Data Science Individual Project Proposal
Differentially Private Location-based Histogram Publication

Name: CHAN, Lo Yuet

Introduction

With prevalence of applications collecting data about life patterns of users, it is made possible for analysts to look into both open datasets and private datasets for intelligence. However, it is of prime importance to maintain individual privacy while distributing any of the data and findings. In this project, the problem of releasing one-dimension dataset in its histogram representations under differential privacy policies will be addressed.

With sufficient privacy enforced, analyst can then publish their findings without risking any sensitive data of both firms and users. Such work will contribute not only to daily life of everyone, but also to development strategies of firms. Taking restaurant reviews and check-in data as an example, analyst can then segment users into different preferences and provide them with suggestions of new restaurants when they are travelling to a new place; analysts can also look into the positioning of their firm and adopt strategies to either consolidate their reputation or mitigating to the desired market segments.

With the possible benefits said, this project covers over a broad domain. By modifying differential private algorithm, our work would be applicable for applications which uses histograms as query answers. Our work would hopefully provide insights on clustering and classification as well. Among all the datasets, our work will be focusing on one-dimension datasets and specifically on datasets containing check-in records.

Dataset Description

In this project, three anonymised open datasets from Foursquares will be used. The raw dataset contains tips records, reviews on restaurants, and check-in records of users at different restaurants. There would be a total of 34,091,913 check-in records. To tailor-made the datasets for this projects, excess data are dropped leaving behind datasets with only user ID, venue ID. Details of the cured dataset can be found in Table 1.

dataset_TIST2015_Checkins.csv (33,263,633 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_TSMC2014_NYC.csv (227,428 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID
dataset_TSMC2014_TKY.csv (573,703 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID

dataset_ubicomp2013_checkins.csv (27,149 records)			
Column	Variable	Type	Details
0	User ID	Integer	Anonymous ID
1	Venue ID	String	Anonymous ID

Table 1

Literature Review

There has been a number of new knowledge in the field of differential privacy recently. To be specific, this project is inspired by the work of Zhang et. al. [1] on AHP and its edge over other clustering algorithms. To obtain a broad picture of the latest development, a series of research papers, which cover different histogram settings, dimensionality, and nature of data are reviewed. Apart from that, work focusing on calibrating and benchmarking is also studied.

Clustering Algorithms in Histogram

The principle behind Zhang et. al. [1] is to minimise Laplace error applied onto the dataset with help of AHP. To come up with the optimal solution, three candidate algorithms are put into comparisons.

The AHP suggested is still, however, questionable. Although the algorithm shows a clear edge over others like *PHP* [11], *GS* [12], *NoiseFirst* [13], *StructureFirst* [13], these comparisons are all made with experimental datasets of scale of 10^3 and 10^4. Whether the algorithm excels in datasets of other scales is still questionable.

Free parameters are also not fully discussed in the research. The work focuses on comparisons of algorithm over different privacy budget. However, the fact that privacy budget can be freely allocated at different stage is not fully exploited, but is preset and hard-coded instead makes the work extendable in the future.

As the work focuses on Laplace noise only, approximation error is almost not addressed in the paper. This leaves a question of whether it is better to have reduced Laplace noise or to have reduced approximation error. Another problem would be that is it better to have a highly reduced Laplace noise or slight reduction in both Laplace noise and approximation error. In other words, a better utility scheme would be needed to measure the improvement.

Node Differential Privacy

In this domain, the general concept towards differential privacy is base on graph and networks. Once differential privacy is enforced, nodes and all its edges should not be distinguishable from other nodes and edges [4].

Day et. al. [4] focuses on graph data representation. In this domain, they have suggested two solutions, that has an edge over the current state-of-art in lower-degree nodes. These are all made possible with their novel graph projection mechanism.

Although the work is outside the domain of this project, Day et. al. had reminded that for a new algorithm to be successfully developed, sensitivity of each attribute should be carefully reviewed such that altering one attribute will not lead to significant error.

Apart from the reminder on sensitivity, the way Day et. al. compared free tuning parameters can also be applicable on this project given some improvement. They have selected a few values for testing their new algorithm before finalising the preset value. This, however, has restricted the flexibility nature of the free tuning parameter, and thus missed that opportunity to optimise and generalise the algorithm with help of a function that help tune these attributes.

High-dimension data

For datasets with higher dimensions, enforcing differential privacy will usually suffer from what is called 'curse of high dimensionality' resulting in either heavy computation power requirement, excessive error, or low privacy. The current state-of-art algorithm MWEM is reviewed to be a NP-hard computationally complex algorithm by Hardt et. al. [10].

Chen et. al. [5] proposed to solve the 'curse' with a top-down partitioning algorithm on a set-valued dataset. With the novel partition, they yielded a stronger privacy protection with the same amount of relative error. They also suggested that a non-interative data sanitisation can be achieved situationally if underlying data is carefully used. This which would allow users to directly interact with the sanitised dataset without needing a sanitising mechanism as a middleman to modify queries from users.

As their proof for non-interactive data sanitisation is fairly situational, there would still be a big gap of generalising such a sanitisation scheme. On the other hand, their utility is only based on overall amount of relative error. This would have allowed results with low relative error, while having a highly varied error slip through. Further work would be needed to improve the scheme such that error can be kept small and concentrated, and thus, for preserving accuracy of any queries.

In another research, Chen et. al. [3] focuses on solving the problem with a sampling-based inference which allows thicker spread of privacy budget. The solution depends on the conditional independence and the one dimensional association across attributes. With a higher privacy budget at each step, the accumulated noise in the sanitised dataset would be highly reduced.

As the solution depends on the important assumption of conditional independence, its effectiveness would be highly situational, and should be reconsidered before using it as a general approach for enforcing differential privacy.

Despite the presumption, this piece of work brings a new insight towards efficient spending of privacy budget. Sampling-based inference should be further studied so that it may be applied onto other algorithms to improve the overall accuracy.

Dimitrakakis et. al. [7] looked into sampling inference with posterior sampling in Bayesian network. With little previous differential privacy research in the field of Bayesian inference,

this piece provided a building block and a proof for using posterior sampling for enforcing differential privacy.

This piece of work is mostly out of the domain of our project as it mainly focuses on theoretical use of Bayesian inference on high dimension datasets. It however shows that the field of differential privacy is evolving so quick that new scheme with different statistics skills are being looked into.

Gaboardi et. al. [8] proposed an algorithm, named DualQuery, with worst-case complexity in exponential relation against dataset dimension. Despite the exponential increase, they have shown that their algorithm is computationally more concise than the state-of-art by applying the algorithm on a synthetic dataset with more than 500,000 attributes.

Their algorithm, however, only shows edge over in high-dimension queries. In other words, MWEM still remains to be the state-of-art in lower-dimensions queries, while Gaboardi et. al. did not provide the boundary for switching from MWEM to DualQuery for saving computational power.

Time-series data in Histogram

In the domain of time-series data, Chen et. al. [2] considered it as an infinite stream of data. Their work focuses on generating differentially private histograms continuously. A sampling based algorithm with a retroactive grouping mechanism is adopted to enhance computational, and spatial efficiency, and thus incurring a smaller delay when publishing histograms.

As timestamps are not the major focus of this project, the work from Chen et. al. is mainly out of scope. However, the use of a Bernoulli sampling scheme may be applicable on this project for getting a more accurate differentially private result while suffering from reduced Laplace noise.

Differential Privacy without External Noise

While most of the works are focusing on enforcing differential privacy with help of introducing noises into the dataset, Duan [9] worked on answering sum queries without any external noises. Duan pointed out that state-of-art differential privacy techniques is flawed for aggregate queries due to the zero-mean symmetric nature of Laplace distribution. He then argued that if the dataset is sufficiently large, aggregate queries would be private enough itself given that sum of multivariate Gaussian distributions converges when n is large as stated in central limit theorem.

Although aggregate queries are not the main focus of this project, Duan's work has opened a big gap in the field of differential privacy as a possible flaw has been discovered. The work also provides an insight of whether the original data is private enough once the dataset is large enough. If the dataset is, indeed, private enough, publishing an unsanitised anonymised would ensure the most accurate result for various data mining tasks.

Benchmark and Metrics

While all the above works are focusing on varies algorithms optimising in different scenarios, there are actually limited work on benchmarking algorithms across the table. Hay et. al. [6] suggested DPBench by considering performance of different data-dependent and data-independent algorithms under different scales and sizes.

The work from Hay et. al. adopted a very objective comparison scheme between different algorithms. It also addressed that as the algorithm will provide random noises, it would be biased if comparisons are drawn based on the shape of noise graph across different scale and attributes among different different algorithms. Such standard should also be upheld in this research should a similar comparison be adopted.

In order to compare between different algorithms, Hay et. al. actually exposed a gap in the optimisation of each algorithm under different scenarios. Such gap would lead to a biased decision over which algorithm should be adopted for an input dataset of interest.

Research Question

Allocation of Privacy Budget

When a limited privacy budget over a series of senitising procedures, great amount of noise would be introduced to the input dataset at every step. This, however, would lower the accuracy for further analysis, and, thus, is undesirable. There has been recent works focusing on spreading privacy budget in high-dimension dataset [3], while gaps still exist for lower ones.

One of the possible direction for this project would be to look into approaches for better allocation of privacy budget in one-dimension datasets such that privacy can be enforced with preserving accuracy. A candidate solution would be sampling based approach [3].

Threshold Function

Under the current AHP [1], random noises would be added to every group or cluster which lead to inevitable artificial non-zero counts including both positive ones and negative ones. Therefore, a threshold function would be needed to trim off some of the values that are probably invalid to preserve accuracy. The current status of the function in AHP is just to trim off any value lower than a coded threshold and set them to zero. This may not be the optimal solution as there may be clusters with exceptionally high counts which would also need to be sanitised.

This project can look into approaches than consider both absolute values and variation of value across clusters such that an objective threshold function can be adopted in the future. Some of the possible approaches includes trimming off outliers, expressing the threshold function in terms of variance of the cluster of interest, preserving only the most significant records after applying noises, etc.

Scaling Function

As addressed by Hay et. al. [6], data-dependent algorithm like AHP and AHP* would be easily affected by the scale of dataset. However, there were limited work on optimising the algorithm as scale varies. This project may fill the gap in previous works by smoothing the AHP algorithm across different scale such that data-dependent algorithm would be competitive enough against data-independent algorithms as scale increases. This, however, would be expected to be non-trivial if the number of attributes of interest increases.

Utility and Risk Appetite

It is understandable that users would have different risk appetite and, thus, would like to have various extent of protection of sensitive data and privacy. However, there was a gap in the previous work concerning quantification of risk appetite. A possible issue to be addressed would be quantifying risk appetite into utility and develop a risk profile for users to follow when choosing different algorithms.

With a good utility function, it can not only help users choosing the optimal algorithm settings, but also help evaluating the progress and findings in this project. Therefore, this should be a very important area to look into in this project.

Proposed Approach

Allocation of Privacy Budget

As the dataset is only of one-dimension, reallocation of privacy budget would be very limited at this stage. The room for development would be greatly expanded as the algorithm becomes more complex throughout the project. Therefore, this topic should be studied together with the rest of the possible research questions instead of investigating into it alone.

Threshold Function

As mentioned in previous section, a possible approach would be to look into trimming off outliers, expressing the threshold function in terms of variance of the cluster of interest, preserving only the most significant records after applying noises, etc.

In order to gain more insight on developing a more complete threshold function, spare vector technique [3] and threshold query techniques [3] should also be further studied in the future.

The evaluate the impact of the novel threshold function, mean error and its variation under different scale and across different experiment dataset need to be recorded. Assuming differential privacy can be enforced alongside this function, comparing the expected error and variation with those of the state of art threshold function would give an objective conclusion over the edge.

Scaling Function

In order to make the scaling function possible, we would need to first look into the amount of parameters of interest. At this stage, the possible attributes to be included in the function would be ϵ_1, ϵ_2, and η (the tuning parameter).

Base on the initial choice of parameters, different combination of their values should be tried on different scale to obtain an initial scaling function. However, if more parameters are included, for instance, adding a ϵ_3 as a result of introducing sampling inference, the function would have to be re-optimised.

To enable the investigation, datasets of different scales would be needed. Given the current dataset, synthetic datasets would be created to support the study. However, further studies in related field would be needed before doing so to ensure the synthetic dataset is not biased. In case if such process cannot be completed, additional dataset is always obtainable from tech giants like Facebook or Twitter.

While for evaluation, we would need to look into the noise incurred and its variation for both the novel function and the current state of art AHP function under preset values. In real life scenario, users should not be trying different combination of parameter values for the best outcome as that would violate the principle of differential privacy. Therefore, the current state-of-art would be presetting values for parameters. For objective evaluation, further studies would be needed to obtain the usual preset values.

Utility and Risk Appetite

Based on the work by Hay et. al. [6], a utility framework is developed with IDENTITY as the baseline for comparison. To expand their benchmark utility framework, we can look into quantifying the benchmark utility by considering both, but not restricted to, privacy level, accuracy, variation of error, and also computational complexity.

After the utility function is formulated, it can be calibrated against a benchmark algorithm, say IDENTITY algorithm, to stay consistent with previous work by Hay et. al. [6]. By putting the utility of the modified algorithm in the project into comparison, we would be able to evaluate any improvements.

It would be tricky, however, to evaluate a utility function as it would be a subjective weighting between factors. There can be two ways in evaluating the function, including the shape of the function, and whether monotonic points are reasonable. As a result, graphical representation of utility functions and monotonicity lines would be needed.

If developing a single utility function that aligns risk appetite, and complexity is not possible, an alternative direction would be to provide two scores, one for risk appetite while the other for complexity, in a form of a two-value vector. This would however need users to subjectively balance between the two scores.

Timeline

Proposed Deadline	Tasks
5/6/2017	Completion of gathering required readings on: - Differential privacy - Clustering - Synthetic dataset
12/6/2017	Completion of reviewing required readings Start writing preliminaries section of the final thesis
3/7/2017	Completion of all necessary coding tasks - Threshold function - Scaling function In case of good progress, utility function should also be looked into
17/7/2017	Complete test, review, and documentation of code
31/7/2017	Complete generation of all graphical output
20/8/2017	Completion of final report

Bibliography

1. Xiaojian Zhang, Rui Chen, Jianliang Xu, Xiaofeng Meng, Yingtao Xie. "Towards Accurate Histogram Publication under Differential Privacy." *2014 SIAM International Conference on Data Mining*, 2014: 9.
2. Rui Chen, Yilin Shen, Hongxia Jin. "Private Analysis of Infinite Data Streams via Retroactive Grouping." *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015: 1061-1070.
3. Rui Chen, Qian Xiao, Yu Zhang, Jianliang Xu. "Differentially Private High-Dimensional Data Publication via Sampling-Based Inference." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 129-138.
4. Weiyen Day, Ninghui Li, Min Lyu. "Publishing Graph Degree Distribution with Node Differential Privacy." *Proceedings of the 2016 International Conference on Management of Data*, 2016: 123-138.
5. Rui Chen, Benjamin C. M. Fung, Li Xiong. "Publishing Set-Valued Data via Differential Privacy." *Proceedings of the 37th International Conference on Very Large Data Bases*, 2011: 1087.
6. Michael Hay, Ashwin Machanavajjhala, Gerome Miklau, Yan Chen, Dan Zhang. "Principled Evaluation of Differentially Private Algorithms using DPBench." *Proceedings of the 2016 International Conference on Management of Data*, 2016: 139-154.
7. Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikaterini Mitrokotsa, Benjamin I. P. Rubinstein. "Differential Privacy for Bayesian Inference through Posterior Sampling." *Journal of Machine Learning Research* 18, no. 11 (2017): 1-39.
8. Marco Gaboardi, Emilio Jesu´s Gallego Arias, Justin Hsu, Aaron Roth, Zhiwei Steven Wu. "Dual Query: Practical Private Query Release for High Dimensional Data." *Proceedings of the 31st International Conference on Machine Learning*, 2014: 1170-1178.

9. Duan, Yitao. "Differential Privacy for Sum Queries without External Noise." *Proceedings of the ACM Conference on Information and Knowledge Management*, 2009: 1517-1520.
10. Moritz Hardt, Katrina Ligett, Frank Mcsherry. "A Simple and Practical Algorithm for Differentially Private Data Release." *Advances in Neural Information Processing Systems 25*, 2012.
11. G. Acs, C. Castelluccia, R. Chen. "Differentially private histogram publishing through lossy compression." *Proceedings of ICDM*, 2012: 1-10.
12. Papadopoulos, G. Kellaris and S. "Practical differential privacy via grouping and smoothing." *Proceedings of VLDB Endow* 6, no. 5 (2013): 301-312.
13. J. Xu, Z. Zhang, X. Xiao, and G. Yu. "Differentially private histogram publicaiton." *Proceedings of ICDE*, 2102: 32-43.