DIT IS DE TITEL VAN MIJN

by

Tjibbe van der Ede

in partial fulfillment of the requirements for the degree of

Master of Science
in Software Engineering

at the Open University, faculty of Management, Science and Technology
Master Software Engineering
to be defended publicly on Day Month DD, YYYY at HH:00 PM.

Student number:	student number	
Course code:	IMA0002	
Thesis committee:	titles and name of the chairman (chairman),	Open University
	titles and name of the supervisor (supervisor),	Open University

Contents

1

TITLE OF THE CHAPTER

This is just to show how to include a test file for a chapter, a reference **?**

2
Related work

Distributed clustering

Recent work for clustering is authored by Xia et al. Xia et al. [2020]. Their work focuses specifically on a distributed variant of the K-means cluster algorithm. They propose a method for applying LDP for distributed K-means. For this purpose, a modified LDP algorithm is proposed. To fit K-means clustering, the features are converted to binary strings rather than boolean values. RR is then used to perturb each feature and is combined into a feature vector. The privacy cost is dependent on the length of bits of each feature transformation. This means a longer length yields more information, at the cost of the privacy budget. In every iteration, the data is aggregated (server-side) and the K-means algorithm calculates and sends centroids to each user. In response, the user (client-side) re-calculates the distances until the centroid becomes stable. The disadvantage of the approach is the high correlation between user data ($V = v^0, v^1 ... v^n$) and the clusters ($U = u^0, u^1 ... u^n$). To solve this an enhancement to the algorithm is proposed. The client-side still sends v^i, but now also includes a random zero string set of $S = z, ... v^i ... x$. The perturbed version S' is sent to the server-side along with v^i. Now, the server-side conducts similar calculations to receive the real cluster. *research why this is not a suitable solution.*

A paper written by Huang et al. Huang et al. [2021] extends Xia et al.'s work. Like the previous paper, the algorithm is used for 2-dimensional distributed clustering. The authors aim at fixing a shortcoming of the solution provided by Xia et al. Namely, a too big difference between actual and perturbed values. Their solution is to use Condensed Local Differential Privacy (CLDP) for small-scale values and LDP for large-scale values. To reach CLDP a distance-aware LDP is used. The method for the client side to calculate the distance between two values is Square Wave (SW) mechanism, which is also used for large-scale data. Another method that also satisfies $\alpha - CLDP$ is using the Manhattan distance. Finally, the perturbed data is used with a classical K-Means algorithm on the server-side.

K-anonimity

TODO

Location based clustering

The next two papers apply LDP as well, but it is more related to location mapping. However, the way they take care of the problem corresponds to clustering. The paper written by

Andrés et al. focuses on clustering for 2-dimensional data Andrés et al. [2012]. Especially meant for location-based systems (LBS). This location data has a big impact on privacy for individuals, which is why this paper focuses on integrating LDP with 2-d clustering.
They provide a solution by perturbing real-time location data by sending a random location in the same radius r. The level of privacy l is dependent on the size of r and ϵ-geo-indistinguishability is then defined as $\epsilon = l/r$.
LDP is used by randomly selecting a location on the device of a user. A set of interesting locations X based on the original $x \in X$ is extracted from a GPS service. These values are calculated based on the Euclidian distance $d(x, x')$ for which everything should be within a radius r. Finally, the noise is added based on a probabilistic method K, which is similar to that of the Laplace distribution. To support higher dimensions, they used the polar Laplace by using the distance. This mechanism was then further improved and named "Planar Laplace mechanism" Andrés et al. [2012]
The same mechanism could be applied to a tuple of points, by applying the distance between tuples of points. Method K can be individually applied to each point to achieve ϵ-geo-indistinguishability. The calculation of multiple locations is computationally heavy. Their current solution is to apply the technique for single locations for each location independently, but suffers from performance issues. Therefore, future studies should research the appliance of existing methods to solve this issue.

Research conducted by Min et al. Min et al. [2022] extends geo-indistinguishability for 3-dimensional data. Instead of using a plane for cartesian coordinates, they use a sphere for projecting 3D coordinates (X, Y, Z). Similar to the 2D-variant the La place algorithm was used. Therefore, in extension to $d(x, x')$ they define $d^3(x, x')$.

Summary

A few papers researched privacy-oriented distributed K-means. Their solutions yielded promising results but still needed a server in between. Although the results were perturbed, there was still information leaked regarding cluster information. Because K-means is a distance-based algorithm the geo-indistinguishability has shown promising results. Notably, the paper that is written by Min et al. Min et al. [2022] is very promising for K-means clustering. This paper is recent and there is currently no research conducted for applying LDP on n-dimensional data with K-means clustering. Therefore, this thesis will research the possibility of extending the 3-dimensional LDP to use for n-dimensional data. And the effectiveness of this data for usage with K-means clustering.

Validation

A common way of calculating the performance of LDP with K-means is calculating the error-rate against an original K-means algorithm. One way is to calculate the Relative Error (RE) Xia et al. [2020], Huang et al. [2021]. This is calculated to measure the deviation of centroids to the actual centroids (centroids produced by non LDP K-means). Another metric is named the adjusted mutual information (AMI) score. This score is used to quantify similarities between the original and perturbed cluster results Huang et al. [2021].

Bibliography

Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *CoRR*, abs/1212.1984, 2012. 3

D. Huang, X. Yao, S. An, and S. Ren. Private distributed K-means clustering on interval data. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 1–9, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/IPCCC51483.2021.9679364. 2, 3

Minghui Min, Liang Xiao, Jiahao Ding, Hongliang Zhang, Shiyin Li, Miao Pan, and Zhu Han. 3D geo-indistinguishability for indoor location-based services. *IEEE Transactions on Wireless Communications*, 21(7):4682–4694, 2022. doi: 10.1109/TWC.2021.3132464. 3

Chang Xia, Jingyu Hua, Wei Tong, and Sheng Zhong. Distributed K-Means clustering guaranteeing local differential privacy. *Computers & Security*, 90:101699, March 2020. ISSN 0167-4048. doi: 10.1016/j.cose.2019.101699. 2, 3