Local Differential Privacy for Data Clustering

Lisa Bruder[1] and Mina Alishahi[2]
[1]*Informatics Institute, University of Amsterdam, The Netherlands*
[2]*Department of Computer Science, Open Universiteit, The Netherlands*

Keywords: Local Differential Privacy, Clustering, Privacy, Non-Interactive LDP.

Abstract: This study presents an innovative framework that utilizes Local Differential Privacy (LDP) to address the challenge of data privacy in practical applications of data clustering. Our framework is designed to prioritize the protection of individual data privacy by empowering users to proactively safeguard their information before it is shared to any third party. Through a series of experiments, we demonstrate the effectiveness of our approach in preserving data privacy while simultaneously facilitating insightful clustering analysis.

1 INTRODUCTION

The widespread use of digital technology has led to a massive amount of data being available, bringing with it a significant responsibility to handle this data carefully. While data collection for statistical purposes is not new, the exponential growth in data volume coupled with the escalating threat of data breaches has intensified concerns regarding the confidentiality of sensitive information in recent years (Seh et al., 2020). Data breaches have highlighted the risks associated with unauthorized access to user data, raising awareness about the potential consequences of such breaches on individuals and organizations. Consequently, there is a growing demand to develop robust solutions to safeguard sensitive data and protect user privacy (Kasiviswanathan et al., 2011) (Xia et al., 2020).

One such task in data analysis is data clustering, which falls within the realm of unsupervised learning methods wherein patterns are discerned from unlabeled data points. The primary objective of data clustering is to unveil underlying patterns within a dataset by grouping data points into distinct clusters (Xu and Tian, 2015). However, given that data clustering often involves accessing sensitive user information, ensuring the protection of users' privacy is paramount.

Although numerous methods in the literature have been proposed to address the challenges of privacy-preserving clustering, these proposed solutions often fail to meet the following conditions:

- *Lack of Individual Privacy Protection:* Existing solutions often fail to protect single individual privacy locally on users' devices, necessitating trust in third-party entities, such as Differential Privacy in private clustering (Li et al., 2024).
- *Interactive Approach:* In cases where individual privacy is preserved, the proposed methods typically require continuous user involvement in training the clustering algorithm (Yuan et al., 2023)(He et al., 2024).
- *Narrow focus:* Many solutions are tailored for specific use cases, limiting their applicability to specific clustering training, such as exclusively for K-means clustering (Hamidi et al., 2018).
- *Computational Overhead or Loss of Utility:* These solutions may entail computationally intensive processes, such as encryption techniques (Sheikhalishahi and Martinelli, 2017b), or lead to a loss of utility, such as through anonymization methods (Sheikhalishahi and Martinelli, 2017a).

To overcome the mentioned constraints, our study introduces a novel framework leveraging Local Differential Privacy (LDP), wherein individual users protect their information by perturbing their data locally on their devices before sharing it with a third party, such as an aggregator (Alishahi et al., 2022). The aim of perturbation is to ensure that the estimation expectation remains unbiased and to minimize statistical variance as much as possible. Specifically, we employ an LDP-based frequency estimation technique, a fundamental statistical objective under local differential privacy protection.

To explore the effectiveness of LDP-based frequency estimation for private clustering, we conduct

Bruder, L. and Alishahi, M.
Local Differential Privacy for Data Clustering.
DOI: 10.5220/0012838800003767

In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pages 820-825
ISBN: 978-989-758-709-2; ISSN: 2184-7711

a series of experiments. These experiments focus on investigating the impact of key parameters, including the number of cells used for discretization, the size of the input dataset, and the privacy budget. By systematically varying these parameters, we aim to gain insights into their influence on the LDP-based clustering process and assess the performance of our approach under different settings.

2 PRELIMINARIES

This section presents clustering and local differential privacy as preliminary concepts employed in our proposed framework.

***k*-means Clustering:** is a method for partitioning a dataset into k clusters based on similarity. It iteratively assigns data points to the nearest cluster centroid and updates centroids to minimize intra-cluster variance. Formally, given a dataset X consisting of n data points $\{x_1, x_2, \ldots, x_n\}$ and the desired number of clusters k, the K-means clustering algorithm aims to partition the data into K clusters, $C = \{C_1, C_2, \ldots, C_K\}$, such that it minimizes the within-cluster sum of squares (WCSS).

Local Differential Privacy (DP): is a privacy-preserving mechanism in which an aggregator gathers information from users who has some level of distrust but are still willing to engage in the aggregator's analysis.

Formally, a randomized mechanism $\mathcal{M}$ adheres to ε-LDP if and only if, for any pair of input values $v, v' \in \mathcal{D}$ and for any possible output $S \subseteq Range(\mathcal{M})$, the following inequality holds:

$$Pr[\mathcal{M}(v) \in S] \leq e^{\varepsilon} Pr[\mathcal{M}(v') \in S] \tag{1}$$

when ε is understood from the context, we refer to ε-LDP simply as LDP.

Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) (Erlingsson et al., 2014): introduced by Google, is a hash-based frequency statistical method that randomly selects a hash function $\mathcal{H}$ from a hash function family $\mathbb{H} = \{\mathcal{H}_1, \ldots, \mathcal{H}_m\}$, where each function outputs an integer in $[k] = \{0, 1, \ldots, k-1\}$. RAPPOR then encodes the hash value $\mathcal{H}(v)$ as a k-bit binary vector and randomized response is performed on each bit. Accordingly, the encoded vector v_t is shaped as follows:

$$v_t[i] = \begin{cases} 1 & if \quad \mathcal{H}(v) = 1 \\ 0 & otherwise \end{cases} \tag{2}$$

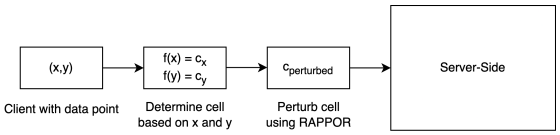


Figure 1: Server side.

The encoded vector is then perturbed as:

$$Pr[\hat{v}_t[i] = 1] = \begin{cases} 1 - \frac{1}{2}f & if \quad v_t[i] = 1 \\ \frac{1}{2}f & if \quad v_t[i] = 0 \end{cases} \tag{3}$$

where $f = 2/(e^{\frac{\varepsilon}{2}} + 1)$. The aggregator employs Lasso regression to improve the estimated frequency value out of collected reports.

3 METHODOLOGY

Our approach is constituted of the following steps as shown in Figures 1 and 2:

Discretization (Shaping the Cells): In this critical phase, the aggregator employs domain knowledge regarding feature ranges to discretize the dataset, thus delineating cells to accommodate data points falling within these ranges. The process entails setting intervals by uniformly partitioning the anticipated range of continuous values present in the dataset. For instance, in the context of adult height values expected to range between 1.4 and 2.0 meters, the process may involve dividing this range uniformly into three intervals, yielding non-overlapping boundaries such as $[1.4, 1.6)$, $[1.6, 1.8)$, and $[1.8, 2.0]$. Each interval demarcates a cell boundary, determining where data points align in relation to these boundaries. Consequently, data points are assigned to specific cells based on their relative positioning within these intervals. This process serves as a foundational step in subsequent aggregation and analysis tasks, enabling effective handling of continuous data.

The aggregator assigns integer identifiers to each cell, and subsequently discloses both the boundaries of these cells and their respective identifiers to users.

LDP-Based Frequency Estimation: LDP-based frequency estimation offers a means for the aggregator to approximate the count of individuals within specific cells, all while maintaining user privacy regarding their cell associations. Users start by pinpointing the cell where their data resides and extracting its corresponding integer identifier. Through the LDP-based frequency estimation algorithm, users perturb their answer by an alternative integer identifier (out of the list of existing identifiers). This method ensures that individual contributions remain confidential, yet

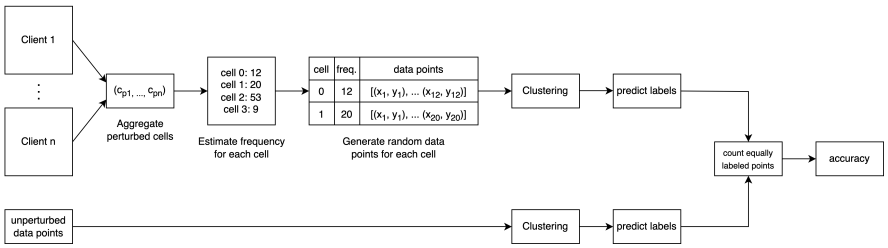


Figure 2: Client-side (top) and Server-Side Architecture (bottom).

still allows for reliable estimation of aggregate frequencies within cells. In this study, we utilize RAPPOR as the chosen LDP-based frequency estimation protocol due to its demonstrated accuracy in this context. Nonetheless, our proposed framework is flexible and can accommodate any alternative LDP-based frequency estimation technique, providing versatility and adaptability to different privacy requirements and data characteristics.

Generating New Dataset: The aggregator gathers cell identifiers submitted by users and then associates a new data point with each user based on this information. Specifically, if a user reports that their data point d belongs to cell c^*, the aggregator assigns a new point, denoted as d^p, randomly within this identified cell. This process ensures that each user's reported data point is mapped to a representative point within the corresponding cell.

Clustering: Now, armed with the newly generated dataset, the aggregator proceeds to train a clustering algorithm. This algorithm's structure can subsequently be shared with any third party for their use. Our proposed approach offers flexibility by remaining agnostic to the choice of clustering algorithm. In this particular study, we opt for the widely used k-means clustering method. This selection, however, does not constrain the applicability of our approach, as it can seamlessly integrate with various clustering techniques depending on specific analysis requirements and preferences.

Theorem 1 *As the number of cells and privacy budget* ε *increase, the error of* ε*-LDP frequency estimation techniques, such as RAPPOR, also increases. Conversely, increasing the number of data points reduces the error rate of these techniques.*

Proof: This conclusion directly stems from the error characteristics of frequency estimation techniques. The Mean Squared Error bound of RAPPOR is expressed as $\Theta\left(\frac{e^{\varepsilon} r}{n\left(e^{\frac{\varepsilon}{2}}-1\right)^2}\right)$, where r denotes the number of cells and n represents the size of the dataset (Wang et al., 2020). From this formula, it is evident that an increase in ε and r leads to a higher error bound, while an increase in the number of users (data points) decreases the error bound.

It is noteworthy that while we specifically discuss the error bound of RAPPOR here, this argument holds true for other frequency estimation protocols, such as Hadamard and RR techniques (Wang et al., 2020).

Proposition 1 *While increasing the number of cells negatively impacts the accuracy of frequency estimation techniques, it improves the accuracy of clustering algorithms trained on published data within our methodology.*

Proof: If the frequency estimation technique adequately preserves the distribution of each cell, the size of cells still affects the accuracy of clustering. This is because if a cluster boundary intersects the middle of a cell, a larger cell size increases the risk of a data point falling outside the cluster boundary when the aggregator returns a randomized point based on the shared cell identifier. In contrast, smaller cells increase the likelihood of cells aligning more closely with cluster boundaries, thereby enhancing clustering accuracy. This can be formulated as following:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x-\mu_i\|^2 \stackrel{?}{=} \arg\min_{S} \sum_{i=1}^{k} \sum_{S_i} \sum_{x \in \Delta_j} \|x-\mu_i\|^2$$

where S_i is the i'th cluster, and $S_i = \sum_j \Delta_j$, for Δ_j considered as the cells overlapping with cluster S_i.

Proposition 2 *Increasing the number of data points and privacy budget improves the clustering accuracy.*

Proof: The variations in the number of data points and privacy budget do not directly influence the accuracy of clustering. However, they indirectly impact clustering accuracy through their effect on accurate frequency estimation. Since clustering serves as a post-processing step in our methodology, the improvements in frequency estimation resulting from an increase in data points and privacy budget should also lead to enhanced clustering accuracy.

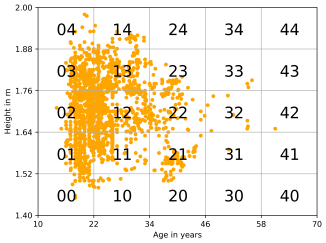


Figure 3: Original data points with 25 cell grid.

4 EXPERIMENTS

4.1 Experimental Set-Up

Experiment Environment: We program the code for our experiments using the programming language Python. We use the pure-ldp[1] package's RAPPOR client- and server-side implementation as well as scikit-learn[2] for k-means clustering and matplotlib[3] to plot the data we gathered.

Dataset: For our experiments, we utilize the "Estimation of obesity levels based on eating habits and physical condition" dataset sourced from the UCI Machine Learning Repository[4]. This dataset encompasses health information gathered from individuals in Mexico, Peru, and Colombia, supplemented with synthetically generated data derived from the original dataset. Its relevance to our research lies in its inclusion of sensitive health-related data, presenting a realistic scenario where data privacy is of paramount importance. The dataset contains 2111 records, and it is described with 17 features. Out of the 17 features available, we specifically focus on utilizing the two continuous features, namely 'Age' and 'Height', for visualization purposes. Additionally, we set aside 20% of our original data for future evaluation, preserving these data points to assess the accuracy of our newly developed model.

Discretization: We discretize the values into uniform cells, ranging from 1.4 to 2.0 meters for height and 10 to 70 years for age, based on the feature value ranges. Each discretized value is then assigned an integer identifier corresponding to the cell it falls into. For example, a data point with a discretized age value of "1" (age in the interval [22, 34)) and a discretized height value of "2" (height in the interval [1.64, 1.76)) would be associated with the cell "12". Figure 3 shows how the cells are shaped on our original data.

[1]https://pypi.org/project/pure-ldp/
[2]https://pypi.org/project/scikit-learn/
[3]https://pypi.org/project/matplotlib/
[4]https://doi.org/10.24432/C5H31Z

Table 1: Parameters tuning.

Parameter	Range of values
Number of cells	4 - 100
Fraction of data points	0.1 - 1.0
RAPPOR Epsilon	1, 2, 4, 8

Frequency Estimation: Upon determining the cell identifier where their data resides, each user employs RAPPOR locally on their device to perturb the cell number. For RAPPOR implementation, we adhere to the defaults established by Google in their RAPPOR demo: setting the bloom filter size to 16 and utilizing 2 hash functions.

Clustering: Initially, we randomly assign a new point for each point that falls within a cell. Subsequently, we train the k-means clustering algorithm on both the original data and the data generated by RAPPOR. We opt for $k = 5$, determined through the elbow technique applied to the original dataset. It's worth noting that while k-means is an iterative clustering algorithm, this iterative process occurs solely on the aggregator side, with only interacting users once during the collection of their perturbed data.

Evaluation Metric: To assess the efficacy of our LDP-based clustering approach, we conduct a comparison of labelings between the 20% reserved original data and the RAPPOR-generated data. Firstly, we determine centroids for both datasets through k-means clustering. Each data point is then assigned to the nearest centroid, allowing us to evaluate the consistency of labels. We measure the accuracy by calculating the percentage of data points accurately labeled. This is achieved by dividing the number of data points labeled the same between the original and RAPPOR-generated data by the total number of data points used for prediction. The resulting percentage represents our utility metric. A higher percentage indicates a closer alignment between the RAPPOR-based model and the original dataset.

Parameter Tuning: To evaluate the effects of various parameters within our framework, we explore different values for the number of cells, fraction of dataset used as input source, and privacy budgets. The range of parameters varied for our experiments is summarized in Table 1. It is noteworthy that while one parameter varies, the other two parameters remain fixed. For enhanced reliability, we repeat our experiments 50 times and report the average results.

4.2 Results

The Influence of Cell Count and privacy Budget: Our initial experiments aims to investigate the influence of cell count and privacy budget on accuracy. We

tested cell counts ranging from 4 to 100 cells, specifically focusing on square numbers within this range. Figure 4a shows the outcome of this experiment for four different epsilon values 1, 2, 4, and 8. The trend reveals that smaller cell sizes, corresponding to higher cell counts, generally result in greater accuracy across all epsilon values up to 81 cell counts. This observation underscores the existence of a trade-off associated with the number of cells, where simply increasing cell counts does not guarantee accuracy enhancement. This observation resonates with the findings of our theoretical analysis. The variation of privacy budget does not show a significant change in accuracy. This can be resulted from the distribution of data under analysis. In other words, even the higher randomness noise still keeps the structure of data properly for clustering. It of course needs more investigation in future studies for variety of datasets. This outcome specifically suggests the use of lower epsilon values (higher privacy gain) in our methodology. To gain a better insight on the dispersion of accuracy on 50 runs of experiments, we depict this variance in Figure 4b for epsilon 1 and different cell counts. It can be seen that the widest range of values is observed with 64 cells, where accuracy spans from a minimum of 18% to a maximum of 90%, reflecting a variance of 72%. While this dispersion is unavoidable due to the randomness inherent property of LDP, it can be seen that yet increasing the number of cell counts leads to the improvement in accuracy.

The Influence of Dataset Size: Figure 6 shows the impact of dataset size by considering fractions of data on the accuracy of our methodology. For this dataset, the amount of data points in the data set does not seem to consistently influence the accuracy we can achieve using RAPPOR. The values stay pretty consistent across all data set sizes and there is no consistent trend. The chosen Epsilon value does not seem to have much influence on the accuracy either. This can be resulted from the distribution of our dataset and the precision of RAPPOR in preserving the distribution of data even in smaller sizes.

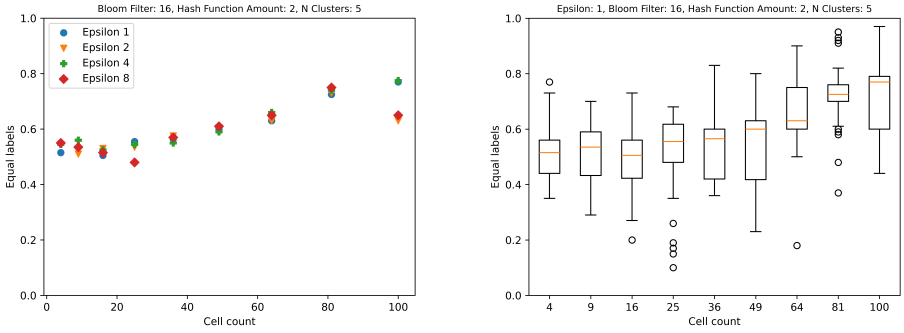
Once again, Figure 6 reveals a notable disparity among accuracy values across experiment runs. Some runs exhibit considerably low accuracy rates, while others nearly achieve 100%. Particularly intriguing is the scenario involving the smallest dataset size, comprising only a fraction of 0.1 relative to the original dataset size. Here, the attained accuracy spans from a minimum of 14% to a maximum of 95%. Despite this variance, computing the median accuracy across all experiment runs still yields commendable results. Interestingly, the dataset sample size of 0.9 of the original dataset demonstrates the least dispersion, with accuracy ranging from 62% to 96%.

Discussion. We present the guidelines for using our framework, our experimental findings, and our plan for future directions.

- We have assumed that the features are continuous. However, our methodology can also be applied on discrete features. To this end, it is enough that we use the boundary of cells as shared value and as randomized value by the aggregator.
- We found that the number of cells has impact on the accuracy of our methodology both in negative and positive way. There is a trade-off in the number of cells for each dataset that the accuracy is optimized. However, it should be noted that the aggregator has no knowledge about the data to offer the optimum number of cells in advance. To this end, in the future directions, we plan to design a privacy-preserving mechanism to infer the optimum number of cells without accessing the original data.
- Although our experiments did not show a considerable impact of dataset size on accuracy, we believe that this requires more extensive experiments when also the dataset distribution also comes under consideration. Given the inherent property of LDP mechanism, we expect that the size of dataset single alone might night affect the outcome if the dataset if the data is almost well evenly distributed across all cells.
- We found the optimum number of clusters using elbow technique on original datset. This is something that the aggregator does not know without accessing the original data. In future direction, we plan to investigate the impact of the number of clusters on accuracy.

5 CONCLUSION

This study introduces a novel framework that leverages Local Differential Privacy (LDP) to safeguard individual data privacy, empowering users to take proactive measures to protect their information before any sharing occurs with third parties in a non-interactive engagement of users. Through a comprehensive series of experiments, we provide compelling evidence of the efficacy of our approach in preserving data privacy while also enabling meaningful and insightful clustering analysis.



(a) Varying cell counts over 50 runs.

(b) Box plot for epsilon 1.

Figure 4: The impact of cell counts and privacy budget.

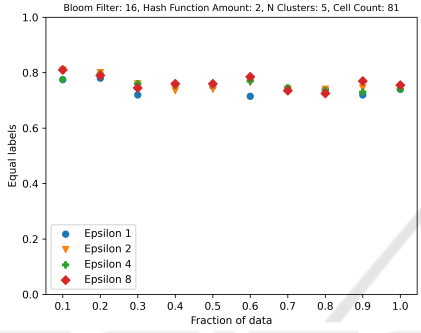


Figure 5: Varying dataset size for Epsilon 1, 2, 4, 8 with median accuracy over 50 runs.

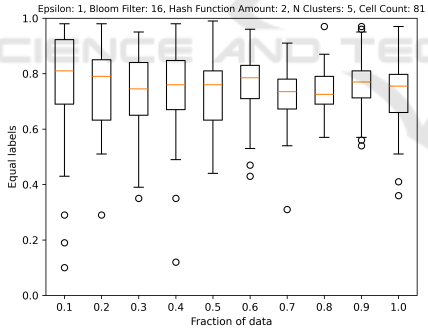


Figure 6: Epsilon 1 box plot based on all 50 accuracy values.

REFERENCES

Alishahi, M., Moghtadaiee, V., and Navidan, H. (2022). Add noise to remove noise: Local differential privacy for feature selection. *Comput. Secur.*, 123:102934.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067.

Hamidi, M., Sheikhalishahi, M., and Martinelli, F. (2018). Privacy preserving expectation maximization (EM) clustering construction. In *DCAI conference*, volume 800 of *Advances in Intelligent Systems and Computing*, pages 255–263. Springer.

He, Z., Wang, L., and Cai, Z. (2024). Clustered federated learning with adaptive local differential privacy on heterogeneous iot data. *IEEE Internet of Things Journal*, 11(1):137–146.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

Li, Y., Wang, S., Chi, C.-Y., and Quek, T. Q. S. (2024). Differentially private federated clustering over non-iid data. *IEEE Internet of Things Journal*, 11(4):6705–6721.

Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., and Ahmad Khan, R. (2020). Healthcare data breaches: Insights and implications. *Healthcare*, 8(2).

Sheikhalishahi, M. and Martinelli, F. (2017a). Privacy preserving clustering over horizontal and vertical partitioned data. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 1237–1244.

Sheikhalishahi, M. and Martinelli, F. (2017b). Privacy preserving hierarchical clustering over multi-party data distribution. In *Security, Privacy, and Anonymity in Computation, Communication, and Storage*, volume 10656 of *Lecture Notes in Computer Science*, pages 530–544. Springer.

Wang, S., Qian, Y., Du, J., Yang, W., Huang, L., and Xu, H. (2020). Set-valued data publication with local privacy: tight error bounds and efficient mechanisms. *Proc. VLDB Endow.*, 13(8):1234–1247.

Xia, C., Hua, J., Tong, W., and Zhong, S. (2020). Distributed k-means clustering guaranteeing local differential privacy. *Computers & Security*, 90:101699.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193.

Yuan, L., Zhang, S., Zhu, G., and Alinani, K. (2023). Privacy-preserving mechanism for mixed data clustering with local differential privacy. *Concurr. Comput. Pract. Exp.*, 35(19).