition. Laplace noise addition enables masking of important numeric data based on the noise that is distributed according to the laplacian distribution. Since it uses laplacian distribution, it provides better utility and privacy while compared to perturbation and it allows differential privacy. It's utility and privacy scores will be compared to k-anonymity using the metrics provided.

2.3 K-anonymity

Applying k-anonymity to a database will disable any person to be distinguished from at least k-1 other people in the same database since there will be people with the same quasi-identifiers. This can be achieved by generalization or suppression and in our project both of them will be used in order to achieve a k-anonymous state. After k-anonymity is achieved, the results will be compared to differential privacy using the privacy and utility metrics as comparison.

2.4 Spatial Cloaking

This technique is known as location cloaking, spatial blurring or location obfuscation, and the user location is represented as a region that contains the exact user location. An adversary knows that the user is located in the cloaked region, but has no clue where the user is exactly located. The area of the cloaked region achieves a trade-off between the user privacy and the user service. Spatial cloaking methods give results similar to k-anonymity and will be used in order to test the correctness of the k-anonymity by testing each of them using the same area for cloaking.

3 Proposed Solution

The main goal of the project is to store location data in a leakproof manner and to prevent data about a specific user or small group of users from being obtained from outside sources. For this, the techniques mentioned in the above section were used in detail in the project. In the rest of the section, information about how and where these techniques are used will be shared. Examples of what is used in current progress and implementation outputs will be shown. In addition, in the next section, how the various techniques used are compared using various metrics will be discussed, the evaluation of the obtained data will be roughly explained and the details will be left in the next section.

First of all, a literature search was conducted on the techniques to be used in the project. As a result of detailed research, it was decided how to adapt the 2 main techniques. The first is noise adding and the other is k-anonymity. However, before the implementation of these techniques, two more auxiliary techniques were used to further improve data privacy in an easy way. These are location perturbation and spatial clocking. These two techniques mentioned are simpler than the other two techniques written before, and they have been tried to be applied to all data in order to further improve data privacy. First of all, these will be mentioned briefly, then the noise addition and k-anonymity methods will be mentioned.

First, let's talk about the location perturbation and spatial clocking techniques we applied when we got the data.Location perturbation is one of the simplest and weakest concepts. The user location is represented with a wrong value, the privacy is achieved from the false reported location. The accuracy and the amount of privacy mainly depends on how far the reported location from the exact location [6]. This technique has been applied to every place data taken from the map, and the real location of the user is kept, but the data quality is still kept healthy. On the other hand, the spatial clocking technique is known as location cloaking, spatial blurring or location obfuscation, and the user location is represented as a region that contains the exact user location. An adversary knows that the user is located in the cloaked region, but has no clue where the user is exactly located. The area of the cloaked region achieves a trade-off between the user privacy and the user service [6]. This technique was used especially when examining the user density on the map, so that the density and identity of the users in certain regions are not completely disclosed. Instead of the exact position, approximate positions and intensity markers represented by various colors are used.

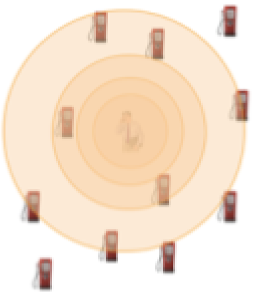


Figure 1: A spatial cloaking example

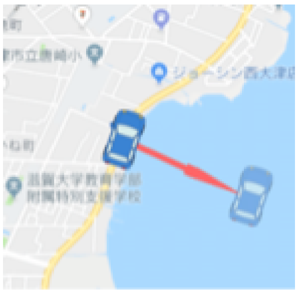
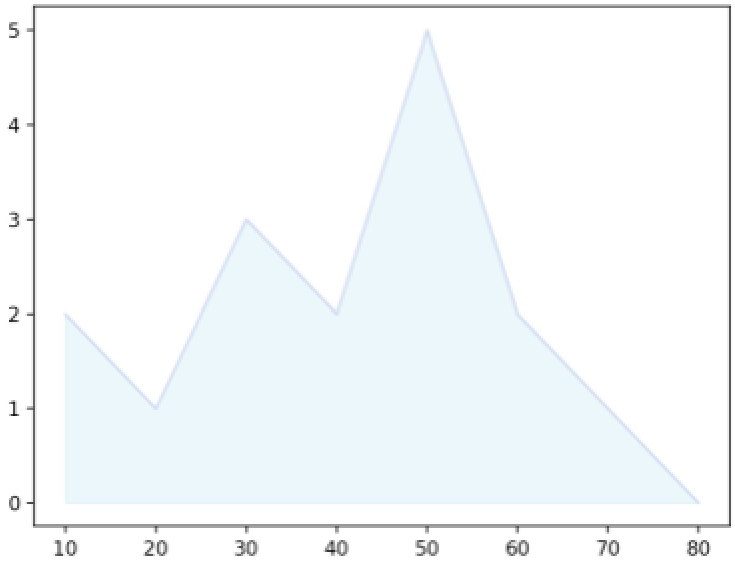


Figure 2: A Location Perturbation example

In this section, the use of the Laplace method added for noise addition in the code will be discussed. Other information about the Laplace method is given in the previous title. Here, its use in the code and what it does will be mentioned. Thanks to the noise added using the Laplace method, data belonging to certain users are hidden in accordance with privacy metrics. As can be seen in the graph, even if there is a single user at any point, a noisy data was sent to avoid revealing it. The use of the Laplace method in the code is as in the appendix A.



Graph 1: Data with thenoise addition

Another technique of ensuring data privacy is K-anonymity. It is aimed to hide the users under one roof by dividing the regions with available information into a specified number of sub-regions. The cloaked region contains at least k users and the user that asks the query is indistinguishable among other k users[3]. The cloaked area largely depends on the surrounding environment. For example, a value of k=100 may result in a very small area if the user is in a stadium but a huge area if he is in the desert [6]. The use of K-anonymity in the code is shown in the appendix. In addition, the literature information of K-anonymity is explained in detail in the previous title. In this way, every k user in the generalized block in the project cannot be distinguished from the remaining k-1 users. As shown in the figures below, the output was obtained by using the colors and parcels on the map. Creating a heat map on the map and compiling it with k-anonymous in accordance with data privacy is in appendix B-C.

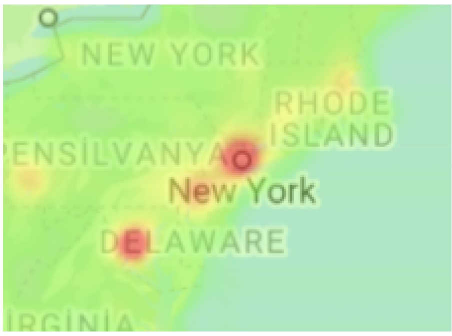


Figure 3

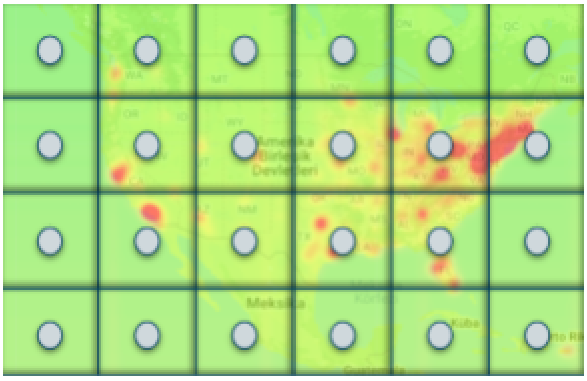


Figure 4: k Parselled map - output example

In the project, these two methods were compared according to 2 main metrics and various criteria below them. The first of these metrics is utility. Utility metric is divided into two subheadings as reaching individual location and counting distance & area. The second metric is the privacy metric. This metric also has two subsections: unlinkability and precision in error expectation. Discussion of these is left to the next section.

4 Evaluations

To compare the results of each one of the methods which is used in this project, some metrics are chosen. These metrics are grouped into two sub-categories as utility and privacy. The comparison between methods are made with outputs of the implemented methods. After the comparison of the methods, the most appropriate method for the metric is stated.

4.1 Utility

Utility metrics are the methods which are used to assess the performance of the methods or solutions. For this project two utility metrics and two sub metrics for each of two are chosen.

4.1.1 Accessing Individual Location

With this utility metric, the outputs of the methods are compared by their relative and real location data.

4.1.1.1 Relative Location

The relative location metric evaluates the performance of the methods when the data location nearby is needed such as a gas station, book store. The results of the comparison of methods shows that spatial cloaking is the most proper one which anonymizes the data and gives the quantity of wanted data nearby.

4.1.1.2 Real Location

The relative location metric evaluates the performance of the methods when the data is located into a data set without any relative location. The comparison of the methods shows that each one of the methods which are location perturbation, laplace noise addition, and k-anonymity could be used.

4.1.2 Count in Distance & Area

This utility metric is used to compare the results of the methods for location and area parameters.

4.1.2.1 Distance

The distance metric was used in the evaluation process to observe the performance of the methods when locating the places' distance from a certain location. According to comparison, spatial cloaking is the most effective method to use in contrast with the other methods.

4.1.2.2 Area

Applied methods were evaluated by the area metric to observe the performance of the methods while checking the quantity of locations in a certain place without knowing their exact location. The comparison outputs that k-anonymity is the perfect method to use for the area metric.

4.2 Privacy

The privacy metrics checks how secure the methods make the data in contrast with each other. For privacy evaluation, two use cases have been used as part of utility evaluation.

4.2.1 Unlinkability

The first use case of privacy iis unlinkability. With this metric,the privacy of the anonymized data was evaluated by checking if the data is linkable to its exact location or from a distant place. There are two parameters to check unlinkability which are distance and exact location.

4.2.1.1 Distance

For distance, spatial cloaking and location perturbation both anonymize the data while linking to a location because the implementation by adding random noise and using the heat map , the real datas possible location in a certain diameter without knowing its exact location.

4.2.1.2 Exact Location

For exact location case, K-anonymity and adding laplace noise anonymie location by putting each data into a certain square or point as it was done in the project which can be used to receive an anonymized location of asked data.

4.2.2 Precision in Error Expectation

This metric evaluates the methods which are implemented in the project by evaluating the accuracy of errors each method gives. The error expectation part has been divided into two categories as systematic and random errors, which are the parameters for this metric.

4.2.2.1 Systematic Errors

This metric category checks if the anonymized data gives accurate, similar errors. Laplace noise addition and spatial cloaking gives systematic error according to its real location because these methods secures the location of the data by a certain parameter or diameter.

4.2.2.2 Random Errors

Random errors checks if the anonymized data gives a random error without having any regularity. Using the location perturbation method in our experimental data gives the most random error because it simply changes the location in a random diameter.

5 Future work and conclusion

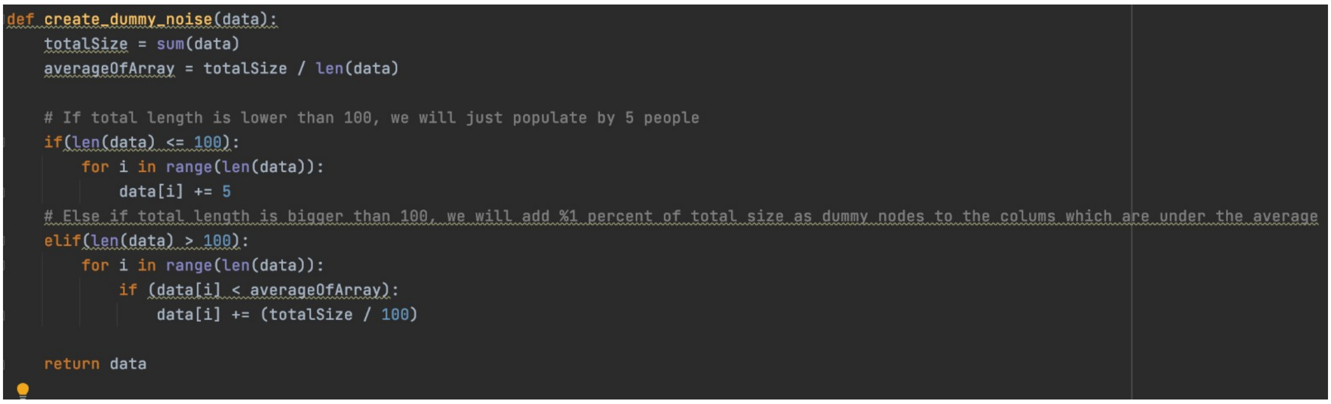
The concept of location privacy can be defined as the right of individuals to decide how, when, and for which purposes their location information could be released to other parties.[7] This paper evaluates the methods defined based on their utility and privacy capabilities. The methods defined are location perturbation, laplace noise addition, k-anonymity and spatial cloaking. Location Perturbation gives protection to people in the bogus revealed area. In this methodology, the client area is addressed with an off-base worth. The exactness and how much protection principally relies upon how far the revealed area from the specific area. Noise adding is a masking method for statistical disclosure control of numerical microdata that consists of adding random noise to original microdata. Moreover, K anonymity is a technique to release person-specific data such that the ability to link to other information using the quasi-identifier is limited. k-anonymity achieves this through suppression of identifiers and output perturbation. Also, Spatial Cloaking is known as area shrouding, spatial obscuring or area jumbling, and the client area is addressed as a locale that contains the specific client area.

In order to evaluate the methods mentioned on utility and privacy, four metrics were chosen. For utility, accessing an individual location and counting in distance & area were proposed. The data used for these metrics were computed in the algorithm for each method. For accessing individual location metric, the result for a relative location was achieved using the spatial cloaking and for real location it was achieved with location perturbation, laplace noise and k-anonymity. Count in distance metric gave the best results for spatial cloaking and for area k-anonymity was the best method in terms of area count.

For privacy, unlinkability and precision in error expectation metrics were chosen. Unlinkability was calculated with an algorithm which calculated the distance and exact location scores for each of the methods. For distance, spatial cloaking and location perturbation gave the highest distanced outcome. Exact location scores concluded k-anonymity and laplace noise as the highest exact location unlinkability outcome. Precision in error expectation calculations were done counting the systematic errors and random errors while applying the methods. Laplace noise & spatial cloaking were the best methods in terms of systematic error outcomes and for random errors location perturbation gave the least random errors. In conclusion, the method with higher privacy metric scores gave less utility metric scores and the utility metric leaders had worse scores in privacy.

6 Appendices

Appendix A



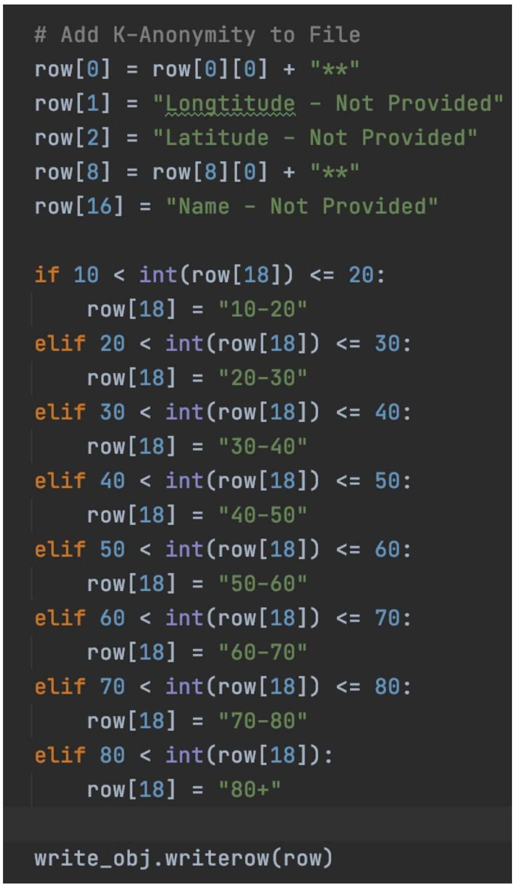
```python
def create_dummy_noise(data):
    totalSize = sum(data)
    averageOfArray = totalSize / len(data)

    # If total length is lower than 100, we will just populate by 5 people
    if(len(data) <= 100):
        for i in range(len(data)):
            data[i] += 5
    # Else if total length is bigger than 100, we will add %1 percent of total size as dummy nodes to the colums which are under the average
    elif(len(data) > 100):
        for i in range(len(data)):
            if (data[i] < averageOfArray):
                data[i] += (totalSize / 100)

    return data
```

This method adds dummy rows to the graphical solutions.

Appendix B



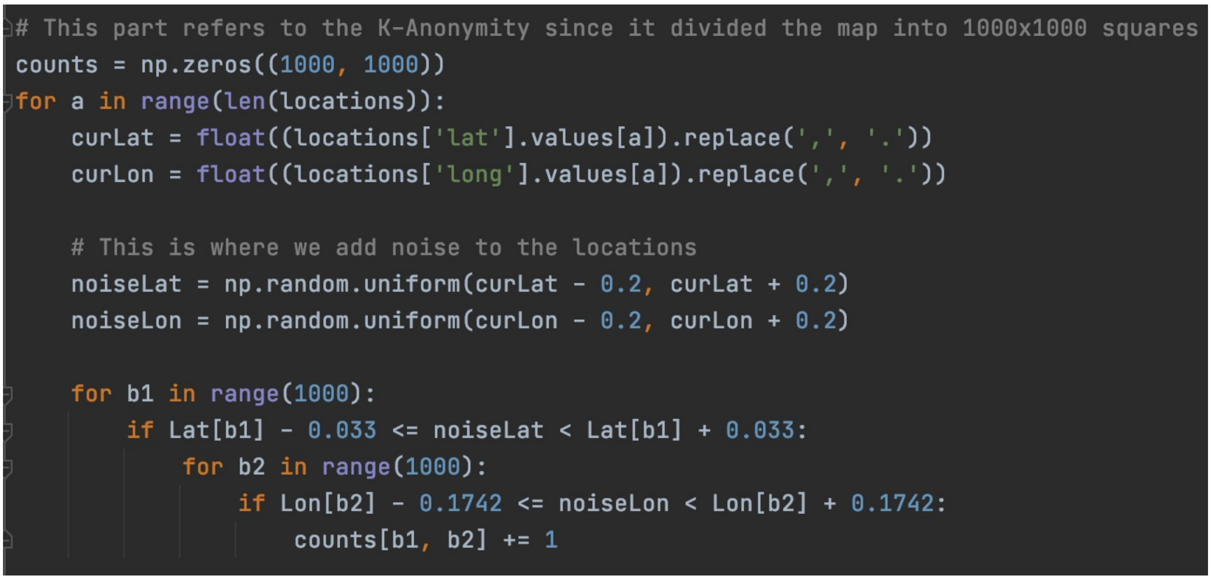
```python
# Add K-Anonymity to File
row[0] = row[0][0] + "**"
row[1] = "Longtitude - Not Provided"
row[2] = "Latitude - Not Provided"
row[8] = row[8][0] + "**"
row[16] = "Name - Not Provided"

if 10 < int(row[18]) <= 20:
    row[18] = "10-20"
elif 20 < int(row[18]) <= 30:
    row[18] = "20-30"
elif 30 < int(row[18]) <= 40:
    row[18] = "30-40"
elif 40 < int(row[18]) <= 50:
    row[18] = "40-50"
elif 50 < int(row[18]) <= 60:
    row[18] = "50-60"
elif 60 < int(row[18]) <= 70:
    row[18] = "60-70"
elif 70 < int(row[18]) <= 80:
    row[18] = "70-80"
elif 80 < int(row[18]):
    row[18] = "80+"

write_obj.writerow(row)
```

This piece of code creates the “results.csv” file by providing privacy.

Appendix C



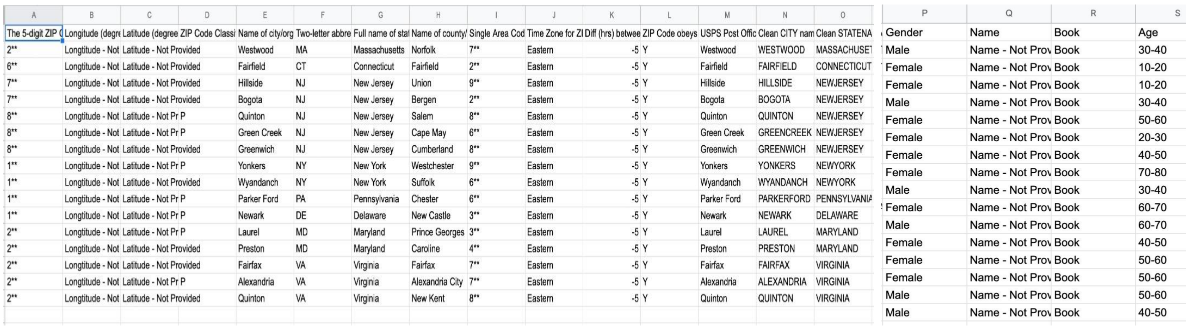
```
# This part refers to the K-Anonymity since it divided the map into 1000x1000 squares
counts = np.zeros((1000, 1000))
for a in range(len(locations)):
    curLat = float((locations['lat'].values[a]).replace(',', '.'))
    curLon = float((locations['long'].values[a]).replace(',', '.'))

    # This is where we add noise to the locations
    noiseLat = np.random.uniform(curLat - 0.2, curLat + 0.2)
    noiseLon = np.random.uniform(curLon - 0.2, curLon + 0.2)

    for b1 in range(1000):
        if Lat[b1] - 0.033 <= noiseLat < Lat[b1] + 0.033:
            for b2 in range(1000):
                if Lon[b2] - 0.1742 <= noiseLon < Lon[b2] + 0.1742:
                    counts[b1, b2] += 1
```

This code part belongs to the heat map implementation.

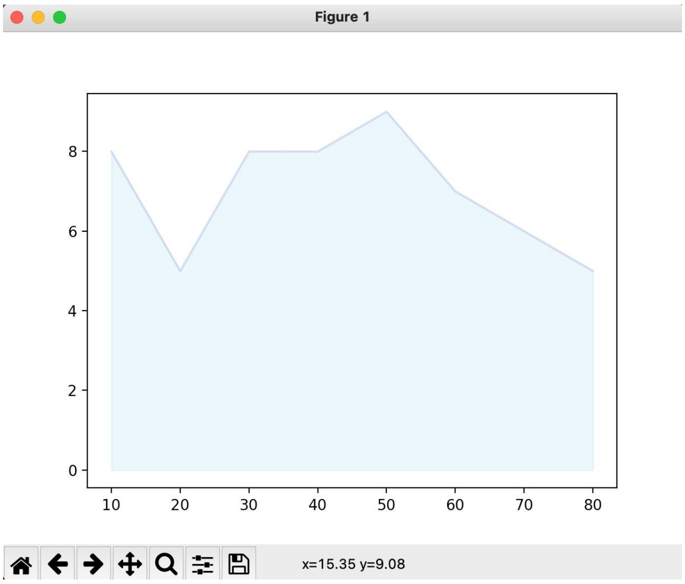
Appendix D

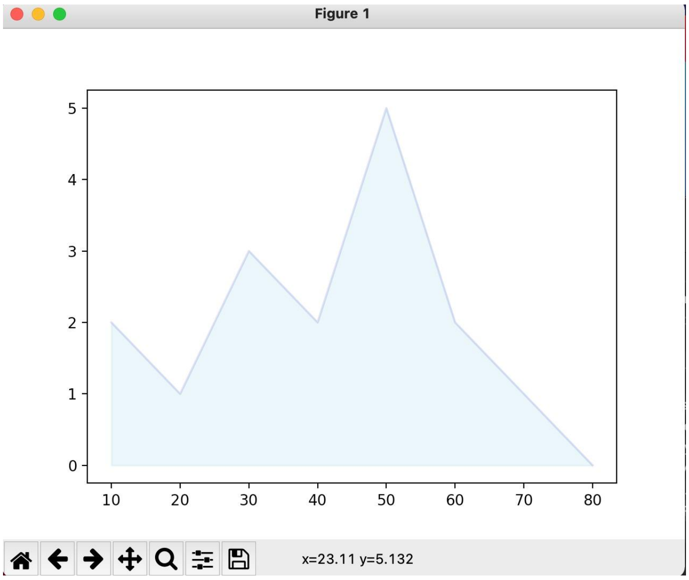


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
The 5-digit ZIP (	Longitude (degre	Latitude (degree	ZIP Code Classi	Name of city/org	Two-letter abbre	Full name of stat	Name of county/	Single Area Cod	Time Zone for ZI	Diff (hrs) betwee	ZIP Code obeys	USPS Post Offic	Clean CITY nam	Clean STATENA	Gender	Name	Book	Age
2**	Longtitude - Not	Latitude - Not Provided		Westwood	MA	Massachusetts	Norfolk	7**	Eastern	-5	Y	Westwood	WESTWOOD	MASSACHUSET	Male	Name - Not Prov	Book	30-40
6**	Longtitude - Not	Latitude - Not Provided		Fairfield	CT	Connecticut	Fairfield	2**	Eastern	-5	Y	Fairfield	FAIRFIELD	CONNECTICUT	Female	Name - Not Prov	Book	10-20
7**	Longtitude - Not	Latitude - Not Provided		Hillside	NJ	New Jersey	Union	9**	Eastern	-5	Y	Hillside	HILLSIDE	NEWJERSEY	Female	Name - Not Prov	Book	10-20
7**	Longtitude - Not	Latitude - Not Provided		Bogota	NJ	New Jersey	Bergen	2**	Eastern	-5	Y	Bogota	BOGOTA	NEWJERSEY	Male	Name - Not Prov	Book	30-40
8**	Longtitude - Not	Latitude - Not Pr	P	Quinton	NJ	New Jersey	Salem	8**	Eastern	-5	Y	Quinton	QUINTON	NEWJERSEY	Female	Name - Not Prov	Book	50-60
8**	Longtitude - Not	Latitude - Not Pr	P	Green Creek	NJ	New Jersey	Cape May	6**	Eastern	-5	Y	Green Creek	GREENCREEK	NEWJERSEY	Female	Name - Not Prov	Book	20-30
8**	Longtitude - Not	Latitude - Not Provided		Greenwich	NJ	New Jersey	Cumberland	8**	Eastern	-5	Y	Greenwich	GREENWICH	NEWJERSEY	Female	Name - Not Prov	Book	40-50
1**	Longtitude - Not	Latitude - Not Pr	P	Yonkers	NY	New York	Westchester	9**	Eastern	-5	Y	Yonkers	YONKERS	NEWYORK	Female	Name - Not Prov	Book	70-80
1**	Longtitude - Not	Latitude - Not Provided		Wyandanch	NY	New York	Suffolk	6**	Eastern	-5	Y	Wyandanch	WYANDANCH	NEWYORK	Male	Name - Not Prov	Book	30-40
1**	Longtitude - Not	Latitude - Not Pr	P	Parker Ford	PA	Pennsylvania	Chester	6**	Eastern	-5	Y	Parker Ford	PARKERFORD	PENNSYLVANIA	Female	Name - Not Prov	Book	60-70
1**	Longtitude - Not	Latitude - Not Pr	P	Newark	DE	Delaware	New Castle	3**	Eastern	-5	Y	Newark	NEWARK	DELAWARE	Male	Name - Not Prov	Book	60-70
2**	Longtitude - Not	Latitude - Not Pr	P	Laurel	MD	Maryland	Prince Georges	3**	Eastern	-5	Y	Laurel	LAUREL	MARYLAND	Female	Name - Not Prov	Book	40-50
2**	Longtitude - Not	Latitude - Not Provided		Preston	MD	Maryland	Caroline	4**	Eastern	-5	Y	Preston	PRESTON	MARYLAND	Female	Name - Not Prov	Book	50-60
2**	Longtitude - Not	Latitude - Not Provided		Fairfax	VA	Virginia	Fairfax	7**	Eastern	-5	Y	Fairfax	FAIRFAX	VIRGINIA	Female	Name - Not Prov	Book	50-60
2**	Longtitude - Not	Latitude - Not Pr	P	Alexandria	VA	Virginia	Alexandria City	7**	Eastern	-5	Y	Alexandria	ALEXANDRIA	VIRGINIA	Male	Name - Not Prov	Book	50-60
2**	Longtitude - Not	Latitude - Not Provided		Quinton	VA	Virginia	New Kent	8**	Eastern	-5	Y	Quinton	QUINTON	VIRGINIA	Male	Name - Not Prov	Book	40-50

This is the sample contents of “results.csv” file.

Appendix E





This is the result of input where

1. Interest area: Book
2. Latitude: -73
3. Longitude: 40
4. Distance: 120

Right hand side is the results that represent people's age who are found in the data by using Laplace noise to columns of longitude and latitude. Dummy rows have been added to the left hand side by Appendix A. As it can be seen, unique people like 20 years old in this table is anonymized.

7 Reference

[1] 14, H. D. A., & Devane, H. (2021, April 14). *K-anonymity: Everything you need to know (2021 guide)*. Immuta. Retrieved January 8, 2022, from https://www.immuta.com/articles/k-anonymity-everything-you-need-to-know-2021-guide/

[2] *K-anonymity location privacy algorithm based on clustering*. ResearchGate. (n.d.). Retrieved January 8, 2022, from https://www.researchgate.net/publication/321638316_k-Anonymity_Location_Privacy_Algorithm_Based_on_Clustering

[3] Gedik, Bugra, and Ling Liu. "Location privacy in mobile systems: A personalized anonymization model." In 25th IEEE International Conference on Distributed Computing Systems (ICDCS'05), pp.620-629. IEEE, 2005

[4] *Privacy protection for users of location-based services*. IEEE Xplore. (n.d.). Retrieved January 8, 2022, from https://ieeexplore.ieee.org/abstract/document/6155874

[5] Polytechnique, M. E. A. E., Andrés, M. E., Polytechnique, E., Nicolás E. Bordenabe INRIA and Ecole Polytechnique, Bordenabe, N. E., Polytechnique, I. N. R. I. A. and E., Konstantinos Chatzikokolakis CNRS and Ecole Polytechnique, Chatzikokolakis, K., Polytechnique, C. N. R. S. and E., Catuscia Palamidessi INRIA and Ecole Polytechnique, Palamidessi, C., Darmstadt, T. U., University, C. M., University, C., & Metrics, O. M. V. A. (2013, November 1). *Geo-indistinguishability: Differential Privacy for location-based systems*. Geo-indistinguishability | Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. Retrieved January 8, 2022, from https://dl.acm.org/doi/10.1145/2508859.2516735

[6] Privacy in location based services - part 1: ~elf11.github.io. Privacy in Location Based Services - Part 1 | ~elf11.github.io. (n.d.). Retrieved December 20, 2021, from https://elf11.github.io/2017/05/06/lbs-part-1.html

[7] *Location privacy*. Location Privacy - an overview | ScienceDirect Topics. (n.d.). Retrieved January 9, 2022, from https://www.sciencedirect.com/topics/computer-science/location-privacy