VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY

MASTER THESIS PROPOSAL

Major: COMPUTER SCIENCE

Topic:

PRESERVING PRIVACY FOR PUBLISHING-TIME-SERIES DATA WITH DIFFERENTIAL PRIVACY

STUDENT NAME: LẠI TRUNG MINH ĐỨC

STUDENT CODE : 2070686

INSTRUCTOR : Assoc. Prof. DANG TRAN KHANH

HCM City, May 2022

COMMENTARY FROM INSTRUCTOR

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

HCM City, …… ……, 2022

Instructor

COMMENTARY OF THE APPROVAL COUNCIL

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

CONCLUSION: *(tick X in the checkbox)*

☐ Approve

☐ Disapprove

Suggestions: .....

.....

.....

HCM City, …… ……, 2022

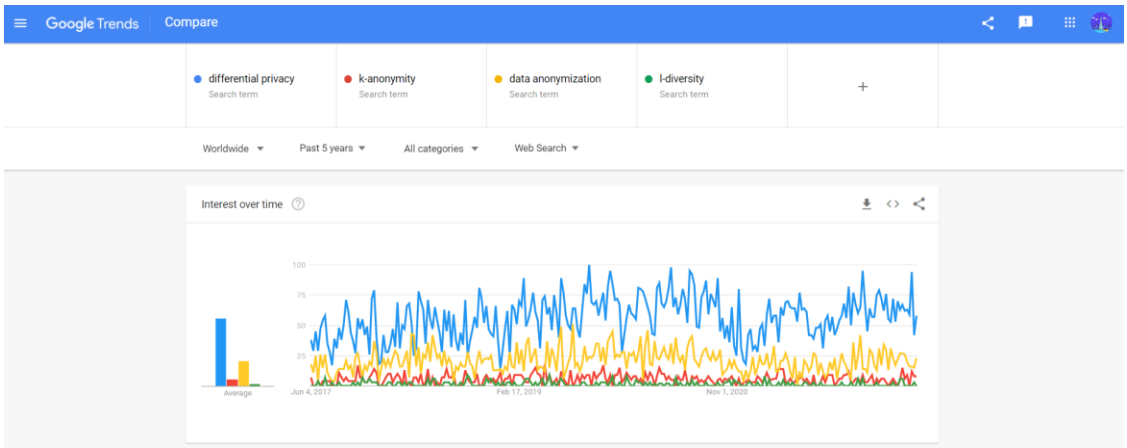
Approval Council

Table of Contents

General Introduction

The urgency of this study

The main reason of choosing this topic: **Preserving privacy for Publishing Time-series data with Differential Privacy** is to apply the knowledge in Differential Privacy with multiple mechanisms into the practical challenges in protecting privacy for Publishing Time-series data to outsourcing analytics services. The ultimate goal is protecting privacy of individual while keeping the Time-series data features *(auto-correlation, trends and seasonality, data distribution, and forecast-able)*. While some of the current solutions likes *Anonymization, k-Anonymity, l-Diversity* aren't enough for fighting against the re-identification process from attacker (mostly, with *linkage attack*), also it's not compatible with the concept of Big Data because of their complexity in algorithms; thereby, it is a huge demand on researching a new mechanism, and Differential Privacy is becoming one of the big research fields since 2006 when it first came.



The Google Search Trends for Differential Privacy - k-Anonymity - l-Diversity - Data Anonymization from 2017-2022 (Source: Google Trends - updated 29 May 2022)

From the context view, we are living in the world of Big Data, and we are providing our personal data to hundreds of applications every-day, which is most

of the time, we don't really aware what we are sharing to the application vendors. From the definition of *GDPR in Art. 4 Definitions, "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*"[@(gdpr)_2018]. And naturally, while getting into the digital world with smart devices/ smart applications/… we provide our data for the vendors to do account registration, account management, improve user experience program, and more.

For instances, if someone uses iOS/Android devices, then they provide their information to Apple/Google for account management; the company also tracks the behaviors on the device to improve the the user experience; improve the relevant advertisement visibility; and if they uses smart-watch devices, the vendor will also know about their daily heart-beat & daily-steps, which can easily inference to the activity of a whole day of the user. Moreover, every-time they used Facebook/Twitter/any-social-app, they provide photos/face identity/hobbies/interests/politics perspectives/… via direct posting or just a reaction button; then every-time they go to any websites in the Internet, surely they will be tracked with Google Analytics/Adobe Tag Manager/… or self-implemented tracker via cookies.

After the scandal of Cambridge Analytics and Facebook in 2018 (Confessore, 2018) released, people are more concerning on their privacy on digital platforms. No-one wants their data to be using without their consent, also, no-one wants to be known so-well and be manipulated / be driven - especially in the politics perspective.

To protect the privacy right of people, one-of the latest and strict regulation is released in 2016 under the name *General Data Protection Regulation (GDPR)* in European - effective for all European companies since May 2018. This regulation, in general, protect the right-to-control (and right-to-delete) personal data of EU-citizen on digital platforms. However, to comply with GDPR and other privacy regulations, the system (e.g: the health-app tracking heartbeat, the recommendation system,…) needs to be designed differently. Take an example of recommendation system, the algorithms will "remember who-is-who", and can be used to re-identify the user; which doesn't protect the privacy at all.

Additionally, the needs of sharing/publishing data between multiple parties for analytics purposes are increasing with the trends of Open Data, Outsourcing data analytics, and these needs also raise more concern about personal privacy - especially the temporal data (or time-series data) and the spatial-temporal data. These kinds of data are special and useful for companies to analyze, and getting insight with behavior of user, to create personalized program to engage and service. However, it also expose the risk of "knowing-too-much" from the analytics vendors.

In detail, the user accept to share their personal data for the application vendors, and they won't aware about the third-parties analytics vendor, and to ask for further consent from user to conduct analytics may costly and inefficient.
Another use-cases in this can be noted is the case of public-data sharing for public-analytics in traffic monitoring, hospital-sickness, trajectories and point-of-interests; although the data is anonymize with some techniques: *remove PII (Personal Identifiable Information); k-anonymity; l-diversity;…* sometimes it can be re-identify via information from other-public-database (in case of Netflix reviews competition in 2007 and case of Massachusetts Group Insurance Commission in 1997).

Therefore, it's obviously a big need for researching in protecting privacy, and Differential Privacy is an emerging research field for solving the privacy issue in Big Data and Machine Learning era - especially publishing Time-series data needs.

About The Thesis

Topic

Preserving privacy for Publishing Time-series data with Differential Privacy

Objectives

1. To study the core principles and multiple mechanisms of Differential Privacy on Time-series Dataset
2. To study the features and properties of Time-series Analysis; and evaluate the impact of data utility of Differential Privacy on Time-series data; and pointing out the use-cases of each algorithms.
3. To build a demonstration that using multiple mechanisms in Differential Privacy on the Time-series data from public-data-source (aggregated traffic data; revenue product sales;…)
4. To write a report and technical document for the demo of this study.

Methodology

1. Research from academic papers in Differential Privacy fields, and Differential Privacy for Time-series sub-fields; thereby, collecting and proposing a list of algorithms that fit the requirement to conduct the deeper analysis.
2. Research from academic papers and from corporate knowledge (in Finance and Supply Chain area) on Time-Series Analysis; thereby, reviewing the

data utility by extracting time-series features of privacy-protection versions versus the original one; then building use-cases of Differential Privacy for each time-series dataset features.

3. Based on the foundation and algorithms from #1 and #2; thereby, implementing multiple Differential Privacy mechanisms for Time-series, and conducting the comparison for those mechanisms on multiple Time-series data on Python.

4. Incrementally write up the report until completion as planned (16 weeks).

Thesis Layout

1. A brief introduction about the social motivation on Data Privacy in the Big Data era; the impact on individual privacy of Time-series analysis on Big Data; the benefits and risks of Publishing Time-series data for outsourcing analytics; and the Differential Privacy research area as the emerging research fields.

2. A detail analysis of current privacy-protection mechanisms: *Anonymization, k-Anonimity, l-Diversity, PCA,…*; the constraint of the algorithms in Big Data era; and some attack techniques to: *de-identification, re-identification, linkage attack, aggregation and statistics,….*

3. A detail introduction Differential Privacy and its state-of-the-art techniques *(Laplace noise, Gaussian noise, Kalman filtering,…)* on Time-series Dataset.

4. Proposing and implementing multiple Differential Privacy techniques for Time-series data on Python.

5. Conducting the comparison between multiple Differential Privacy techniques and auditing the data utility of the output with Time-series

Analysis; then pointing out the characteristics of dataset to choose the best fit algorithms.

6. The conclusion of the study for selected topic.

Implementation Plan

Week	Task	Time
	Thesis proposal defense	**Jun 2022**
W1 to W2	- Conduct the literature review and methodology to conduct the study. - Define scope of work for the main research of the thesis- Write up the report	2 weeks
W3 to W4	- Research of related works/projects on Differential Privacy, Time-series privacy- Write up the report (cont.)	2 weeks
W5 to W14	- Implementing the state-of-the-art algorithms of Differential Privacy on Time-series data- Comparing those algorithms with data utility metrics- Finding the data characteristics to choose the best algorithms- Write up the report (cont.)	10 weeks
W15 to W16	- Finalize the solution package- Finalize the document- Prepare the presentation	2 weeks
	Thesis defense	**Dec 2022**

Related Works Review

Traditional techniques to protect privacy

Anonymization

From the definition of **Pseudonymization in ISO 25237:2017 Health informatics** ("ISO 25237," 2022), it has been defined as a "process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party."

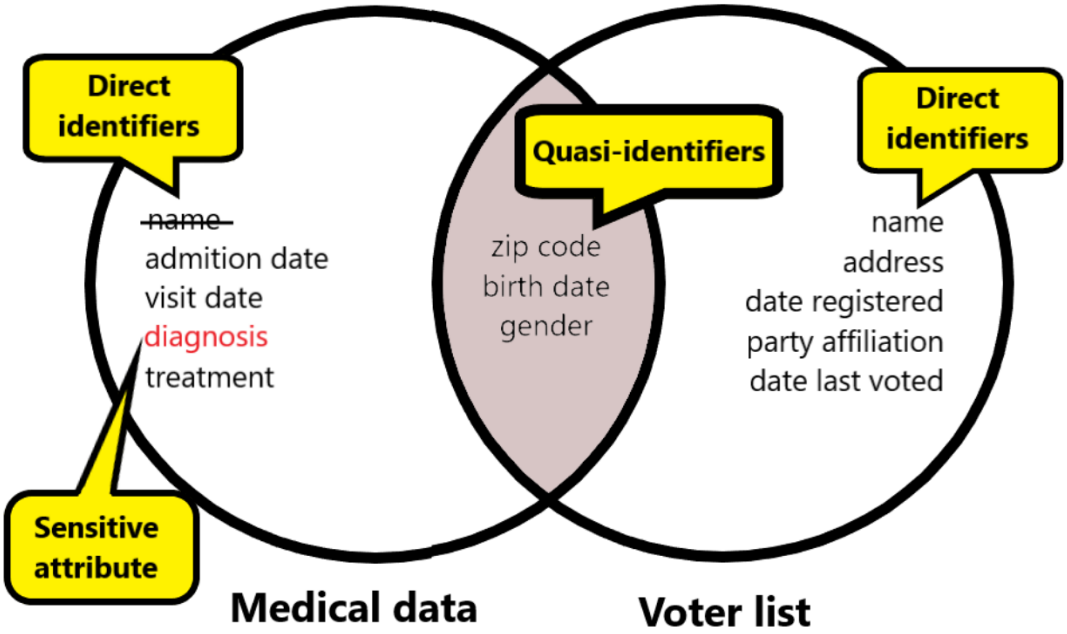
To conduct this process, we will try to remove the *Personally Identifiable Information (PII - e.g.: email, full-name, ID card number, credit card number, phone-number, full-address…)* from the dataset, then we can publish this data to the public and assume no-one can find who-is-who. This technique has been used for a long time, but it is not a good technique for now. To prove that statement, from the research of (Sweeney, 2000), she concluded that 87% of American people can be identified with just Gender - Birthday - ZIP code. This is called linkage attack (or background knowledge attack) - when the attacker knew "something" about their target (This will be discussed in some next section).

k-Anonymity and l-Diversity

The k-Anonymity was first introduced by (Sweeney, 2002) in a paper published in 1998 as an attempt to solve the problem: *"Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful."*

The formal definition of k-Anonymity is: at least k individuals in the dataset share a set of attributes that might become identifying for each individual, such that each member of that k-group shares the same quasi-identifiers (a selected subset

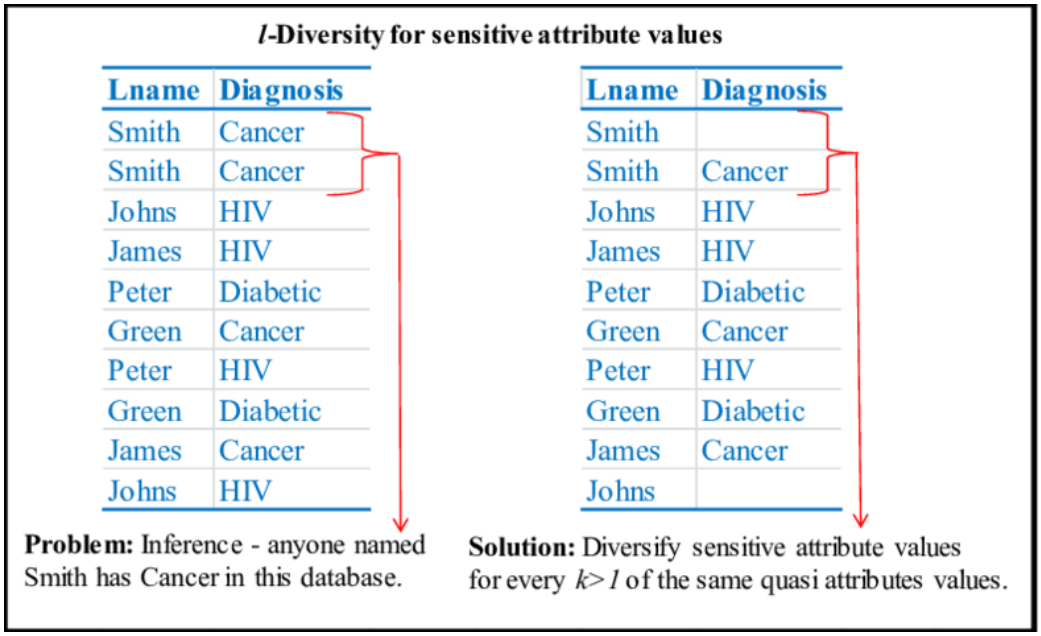
of all the dataset columns) with all of other members of the group. It means, each individual in each group will blend into their group, so it makes harder to identify individual in that k-group.



The terms example for k-Anonymity in the case of Medical records & Voter records

The drawback of k-Anonymity is the extensive computational with complexity level is $O(n^2)$ (which is not fit for Big Data era), also it's a NP-hard problem. Besides, k-Anonymity can be attacked when the data have outliers/abnormal data point or data have only identical records (that is called Homogeneity attack).

To enhance the privacy level of k-Anonymity in the case of Homogeneity attack, there is a suggestion that called l-Diversity - proposed in 2006 by (Machanavajjhala et al., 2006a) from Cornell University. The main idea here is: l-Diversity states that each bucket must have at least **l-distinct-sensitive-values**. However, the usage of l-Diversity is limited because of its worser data utilization ("applying 3-diversity dataset was worse than using 100-anonymity") that was mentioned in the study of (Brickell & Shmatikov, 2008) from University of Texas at Austin in 2008 .



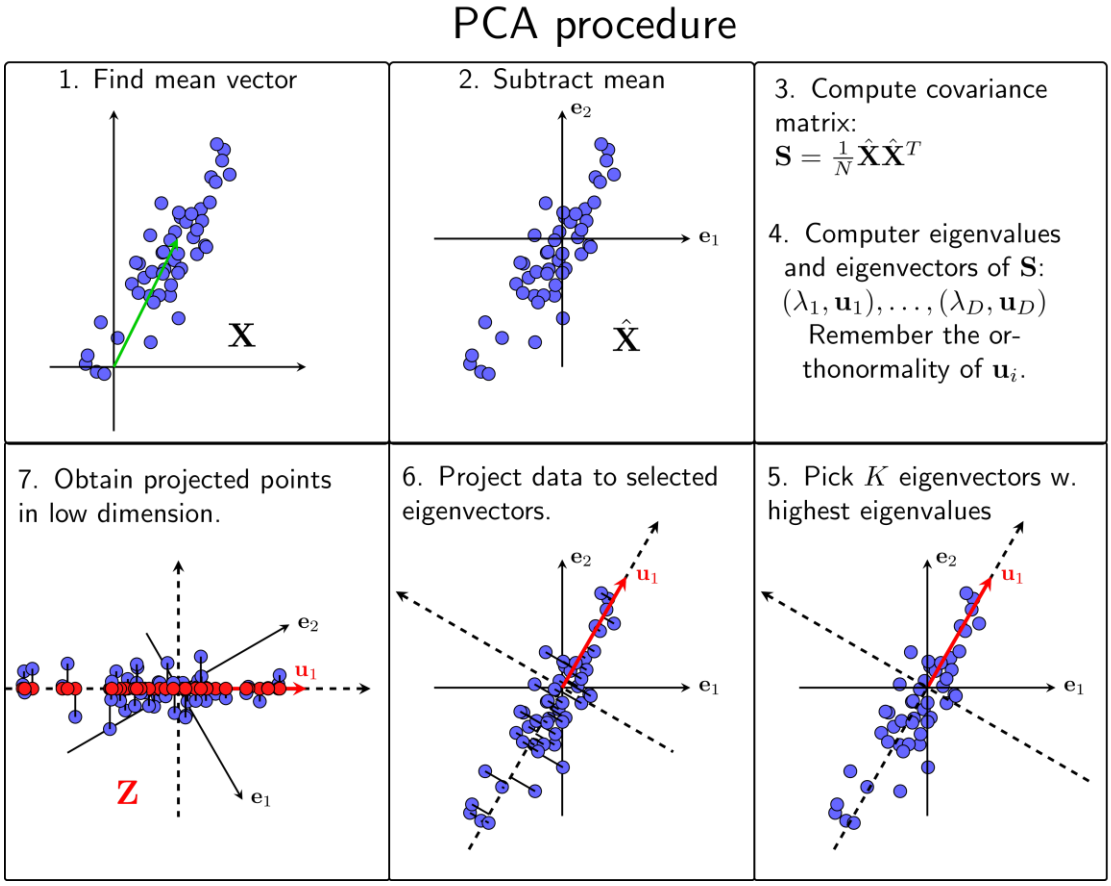
An illustration of l-Diversity for sensitive attribute values - from An Investigation of Data Privacy and Utility Using Machine Learning as a Gauge (Dissertation) - Mivule, Kato - 2014

PCA - Principal Component Analysis

PCA - Principal Component Analysis was invented by Karl Pearson in 1901 with the original purpose is to conduct the dimensional reduction. A formal definition of PCA, is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. This technique helps capture the importance features in the dataset by emphasizing the variation and capture strong patterns in a dataset.

Utilizing PCA as a technique to protect privacy is a reasonable idea in the case of large-multi-columns dataset. It can help hiding the sensitive in the most of columns by converting it into smaller dimensions, and it can't be reversed because we don't know the each column of the result is constructed by which columns from the input. However, the data utility here needs clarity: the result after doing PCA is good for Machine Learning modelling because it helps to

capture important information (in modelling view) and can produce a good quality of model; however, the PCA result can't be understood when doing EDA or analytics, and can't take any insights from here.



PCA Procedure - from Machine Learning Co Ban blog - https://machinelearningcoban.com/2017/06/15/pca/

Statistics aggregation

Most of the time, data will be sharing in the aggregation format (e.g.: average height of class A is 10.5, average salary of Data Science job in company X is $10000, average cars at 5PM on street A in July is 100,…) and the publisher thinks it can't be reverse to find individuals. This technique usually implements in the Data Query system, when the analyst want to get some aggregated data and the system calculate the aggregation on-demand.

However, this technique will fail to protect privacy when the aggregation group with just a few to very few individual, cause it may return the exactly the sensitive value and no-privacy-at-all. And in the world of Big Data and Real-time data querying, this technique will be attacked by the Differencing attacks - which will be mentioned in the next section.

Techniques to re-identify anyone in database

Linking attacks

The Linking attacks (or sometimes it's called Linkage attacks) involves the combination of auxiliary data with de-identified data to re-identified individuals. This is a common attack, and there are some notable efforts using this:

- In 1997, the effort of identifying medical records of William Weld - Massachusetts governor of Latanya Sweeney by combining de-identified public data of Massachusetts Medical Records with the Massachusetts Voter lists. (Sweeney, 2015)
- In 2006, the effort of identifying individual movies reviews from Netflix public dataset of (Narayanan & Shmatikov, 2006) (two PhD students of University of Texas at Austin) by combining with public IMDB reviews and comments. [@]
- In 2008, the effort of identifying search results of individual from AOL public dataset of Michael Barbaro and Tom Zeller (two reporters of The New York Times) by combining phone-book directory and the search queries itself. (Orlowski, 2006)

Differencing attacks

The differencing attacks happens when the attacker try to isolate individual from the statistics aggregation mask. An example for this is using 2 queries, one is finding the sum age of total employee, and one is finding a sum age of total

employee except the target one (by using filtering on some of other data features), then calculate the difference and the exact result show up.

Differential Privacy research

From the summary about Differential Privacy of Harvard Differential Privacy research group ("Differential Privacy," n.d.), the *"Differential privacy is a rigorous mathematical definition of privacy. In the simplest setting, consider an algorithm that analyzes a dataset and computes statistics about it (such as the data's mean, variance, median, mode, etc.). Such an algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not. In other words, the guarantee of a differentially private algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset – anything the algorithm might output on a database containing some individual's information is almost as likely to have come from a database without that individual's information. Most notably, this guarantee holds for any individual and any dataset. Therefore, regardless of how eccentric any single individual's details are, and regardless of the details of anyone else in the database, the guarantee of differential privacy still holds. This gives a formal guarantee that individual-level information about participants in the database is not leaked."*

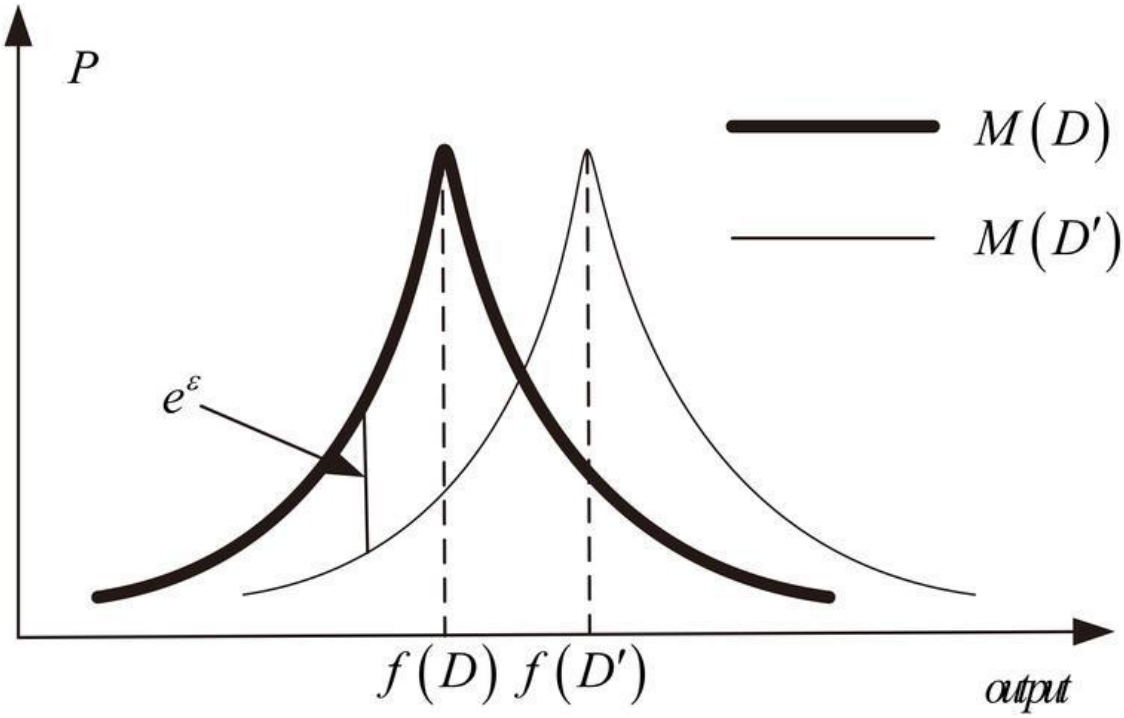
To implement Differential Privacy in the dataset, there are some techniques will explain here, but the main idea is all about adding noise into the data to help it be different.

Differential Privacy with adding noise mechanisms (Laplace and Gaussian random noise)

Laplace mechanism is a mechanism that adding noise from Laplace distribution randomly - with the config of sensitivity and ϵ-differential privacy level.

The general formula of this technique: $F(x) = f(x) + Lap(s/\epsilon)$

The key-point of this technique is to quantify the ϵ-differential privacy level. The larger ϵ, the less privacy, and smaller ϵ (recommendation is less than 1), the better privacy, but less data utility.

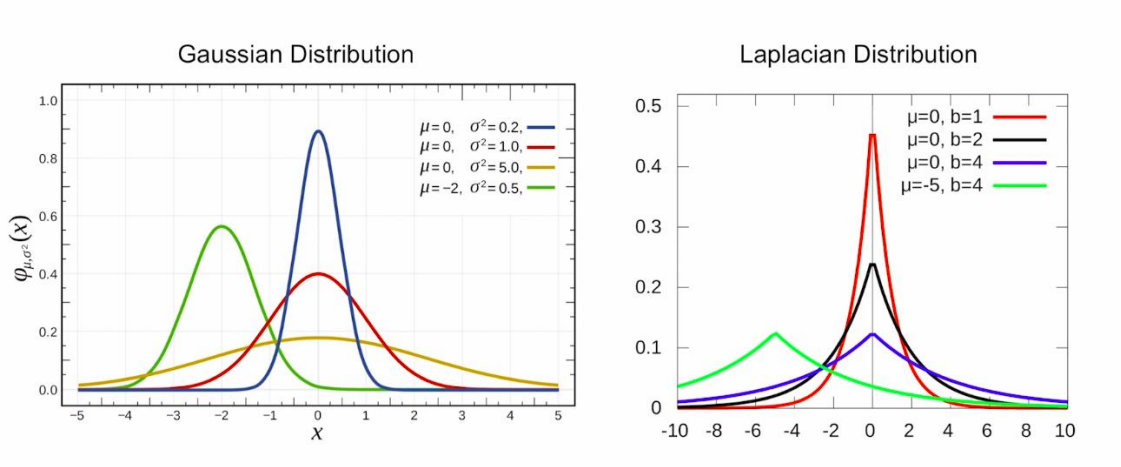


Laplace mechanism demonstration

Explanation: The $M(D)$ denoted as the original distribution of the dataset; the $M(D')$ denoted as the distribution of dataset applying Laplace mechanism; and the gap of data distribution between $M(D)$ and $M(D')$ denoted as e^{ϵ}

Similar to Laplace mechanism, the **Gaussian mechanism** also a noise adding solution to protect privacy - but it create random noise based on Gaussian distribution. A highlight point for this technique, it doesn't work with $\epsilon > 1$.

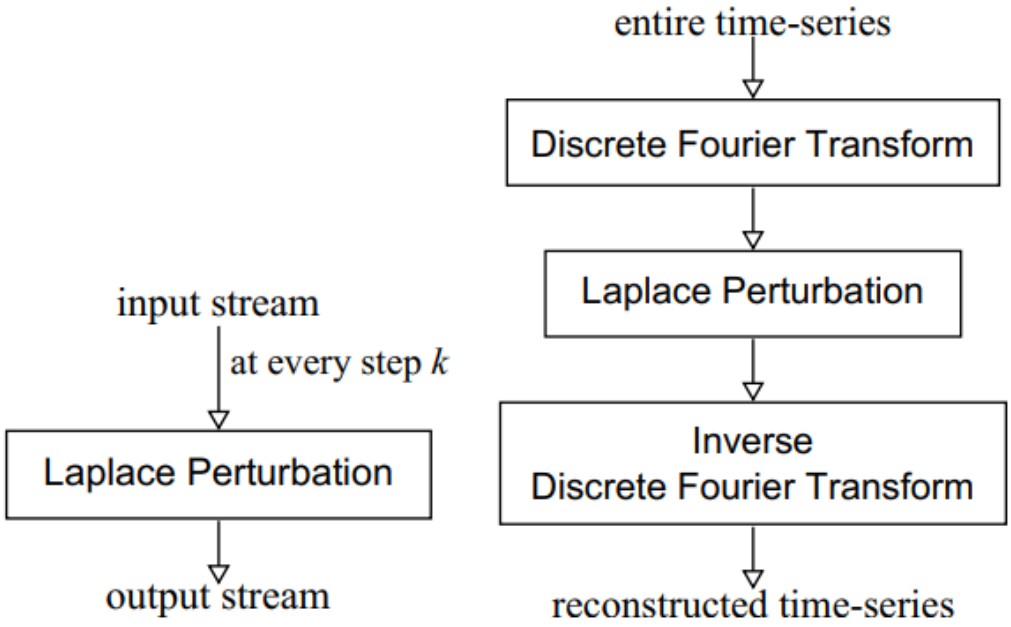
For more detail comparing between Laplace and Gaussian, it will be researched in the full thesis.



Laplace and Gaussian distribution on Differential Privacy

Differential Privacy on Time-series data with Discrete Fourier Transform, and Kalman Filtering

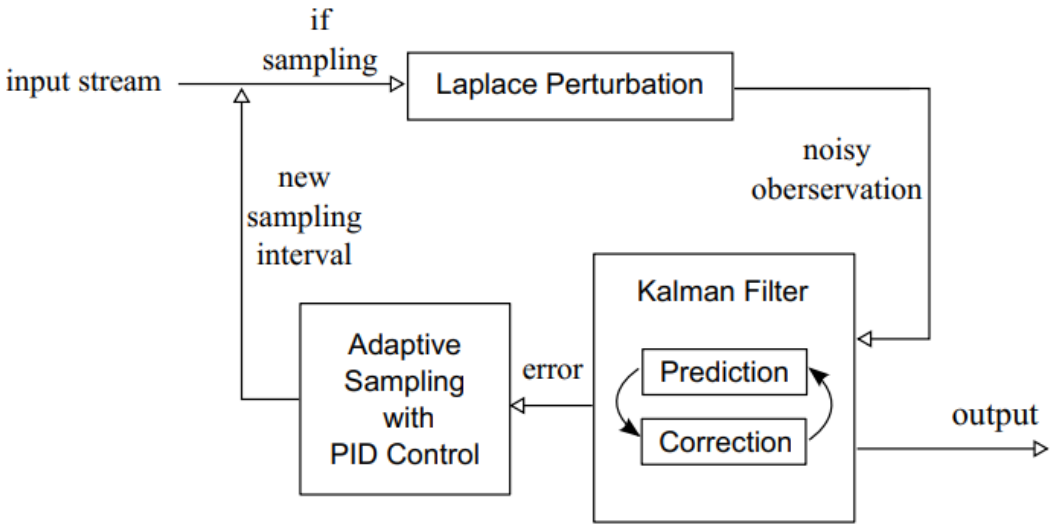
From the recent works of (Rastogi & Nath, 2010) in 2010, they recognized that the time-series data has the characteristics of correlated between data-points, the next-point has been influenced by some previous-points; therefore, just using noise adding mechanisms likes Laplace or Gaussian is not enough to preserve time-series features. To work-around this issue, they proposed to use Discrete Fourier Transform (a technique to convert a time-series into frequency domain, to find a spectrum) and adding noise on top of the result, then using Inverse Discrete Fourier Transform to convert it back to the output time-series.



(a) Workflow of *LPA* Algorithm (b) Workflow of *DFT* Algorithm

Laplace Permutation

After several years, another approach was proposed by (Fan & Xiong, 2012) to improve the data utility of time-series after applying Differential Privacy technique. The main idea is building a controller to manage the perturbation process using Kalman filtering to predict/correct the data-point after adding noise with Laplace mechanism to protect privacy. This technique is proven to be more useful for Time-series data than the DFT, and it will be implemented to compare in thesis.



Adaptive Sharing Time-series with Differential Privacy solution from Liyue Fan and Li Xiong

Time-series Analysis research

To verify the data utility of Time-series after running through the privacy solution, there are some techniques to check the time-series features:

1. Time-series decomposition into Trends and Seasonality
2. Time-series features analysis with Moving Average and Auto-Correlation
3. Time-series forecasting: ARIMA - Supervised Machine Learning - Deep Learning model with LSTM

The purpose of (1) and (2), it will help to verify the ability to conduct Exploratory Descriptive Analysis (EDA) and give reasonable insight from the dataset. And the purpose of (3) is to verify the ability to ingest into Forecasting algorithms - which is a common use-case in Time-series application.

All of these will be mentioned in more detail inside the thesis.

References

Aitsam, M. (2022). *Differential privacy made easy*. arXiv. https://doi.org/10.48550/ARXIV.2201.00099

Arcolezi, H. H., Couchot, J.-F., Renaud, D., Bouna, B. A., & Xiao, X. (2022). *Differentially private multivariate time series forecasting of aggregated human mobility with deep learning: Input or gradient perturbation?* arXiv. https://doi.org/10.48550/ARXIV.2205.00436

Art. 4 gdpr – definitions. (2018). Retrieved from https://gdpr-info.eu/art-4-gdpr/

Balle, B., & Wang, Y.-X. (2018). *Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising*. arXiv. https://doi.org/10.48550/ARXIV.1805.06530

Brickell, J., & Shmatikov, V. (2008). The cost of privacy: Destruction of data-mining utility in anonymized data publishing. *KDD*.

Confessore, N. (2018). Cambridge analytica and facebook: The scandal and the fallout so far. The New York Times. Retrieved from https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Cunningham, T., Cormode, G., Ferhatosmanoglu, H., & Srivastava, D. (2021). Real-world trajectory sharing with local differential privacy. *Proceedings of the VLDB Endowment*, *14*(11), 2283–2295. https://doi.org/10.14778/3476249.3476280

Differential privacy. (n.d.). Retrieved from https://privacytools.seas.harvard.edu/differential-privacy

Fan, L., & Xiong, L. (2012). Adaptively sharing time-series with differential privacy. *ArXiv, abs/1202.3461*.

Fan, L., Xiong, L., & Sunderam, V. (2013). Differentially Private Multi-dimensional Time Series Release for Traffic Monitoring. In L. Wang & B. Shafiq (Eds.), *27th Data and Applications Security and Privacy (DBSec)* (pp. 33–48). Newark, NJ, United States: Springer. https://doi.org/10.1007/978-3-642-39256-6_3

ISO 25237:2017. (2022). Retrieved from https://www.iso.org/standard/63553.html

Le Ny, J. (2020). Differentially private kalman filtering. In *Differential privacy for dynamic data* (pp. 55–75). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-41039-1_5

Liu, Y., Chen, C., Zheng, L., Wang, L., Zhou, J., Liu, G., & Yang, S. (2020). *Privacy preserving pca for multiparty modeling*. arXiv. https://doi.org/10.48550/ARXIV.2002.02091

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006a). L-diversity: Privacy beyond k-anonymity. *22nd International Conference on Data Engineering (Icde'06)*, 24–24. https://doi.org/10.1109/ICDE.2006.1

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006b). L-diversity: Privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*. https://doi.org/10.1109/icde.2006.1

McCarthy, J. (2021). Worries about personal data top facebook users' concerns. Gallup. Retrieved from https://news.gallup.com/poll/232343/worries-personal-data-top-facebook-users-concerns.aspx

Mercier, D., Lucieri, A., Munir, M., Dengel, A., & Sheraz, A. (2021). Evaluating privacy-preserving machine learning in critical infrastructures: A

case study on time-series classification. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/tii.2021.3124476

Narayanan, A., & Shmatikov, V. (2006). *How to break anonymity of the netflix prize dataset*. arXiv. https://doi.org/10.48550/ARXIV.CS/0610105

Ny, J. L., & Pappas, G. J. (2012). *Differentially private kalman filtering*. arXiv. https://doi.org/10.48550/ARXIV.1207.4592

Orlowski, A. (2006). AOL publishes database of users' intentions. The Register. Retrieved from https://www.theregister.com/2006/08/07/aol_search_logs/

Rastogi, V., & Nath, S. (2010, June). *Differentially private aggregation of distributed time-series with transformation and encryption*. 735–746. https://doi.org/10.1145/1807167.1807247

Sweeney, L. (2000). *Simple Demographics Often Identify People Uniquely*. https://doi.org/10.1184/R1/6625769.v1

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *IEEE Security and Privacy*, *10*, 1–14.

Sweeney, L. (2015). Only you, your doctor, and many others may know. Retrieved from https://techscience.org/a/2015092903/

Xiong, W., Xu, Z., & Wang, H. (2017). An attack model on differential privacy preserving methods for correlated time series. *International Journal of Database Theory and Application*, *10*. https://doi.org/10.14257/ijdta.2017.10.1.09