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

Master Thesis Proposal Defense – 28 Jun 2022

Student: Lai Trung Minh Duc – 2070686

Instructor: Assoc. Prof. DANG TRAN KHANH

Greetings the Board of Judges

- Assoc. Prof. NGUYEN THANH BINH
- Assoc. Prof. TRAN MINH QUANG
- Dr. PHAN TRONG NHAN

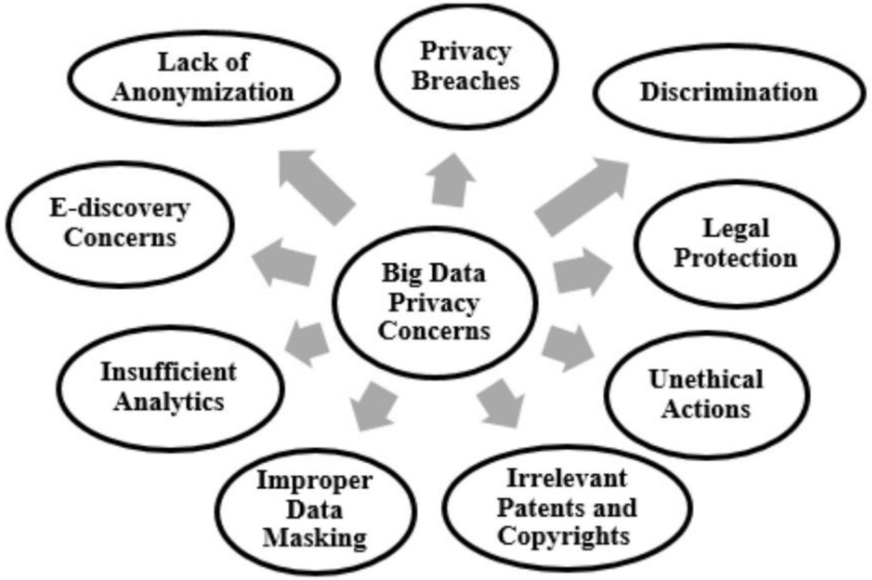
Contents

Urgency of this study

Social – Academic – Corporate

Context - Social

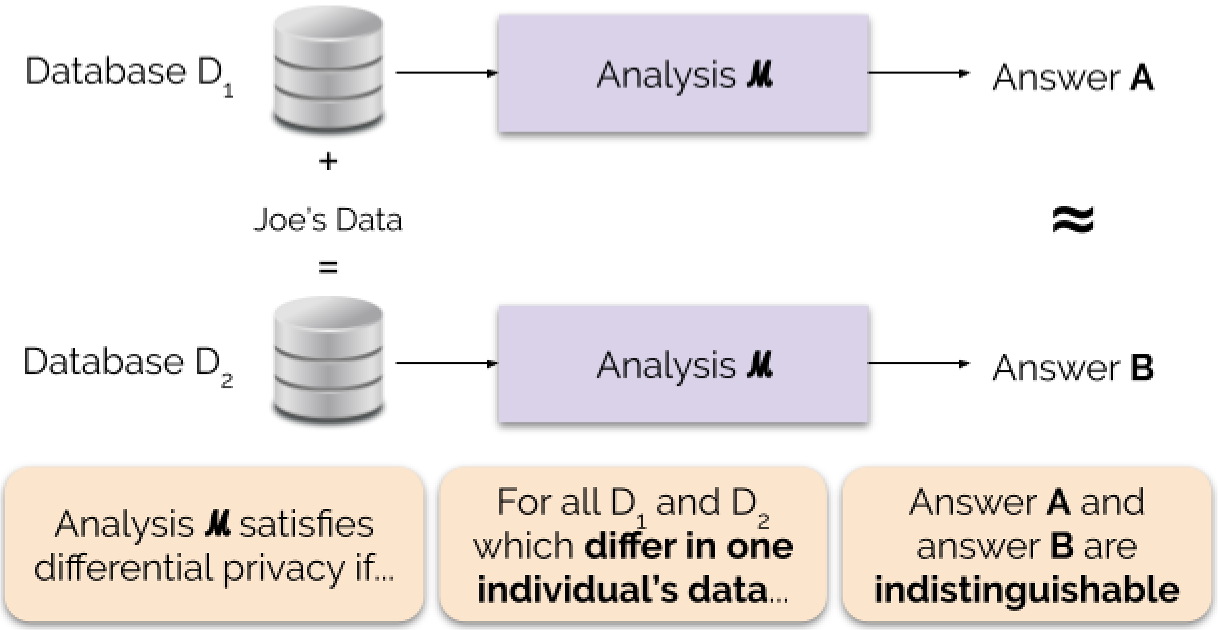
- We are living in the world of Big Data.
- Privacy-concern is getting more attention after Facebook scandal and GDPR effective (2018).
- People don’t want to be analyzed deeply, also don’t want to be known publicly.
- People are severely getting attention to search-privacy & location after Roe v. Wade overturned 26 Jun 2022 *(US Supreme Court ends constitutional rights to abortion)*



Source: Brohi, Sarfraz & Bamiah, Mervat & Brohi, M Nawaz. (2016). Identifying and Analyzing the Transient and Permanent Barriers for Big Data. Journal of Engineering Science and Technology. 11. 1793 - 1807.

Context – Academic

- Differential Privacy (DP) is emerging research area for dealing with privacy in Big Data & Data Analytics.
- DP is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals.
- DP preserves the data distribution & statistics descriptive features.



Source: Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series – NIST Blog - https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our

Context – Corporate

- Data partnership is forming between multi-parties (e.g.: Supermarkets with FMCGs, E-commerce platform with FMCGs,…)
- Anonymization transactions data is sharing privately/publicly → leads to potential data reidentification while doing data analytics – especially for the problems of customer segmentation/customer life time value.

Urgency of this study

- Traditional DP preserves the general data distribution – but not preserves the seasonality & other-TS features → Work badly on Time-series data.
- Ultimate purpose:
  - Protect individual privacy
  - Preserves the analytical features: auto-correlation, seasonality,…
  - Low error data
  - Apply for multiple data-type: continuous, intermittent, long-history, short-term,…

➔ Need a deeper research on the DP method for time-series.

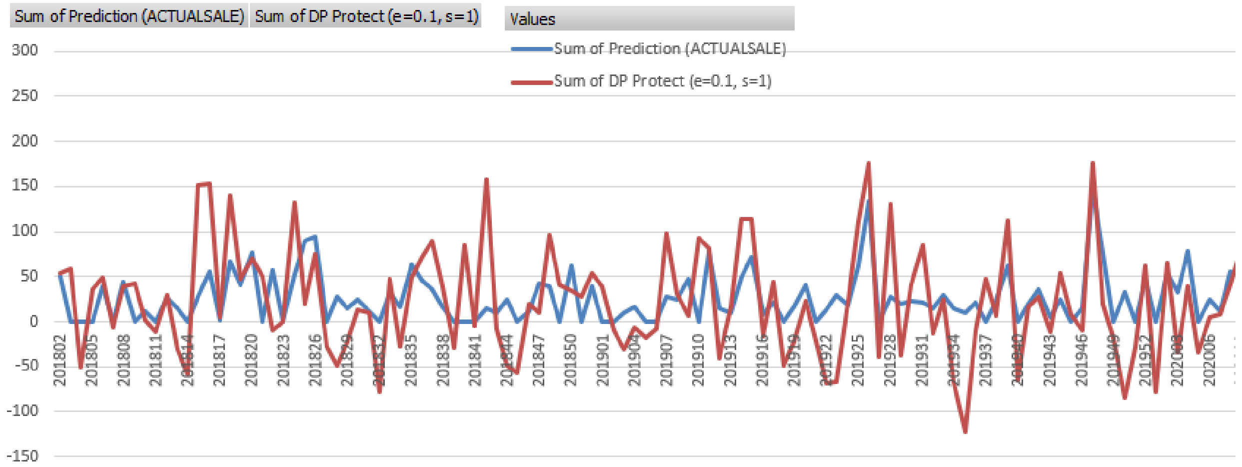


Figure: Personal experiment on corporate retail data – Test Laplace DP technique → Unusable TS

Literature Review

Differential Privacy

- Differential privacy is a mathematical definition of what it means to have privacy → A extension in a process which can help to provide privacy (adopt from NIST)
- The output of a differentially private analysis will be roughly the same, whether or not you contribute your data. A differentially private analysis is often called a mechanism, and we denote it $\mathcal{M}$. (adopt from NIST)

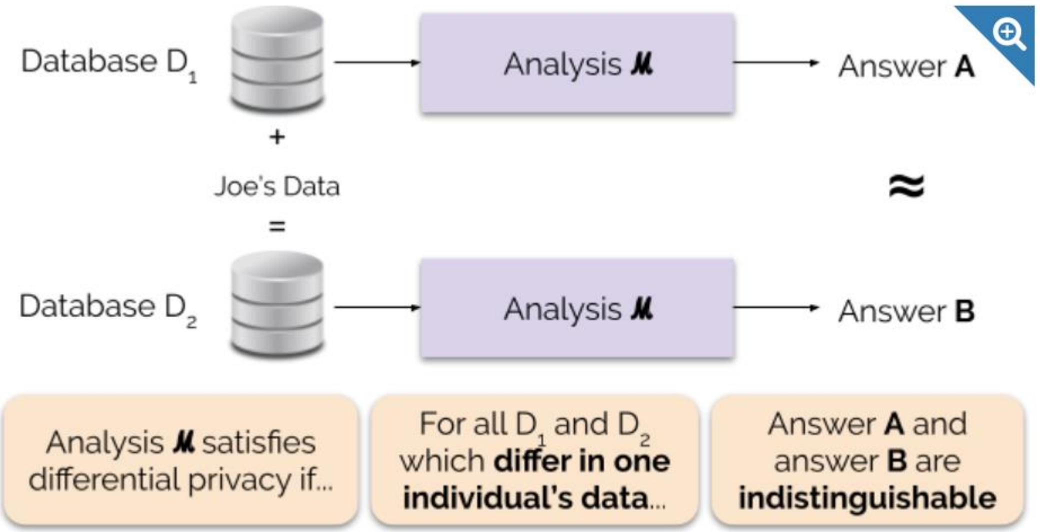


Figure: Informal DP definition - NIST

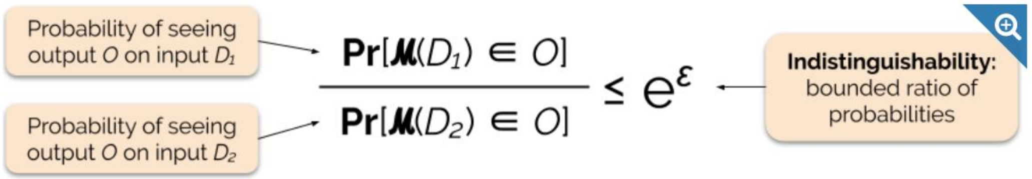
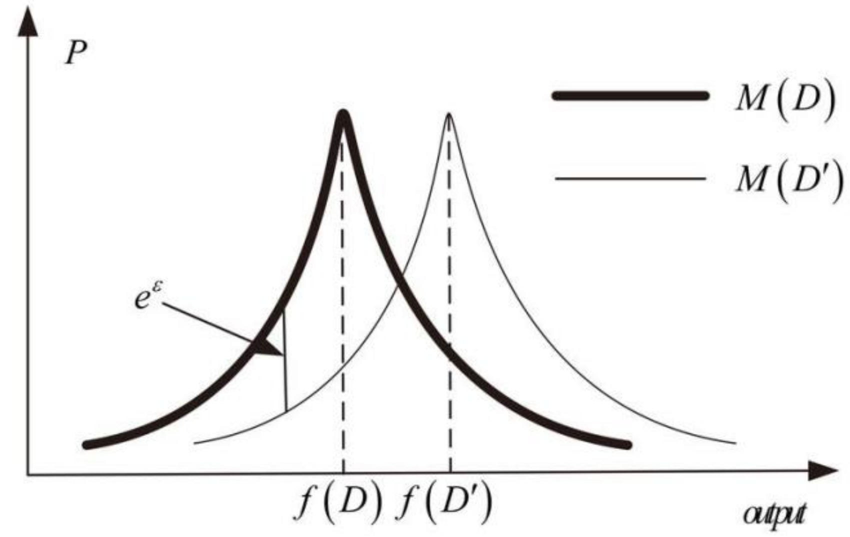


Figure: Formal DP definition - NIST

Some Differential Privacy techniques

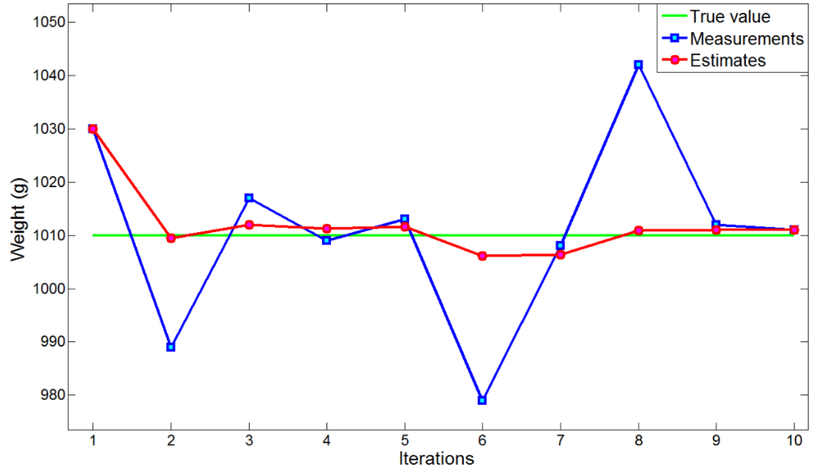
Laplace technique

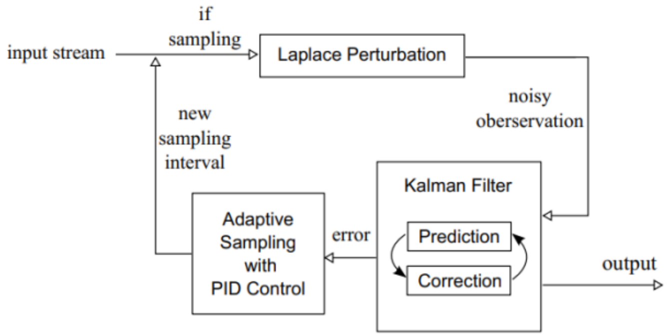


PDF	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$

$$F(x) = f(x) + Lap(\tfrac{s}{\varepsilon}) \quad (1)$$

Kalman Filtering technique





About the thesis

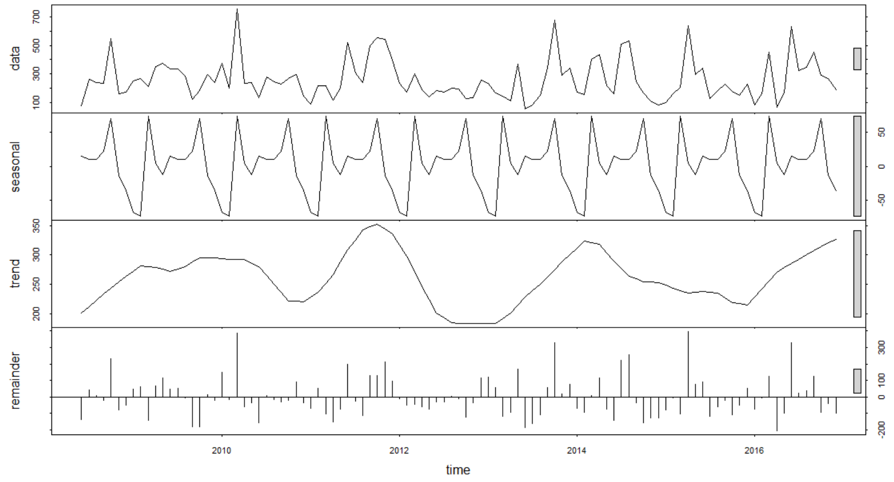
Data information

- Description: Daily customer transaction sales data – after anonymization (Retail TS data).
- Data source: Corporate
- Data structure:
  - CustomerID (anonymization ID)
  - StoreID (anonymization ID)
  - Region
  - Date
  - ProductID
  - ProductCategory
  - Quantities
  - Value

- Potential attack for this kind of dataset:
  - Basket-size attack (attacker knows the usual basket size of target).
  - Basket-item attack (attacker has 1 receipt of target).
  - Location & Date & Product behavior.
  - Correlated attack (inference data-points)
  - Linkage attacks from public database (Facebook/Zalo/Twitter photos/status crawler → define behavior + check-in)
  - …

Data features analytics
+ Data error metrics settings

- Data features analytics check
  - Trends and Seasonality check
  - Data distribution check
  - Intermittent check
  - Stationary check
  - Machine Learning model generator



- Data error metrics
  - Mean Relative Error (MRE)

$$MRE = \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|}{\sum_{j=1}^{n} x_j}$$

→Dimensionless and relative to true values

Check on Trends/Seasonality/Actual-value/Data-distribution/Intermittent.

  - Target data accuracy: >= 60% (Baseline model accuracy is 44%)

- And make sure the proposed solution will satisfied the Differential Privacy definition.

Research outcome expectation

- Compare proposed algorithms in literature review (Traditional Laplace/Gaussian, Kalman Filtering, DFT-IDFT)
- Analyze data features and build algo-classification (which data-type go with which algorithms) by data features analytics check.
- Propose a decision tree on choosing DP algorithms.
- Package the solution and make it run (Python)

Thesis structure

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: *Anoynization, 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.

Expected timeline

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**

Q/A

Thank you for your listening.

Reference (in this slide)

1. 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
2. Ny, J. L., & Pappas, G. J. (2012). Differentially private kalman filtering. arXiv. https://doi.org/10.48550/ARXIV.1207.4592
3. 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
4. Fan, L., Xiong, L., & Sunderam, V. (2013). Differentially Private Multidimensional 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
5. 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
6. Brickell, J., & Shmatikov, V. (2008). The cost of privacy: Destruction of datamining utility in anonymized data publishing. KDD
7. 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