Applying Differential Privacy on Crime Report Dataset for Better Anonymization

Xueru Liu
Khoury College of Computer Science
Northeastern University
Vancouver, Canada
liu.xuer@northeastern.edu

***Abstract*—In the contemporary landscape of data collection, the imperative to safeguard individual privacy amidst extensive data collection has become paramount. This project delves into the domain of data privacy erosion, focusing on the vulnerability of public crime datasets that are not anonymized. Leveraging advanced Differential Privacy (DP) techniques, the project seeks to mitigate risks associated with the exposure of sensitive information. Drawing from recent studies on Differential Privacy, this project employs the Laplace mechanism to introduce privacy-preserving noise into various queries on the dataset. The differential privacy framework ensures that even a single piece of information does not unduly influence query outputs, safeguarding against re-identification. Implementation involves nuanced variations in epsilon values for different queries, highlighting the project's commitment to balancing privacy and utility. While successful in achieving commendable privacy levels, areas for future refinement, including epsilon calculations and specific information disclosures in crime descriptions, are identified.**

***Index Terms*—differential privacy, data anonymization, crime report, re-identification**

I. Introduction

The exponential surge in data has ushered in an era of extensive collection and analysis, offering unprecedented insights into individual lives. However, this transformative shift in data practices has concurrently exposed personal privacy to substantial risks, necessitating robust measures for the protection of sensitive information.

This project is dedicated to addressing the critical issue of data privacy erosion by employing cutting-edge Differential Privacy (DP) techniques. Specifically, the focus lies in anonymizing a publicly accessible yet un-anonymized crime report dataset. The paramount objective is to fortify the privacy and safety of the individuals identified in the dataset, ensuring the attainment of differential privacy and rendering any form of re-identification or privacy attack unfeasible.

The dataset under consideration is sourced from the Columbus, Ohio, Division of Police Unofficial Web Report Portal, encompassing 254 crime reports spanning the period from November 1st to November 3rd. Each entry comprises crucial information, including the CRNumber (Crime report number), Description (crime category), Victim name, Reported date, and crime case location.

Exploiting the blatant display of information in crime cases, adversaries do not require extensive efforts to obtain sensitive victim information. Beyond basic identification, adversaries may exploit vulnerabilities by seeking auxiliary information once victim names are acquired, enabling further targeted attacks based on the gathered information.

The unmitigated display of information in the dataset poses a substantial threat to the privacy of victims and those residing in the nearby area. This project is specifically geared towards shielding the privacy of victims, recognizing the CRNumber and individual's name as explicit identifiers that can potentially disclose sensitive information. Mishandling such information poses a genuine risk to privacy. Additionally, the link between the exact location of each crime case and the victim's name further exposes critical details. The disclosure of such incidents to the public not only compromises individual privacy but also constitutes a form of attack. Consequently, safeguarding victim-sensitive information is the primary focus of this project.

II. Relevant work

Differential privacy is a mathematical framework for ensuring the privacy of individuals in datasets. The core of a differentially private dataset is that the output of a function does not vary whether a record is present or absent from the dataset.

The concept traces back to 2006 when Cynthia Dwork introduced it in the paper "Differential Privacy" [1]. In this influential paper, Dwork proposed a mathematical framework for formally defining and achieving privacy in data analysis, coining the term ”differential privacy.” In the book ”Hands-on Differential Privacy”, the authors state that if a data release does not observably change when any one individual is added/removed/changed, then it provides immunity from reverse-engineering [2]. And in this new perspective, privacy is more a property of the function that computes the release.

There are some basic techniques to achieve differential privacy. The Laplace mechanism is one of them. As the name suggests, the Laplace mechanism simply computes f, and perturb each coordinate with noise drawn from the Laplace distribution. The scale of the noise will be calibrated to the sensitivity of f (divided by epsilon) [3].The epsilon parameter in the definition is called the privacy budget. Epsilon provides a knob to tune the ”amount of privacy” the definition provides. Small values of epsilon require f to provide very similar outputs when given similar inputs, and therefore provide higher levels of privacy; large values of epsilon allow less similarity in the outputs, and therefore provide less privacy.

To make a suitable tradeoff between privacy and utility, deciding how much noise should be acceptable in the release is crucial for the release to be useful, and what expenditure is acceptable for the individuals in the data.

III. Methods

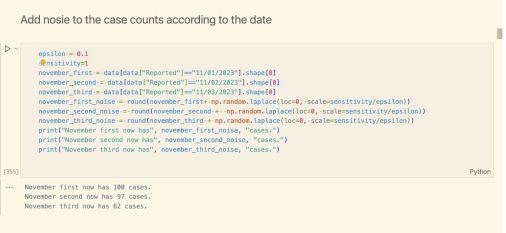
The dataset used in this project puts the victims’ privacy under great risk due to its lack of data anonymization. The goal of this project is to anonymize the dataset so that it satisfies differential privacy. To mitigate potential re-identification threats targeting the dataset, a primary solution involves filtering crime categories and obtaining counts for each category. The risk arises when unique categories have only one case, i.e., a single victim, as this scenario could facilitate re-identification. To counteract this vulnerability, the project implemented two strategic alterations to the data. I split the columns in the dataset into 2 categories, explicit identifers and non-explicit identifiers. Explicit identifers include CRNumber and victim names, and the others are non-explicit identifiers. For the three queries in non-explicit identifiers, the project implemented noise addtion, while for explicit identifiers it replaced the original data with the same generalized information.

A. Adding random noise with parallel composition

The implementation is in the format of code in jupyter notebook using the language of python to add noise to the query. In the jupyter file, where I conducted most of my implementations, I first read the original csv file, and get the columns in the csv file. There are in total 254 rows, 5 columns in the dataset.

First, I get the count values of crime cases by date. And I added noise to the output by selecting my epsilon to be 0.1, and sensitivity to be 1. I chose to use the laplace functions in numpy and random libraries in python to apply the Laplace mechanism. The epsilon value ensures that the modified result at most times is close enough to the original answer as the count by date solely doesn’t disclose too much private information. The original outputs for the 3 dates were: 104, 94, 56. And after we added noise, the outputs are: 100, 97, 62 as shown in Fig 1. The output changes every time the code runs, but most of the times the answer is close enough to the true counts.

Second, I calculated the number of reported cases by the category (“Description” in the csv). There are in total 54 categories in the 254 cases. In order to make the data differentially private, I added noise by defining the epsilon value to be 1, and sensitivity value also to be 1, which ensure that we achieve some privacy protection and preserve certain accuracy. The output difference is still quite small, simply adding a few decimals after the original ouput which are integers (see Fig.2).

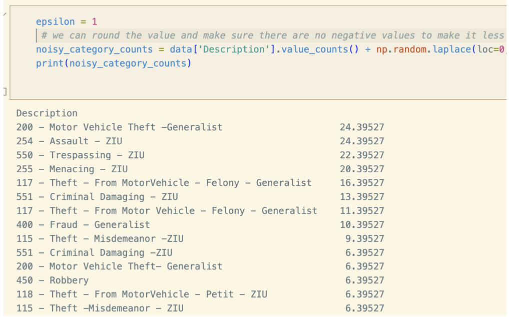


Add nosie to the case counts according to the date

```
epsilon = 0.1
sensitivity=1
november_first = data[data["Reported"]=="11/01/2023"].shape[0]
november_second = data[data["Reported"]=="11/02/2023"].shape[0]
november_third = data[data["Reported"]=="11/03/2023"].shape[0]
november_first_noise = round(november_first+ np.random.laplace(loc=0, scale=sensitivity/epsilon))
november_second_noise = round(november_second +  np.random.laplace(loc=0, scale=sensitivity/epsilon))
november_third_noise = round(november_third + np.random.laplace(loc=0, scale=sensitivity/epsilon))
print("November first now has", november_first_noise, "cases.")
print("November second now has", november_second_noise, "cases.")
print("November third now has", november_third_noise, "cases.")
```

```
November first now has 100 cases.
November second now has 97 cases.
November third now has 62 cases.
```

Fig. 1. Add noise to case count by date



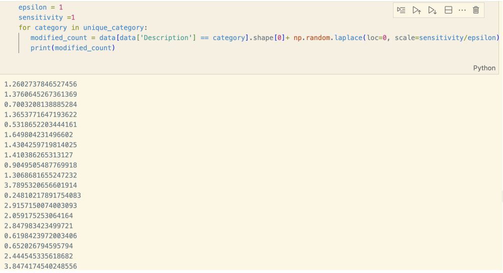
```
epsilon = 1
# we can round the value and make sure there are no negative values to make it less
noisy_category_counts = data['Description'].value_counts() + np.random.laplace(loc=0
print(noisy_category_counts)
```

```
Description
200 – Motor Vehicle Theft –Generalist                    24.39527
254 – Assault – ZIU                                      24.39527
550 – Trespassing – ZIU                                  22.39527
255 – Menacing – ZIU                                     20.39527
117 – Theft – From MotorVehicle – Felony – Generalist    16.39527
551 – Criminal Damaging – ZIU                            13.39527
117 – Theft – From Motor Vehicle – Felony – Generalist   11.39527
400 – Fraud – Generalist                                 10.39527
115 – Theft – Misdemeanor –ZIU                            9.39527
551 – Criminal Damaging –ZIU                              6.39527
200 – Motor Vehicle Theft– Generalist                     6.39527
450 – Robbery                                             6.39527
118 – Theft – From MotorVehicle – Petit – ZIU             6.39527
115 – Theft –Misdemeanor – ZIU                            6.39527
```

Fig. 2. Add noise to case count by category

When reading the category counts, we can see that there are some unique categories where the count value for the reported crime case is only 1. This is detrimental to the victim’s privacy because if one case gets withdrawn, re-identification of one individual can easiky be done. Therefore, I chose to add noise to the unique categories too. I first created a list for the unique categories. To prevent re-identification for the victims in the unique categories, I used an epsilon value of 1, and sensitivity of 1 to add noise to this query. And now when I try to get the each case count in the unique categories, instead of a list of “1”s, the output varies from 1.2602737846527456 to 0.70032081388885284 (see Fig.3).

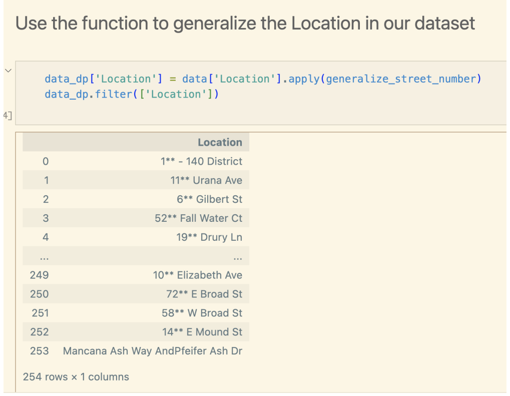
In this case, the loss of accuracy is acceptable since the safety of privacy is important. We cannot tell if the true answer is 0 or 1, because there’s too much noise to be able to reliably tell. But this is how differential privacy is intended to work - the approach does not reject queries which are determined to be malicious, but instead, it adds enough noise that the results



```
epsilon = 1
sensitivity =1
for category in unique_category:
    modified_count = data[data['Description'] == category].shape[0]+ np.random.laplace(loc=0, scale=sensitivity/epsilon)
    print(modified_count)
```

```
1.2602737846527456
1.3760645267361369
0.7003208138885284
1.3653771647193622
0.5318652203444161
1.649804231496602
1.4304259719814025
1.410386265313127
0.9049505487769918
1.3068681655247232
3.7895320656601914
0.24810217891754083
2.9157150074003093
2.059175253064164
2.847983423499721
0.6198423972003406
0.652026794595794
2.444545335618682
3.8474174540248556
```

Fig. 3. Add noise to each unique category



Use the function to generalize the Location in our dataset

```
data_dp['Location'] = data['Location'].apply(generalize_street_number)
data_dp.filter(['Location'])
```

```
                               Location
  0                     1** - 140 District
  1                         11** Urana Ave
  2                         6** Gilbert St
  3                      52** Fall Water Ct
  4                          19** Drury Ln
...                                    ...
249                    10** Elizabeth Ave
250                       72** E Broad St
251                       58** W Broad St
252                       14** E Mound St
253   Mancana Ash Way AndPfeifer Ash Dr

254 rows × 1 columns
```

Fig. 4. Hiding the last two digits in street number

of a malicious query will be useless to the adversary.

After these steps, in order to make the output public and reliable, post-processing measures need to be done. For example, the count values are presented as floats instead of integers, with occasional negativity. Therefore, before the data can be displayed to the public, it is essential that floats will be rounded to integers, and all negative values will be replaced with zero to guarantee reliability and utility.

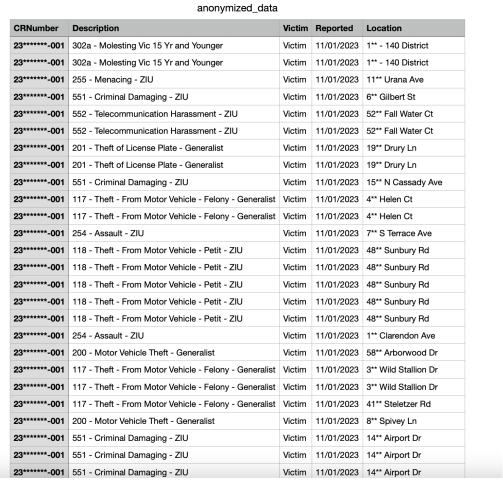
B. *Obfuscating explicit identifiers*

Following adding noise to each query, the next step involves creating a new CSV file to publish the anonymized data simultaneously. The chosen approach involves retaining the obfuscation of ”CRNumber,” ”Victim,” and ”Location,” while ensuring the absence of unique categories to avoid potential malicious attacks and re-identification risks. In this post-processing phase, the goal is to refine the dataset and alleviate any suspicions that might arise. This new csv file strikes a delicate balance between privacy preservation and the provision of trustworthy, accurate information to end-users.

For the two explicit identifiers, I replaced the original data generalized data so that no attack can be done with these two queries. For CRNumbers, all of them start with “23” and ends with “-001”. In between the start and end are the unique number combinations that can identify each specific crime report case. I used a lambda function (.apply(lambda x: x[:2] + ’*’ * 7 + x[9:])) to obfuscate the 7 unique digits so that no identification can be done. The final output of the CRNumber is: 23*******-001 (see Fig. 4).

For Victim, I replaced the victims’ true names with the word “Victim”. This ensures the absolute privacy but also in some degree is a trade-off for usability. Moreover, I also generalized the location by replacing the last two digits of the street number with 2 asterisks so that the information of the address is less revealing. I defined a function to replace the last two digits in the street number, which takes in the address as a parameter, and replace the last two digits with 2 asterisks, e.g. 52** Fall Water Ct.

As for the categories and date in the original dataset, I chose to add noise by duplicating the case of unique categories



anonymized_data

CRNumber	Description	Victim	Reported	Location
23*******-001	302a - Molesting Vic 15 Yr and Younger	Victim	11/01/2023	1** - 140 District
23*******-001	302a - Molesting Vic 15 Yr and Younger	Victim	11/01/2023	1** - 140 District
23*******-001	255 - Menacing - ZIU	Victim	11/01/2023	11** Urana Ave
23*******-001	551 - Criminal Damaging - ZIU	Victim	11/01/2023	6** Gilbert St
23*******-001	552 - Telecommunication Harassment - ZIU	Victim	11/01/2023	52** Fall Water Ct
23*******-001	552 - Telecommunication Harassment - ZIU	Victim	11/01/2023	52** Fall Water Ct
23*******-001	201 - Theft of License Plate - Generalist	Victim	11/01/2023	19** Drury Ln
23*******-001	201 - Theft of License Plate - Generalist	Victim	11/01/2023	19** Drury Ln
23*******-001	551 - Criminal Damaging - ZIU	Victim	11/01/2023	15** N Cassady Ave
23*******-001	117 - Theft - From Motor Vehicle - Felony - Generalist	Victim	11/01/2023	4** Helen Ct
23*******-001	117 - Theft - From Motor Vehicle - Felony - Generalist	Victim	11/01/2023	4** Helen Ct
23*******-001	254 - Assault - ZIU	Victim	11/01/2023	7** S Terrace Ave
23*******-001	118 - Theft - From Motor Vehicle - Petit - ZIU	Victim	11/01/2023	48** Sunbury Rd
23*******-001	118 - Theft - From Motor Vehicle - Petit - ZIU	Victim	11/01/2023	48** Sunbury Rd
23*******-001	118 - Theft - From Motor Vehicle - Petit - ZIU	Victim	11/01/2023	48** Sunbury Rd
23*******-001	118 - Theft - From Motor Vehicle - Petit - ZIU	Victim	11/01/2023	48** Sunbury Rd
23*******-001	118 - Theft - From Motor Vehicle - Petit - ZIU	Victim	11/01/2023	48** Sunbury Rd
23*******-001	254 - Assault - ZIU	Victim	11/01/2023	1** Clarendon Ave
23*******-001	200 - Motor Vehicle Theft - Generalist	Victim	11/01/2023	58** Arborwood Dr
23*******-001	117 - Theft - From Motor Vehicle - Felony - Generalist	Victim	11/01/2023	3** Wild Stallion Dr
23*******-001	117 - Theft - From Motor Vehicle - Felony - Generalist	Victim	11/01/2023	3** Wild Stallion Dr
23*******-001	117 - Theft - From Motor Vehicle - Felony - Generalist	Victim	11/01/2023	41** Steletzer Rd
23*******-001	200 - Motor Vehicle Theft - Generalist	Victim	11/01/2023	8** Spivey Ln
23*******-001	551 - Criminal Damaging - ZIU	Victim	11/01/2023	14** Airport Dr
23*******-001	551 - Criminal Damaging - ZIU	Victim	11/01/2023	14** Airport Dr
23*******-001	551 - Criminal Damaging - ZIU	Victim	11/01/2023	14** Airport Dr

Fig. 5. New anonymized dataset

so that the count becomes 2 instead of 1. Given that the other queries in the new anonymized dataset have already been masked, providing little utility, duplicating the unique categories instead of removing then preserves the overall accuracy in terms of category counts, but still altered the data where the total case count also changes, and so do the date counts. Since there are already cases where the count values are 2, the adversary will face difficulty distinguishing which category has been manipulated and which hasn’t.

IV. Results

Following the implementation of noise addition and the generation of a new CSV file (see Fig.5), our results, both for individual queries and the dataset as a whole, have been effectively anonymized.

As previously illustrated, the output for a single query exhibits variability each time the code is executed, introducing a layer of complexity by presenting results in floating-point values instead of integers. This deliberate variability enhances the difficulty of distinguishing between the actual and modified outcomes.

In the revised CSV file, where the two explicit identifiers have been replaced, potential adversaries are held back in their attempts to directly identify victims or specific cases. Furthermore, the broad generalization of location ensures a wider scope in representation without compromising the real-world locations. Crucially, the absence of unique categories precludes adversaries from executing a malicious attack through a single query, as these categories no longer exist.

The complete project code and new CSV file can be found on my public github repository dedicated to this project (https://github.com/lexiliew/differential-privacy-project).

While stringent measures have been taken to eliminate unique categories in the modified dataset, the overall data pertaining to crime descriptions, reported dates, and general locations retains its utility. This nuanced approach balances the imperative of safeguarding privacy with the continued relevance and usability of the anonymized dataset.

V. Future work

While the current project has achieved a certain level of differential privacy, it is acknowledged that the calculation of epsilon has not undergone meticulous design. This aspect requires further refinement to ensure accurate and appropriate determination of epsilon values, fostering a more robust privacy framework.

Additionally, for the forthcoming iterations of the modified CSV file aimed at publication, the current trade-off between privacy and usability will be revisited. Rather than rendering CRNumber completely unrecognizable, future efforts will focus on preserving some level of accuracy in the numbers while introducing the necessary noise for privacy protection.

Furthermore, an identified area for improvement lies in the handling of crime descriptions that inadvertently disclose information about victims, such as their age, as evident in instances like ”301 - Rape/Sexual AssaultVic 15 Yr and Younger.” Future enhancements will explore techniques like adding noise or generalizing these descriptions to bolster privacy measures.

To address these aspects effectively, the project could benefit from a more standardized approach to best practices in implementing differential privacy. This involves establishing clearer guidelines for determining epsilon and sensitivity values, ensuring a more consistent and reliable implementation across different queries and datasets.

In future endeavors, a critical step would involve providing a formal proof demonstrating the achievement of differential privacy after the implementation phase. This would contribute to the project’s credibility and provide a more comprehensive understanding of the privacy guarantees it offers.

Acknowledgment

I extend my deepest appreciation to Professor Logan for their exceptional guidance, unwavering support, and profound knowledge shared during the Privacy, Security, and Usability course. Their passion for the subject matter and commitment to fostering an environment of intellectual curiosity have been transformative for my academic journey.

I am immensely grateful to our Teaching Assistant, Kay, whose dedication, patience, and responsiveness have played a crucial role in enhancing my understanding of complex concepts. Her commitment to student success and tireless efforts in providing constructive feedback have made a lasting impact on my learning experience.

This academic endeavor has been a journey of growth and discovery, and I feel privileged to have been under the mentorship of such esteemed individuals. Their influence has not only deepened my understanding of privacy and security but has also inspired a commitment to excellence in my future endeavors.

I would also like to thank my friend Meredith and Dora, for supporting me whenever I run into difficulties and being there for me even though we are oceans apart. Your existence is a light to my life.

References

[1] C. Dwork, “Differential privacy,” in Lecture Notes in Computer Science, 2006, pp. 1–12.

[2] E. Cowan, M. Pereira, and M. Shoemate, Hands-On Differential Privacy: Introduction to the theory and practice using OpenDP. O’Reilly Media, 2024.

[3] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, Jan. 2013.