Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 3, 2024

Flexible Local Differential Privacy Mechanism for Collecting Location Data

Tianci Lv *, Mengyuan Cheng

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China
* **Corresponding author**: Tianci Lv (Email: 622220070033@mails.cqjtu.edu.cn)

Abstract: To address issues in existing location data collection methods, such as poor data utility and large deviations between perturbed and true locations, this paper proposes a personalized local differential privacy (LDP) mechanism for location data collection. Users can select their privacy protection range based on their needs, limiting the perturbed output to this range, thereby improving data utility. To address the issue of large deviations between perturbed and true locations, we introduce a new strategy where the location is perturbed with a probability that is higher the closer it is to the true location. By analyzing the mutual information upper bound of the true and estimated location distributions, the optimal perturbation probability range is determined. Finally, a probability transition matrix is generated from the location transfer probabilities, and the true location distribution is estimated from the perturbed location distribution.

Keywords: Location Privacy; Personalization; Differential Privacy.

1. Introduction

With the development of the Internet of Things (IoT), people increasingly rely on location-based services provided by mobile applications, which typically require users to share their location information. Governments and businesses can leverage this location data to predict traffic flow, conduct urban planning, and improve service quality and marketing strategies. However, location information is sensitive, and its leakage can expose users' personal information, posing privacy risks. Therefore, collecting location data while protecting user privacy has become a significant challenge.

To address this, scholars have proposed methods such as Differential Privacy (DP) and Local Differential Privacy (LDP) to protect location privacy. The core idea of DP is to add noise to users' location data to preserve privacy. However, it assumes that third parties will not leak users' original data, overlooking the potential risk of data leakage. To mitigate this, LDP has been introduced, where users locally perturb their location data before submitting it to service providers, thus avoiding the risk of exposing raw data. Currently, there are two main approaches for collecting location data under LDP: one involves dividing the space into grids for frequency estimation perturbation and aggregation, but this method neglects the relationships between grids and variations among users; the other discretizes the spatial region for perturbation, which avoids the errors from grid division but still treats adjacent regions equally. A more advanced approach is continuous perturbation, where location data is normalized before perturbation. This method avoids errors from spatial discretization and preserves the characteristics of neighboring locations. However, because the high-probability regions after normalization are fixed, it may lead to discrepancies in perturbation probabilities across different spatial ranges.

To address the aforementioned issues, this paper proposes an adaptive Circular Fusion Mechanism that combines flexible local differential privacy (FLDP) with user-defined privacy protection ranges for location data collection. Users can select a privacy protection range based on their individual needs, ensuring that the perturbed location data does not leak sensitive information. The true location is perturbed with varying probabilities to a circular region centered around the true location, with the radius defined by the privacy protection range. This circular region is divided into multiple concentric rings, where the probability of perturbation is higher the closer the location is to the true position. To prevent an excessive number of rings from reducing the perturbation probability, a probability lower bound for the rings is set, allowing for adaptive adjustment of the number of rings. Finally, a probability transition matrix is generated based on location transfer probabilities, and the true location distribution is estimated by combining the perturbed location distribution frequencies.

2. Fundamental Knowledge

2.1. Local Differential Privacy

Definition 1 A data perturbation algorithm M on an input set X and output set Y satisfies Local Differential Privacy if and only if for any two inputs $x_1, x_2 \in X$, the probability of obtaining any output $y \in Y$ satisfies:

$$\Pr[f(x_1) = y] \le e^{\varepsilon} \Pr[f(x_2) = y] \tag{1}$$

The parameter ε in the definition is referred to as the privacy budget. The smaller the privacy budget, the higher the protection of the original data, but the utility of the perturbed data is lower.

2.2. Flexible Local Differential Privacy

Definition 2 Given a mechanism M with domain I and range R, and for any two items t and t' having outputs $R(t)$ and $R(t')$ via mechanism M, if the range satisfies inequality (2) and the output satisfies inequality (3), then we say that M satisfies $(\varepsilon, \eta) - FLDP$.

$$\min_{t,t' \in I} \frac{|R(t) \bigcap R(t')|}{\max\{|R(t)|, |R(t')|\}} > \eta \tag{2}$$

$$\max_{s \in R(t) \bigcap R(t')} \frac{\Pr[M(t) = s]}{\Pr[M(t') = s]} \le e^{\varepsilon} \tag{3}$$

In Definition 2, any output s can be transformed by an element of a group $G \subseteq I$ via the perturbation mechanism M; thus, any element $t \in G$ can be concealed within the group G.

3. Privacy-preserving Location Data Collection Scheme

The Circular Fusion Mechanism for location data collection includes the following steps: 1) The client and server negotiate the privacy range R and budget ε; 2) The user perturbs the data based on the range and budget ε, then sends it to the server; 3) The server aggregates the perturbed data to obtain the frequency distribution of perturbed locations; 4) The server uses the transition probability matrix to estimate the true location distribution.

3.1. Perturbation Mechanism

To improve location distribution accuracy, the closer the perturbed location is to the true position, the better the data utility. Incorporating the concept of privacy protection range, we propose a new perturbation mechanism, CFM. In this mechanism, the perturbation output range is determined by the user's privacy protection range, minimizing unnecessary perturbations to reduce errors. To further reduce the bias between the perturbed and true locations, the probability of perturbation varies with distance—closer locations have higher probability, while farther locations have lower probability.

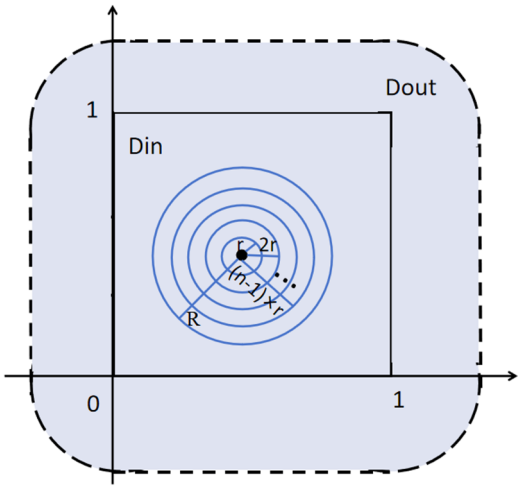


Figure 1. Perturbed Output of CFM

The perturbed output is shown in Fig 1. All true locations are normalized and fall within the input domain Din. The perturbed locations are confined to the output domain Dout, which is shared by all users, with an area of $\pi R^2 + 1 + 4R$. Each user has a selected privacy protection range R, and their output domain is a circular region centered on the true location with a radius R. Perturbed locations must fall within this region. The circular area is divided into a central circular region with radius r and several concentric rings, with outer radii ranging from $2r$ to $(n-1)r$, where R is the largest. The true location is randomly perturbed across the output domain, with the probability decreasing from the innermost circular region to the outer rings, as follows:

$$\Pr(v'|v) = \begin{cases} p_1 & dist(v,v') \le r \\ p_2 & r < dist(v,v') \le 2r \\ \cdots\cdots & \\ p_{n-1} & (n-2)r < dist(v,v') \le (n-1)r \\ p_n & (n-1)r < dist(v,v') \le R \end{cases}$$

Here,

$$p_n = \frac{1}{(n-1)^2 \pi r^2 e^{\varepsilon} + \pi R^2 - [\sum_{i=0}^{n-2} (2i+1) \cdot i + (n-1)^2] \cdot \pi r^2},$$

$p_1 = p_n \cdot e^{\varepsilon}$, $p_i = p_{i-1} - p_n (i \ge 2)$, n represent the number of rings (with the central circular region as a special ring with an inner radius of 0), $dist(v,v')$ is the distance between the true and perturbed locations, and the total probability sums to 1.

The adaptive number of rings n is determined to ensure that the average probability for regions closer to the true location within the output domain does not become too low. To achieve this, it is required that p_i satisfies $p_i \ge (p_1 + p_n)/2$,from which $e^{\varepsilon} - (n-1) \ge (1+e^{\varepsilon})/2$ can be derived, leading to $n \le (1+e^{\varepsilon})/2$. This allows the determination of the number of rings n under the current privacy budget ε. The number of rings increases automatically with the privacy budget, partitioning more probability regions while maintaining privacy protection, thereby further reducing the bias between the true and perturbed locations.

3.2. Frequency Distribution Estimation

After receiving all perturbed locations, the server performs a distribution analysis and estimates the true location distribution in the input domain based on the probability transition between input and output domains. Users are grouped by their chosen privacy protection range R, as locations with the same R share the same transition matrix. Frequency estimation is performed for each group, contributing to the overall true frequency estimation. This process involves dividing the input domain into grids, where each grid represents a location set V_i, and counting the perturbed locations in the corresponding grid of the output domain.

Since any position in the input domain can be perturbed to any location set within the safe range R in the output domain, let $\Pr[D_{\text{in}}(v_i) | D_{out}(v_j)]$ represent the perturbation probability of positions from input domain v_j to output domain v_i. The calculation formula is as follows:

$$\Pr[D_{\text{in}}(v_i) | D_{out}(v_j)] = \frac{\sum_{z=1}^{n-1} cov_{ijz}(e^{\varepsilon} - z) + area_{ij}}{(n-1)^2 \pi r^2 e^{\varepsilon} + \pi R^2 - \left[\sum_{i=0}^{n-2} (2i+1) \cdot i + (n-1)^2\right] \pi r^2} \quad (4)$$

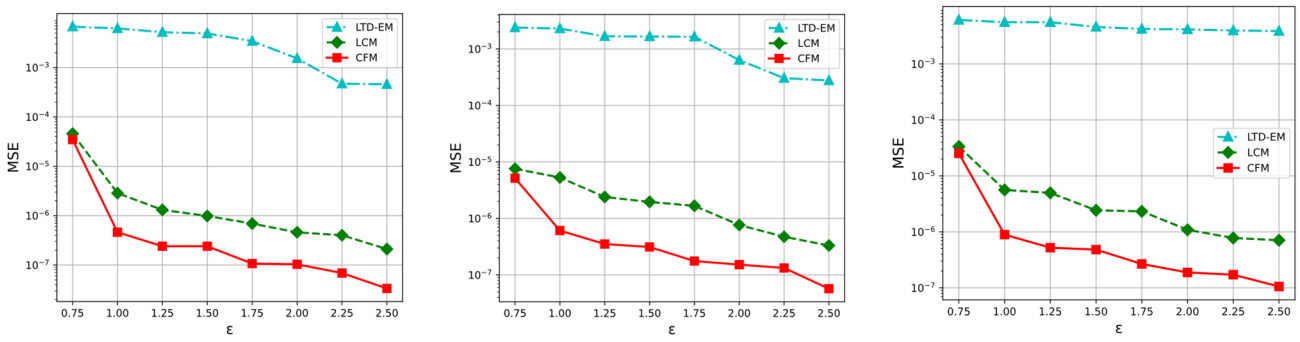
Here, $area_{ij}$ represents the area of the intersection between the output regions of v_j and v_i,and cov_{ijz} represents the

area of the intersection between the circular region in the input domain centered at v_i with radius $z \times r$ and the high-probability region in the output domain.

For users with the same privacy budget ε and privacy protection range R, the following holds for each location set $D_{in}(v_i)$ in the input domain:

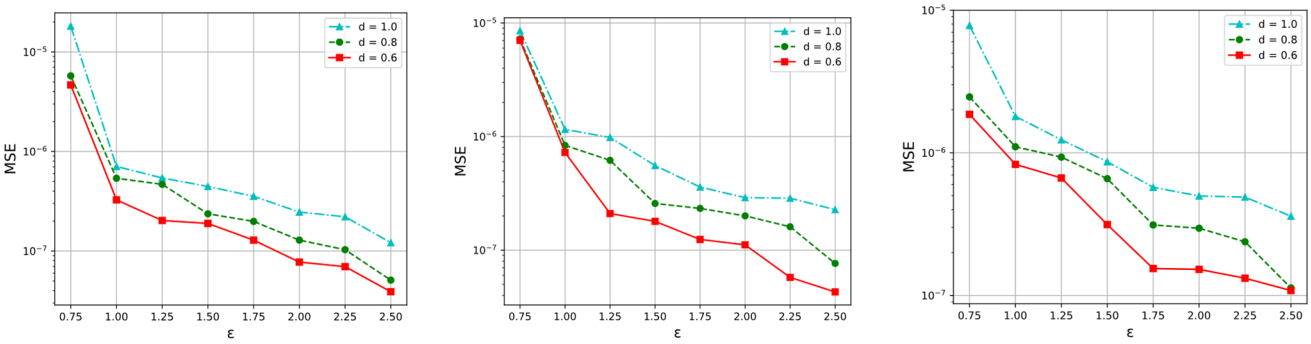
$$f_i^* = \sum_{v_j \in D_{out}} f_j \times \Pr[D_{in}(v_i) \mid D_{out}(v_j)] \quad (5)$$

Here, f_i^* represents the perturbation frequency of this group of users for each location set $D_{in}(v_i)$ in the input domain, f_j is the true user distribution frequency in the input domain, and m is the number of location sets. By constructing m equations based on (4) and (5), the position distribution of this group of users in the input domain can be estimated through solving these equations.



(a)T-Drive (b) Facebook (c) Gowalla

Figure 2. MSE of Different Algorithms on Real Datasets



(a) T-Drive (b) Facebook (c) Gowalla

Figure 3. MSE of CFM with Different Privacy Protection Ranges on Real Dataset

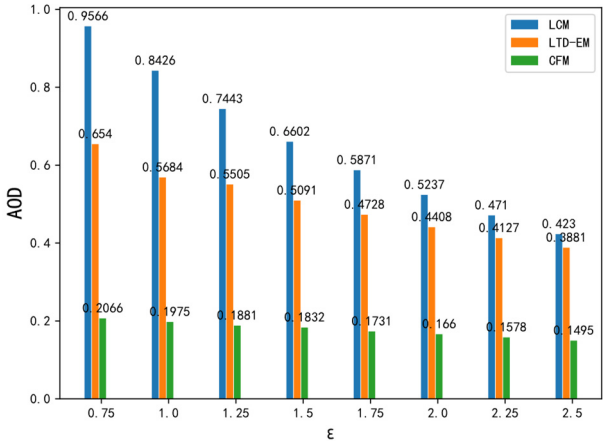


Figure 4. AOD of Different Methods on Real-World Dataset

4. Experimental Results and Analysis

The experimental environment is set as follows: Windows 10, Intel Core i5-8300H CPU @ 2.30GHz, 16GB RAM. The datasets used in the experiment include three real datasets: the Gowalla, which are location check-ins from social media platforms; the Facebook dataset, which contains check-ins over a 100 km² area; and the T-Drive dataset, which consists of GPS trajectories from 10,357 taxis in Beijing between February 2 and 8, 2008. Each position point in the trajectories is treated as an individual location. For this experiment, the spatial range of the datasets is divided into 10×10 grid cells, with the center of each grid serving as a location. The frequency of each grid cell is the total frequency of all positions within that grid.

Fig. 2 shows the relationship between MSE and privacy budget ε in real-world datasets. Among the various methods for estimating the location distribution frequency across the entire map, the performance of CFM significantly outperforms the other two schemes. This indicates that CFM is an effective solution. On real datasets, CFM exhibits the smallest error, followed by LCM, while LTD-EM shows the largest error. It can be observed that, for the three datasets, the error of LTD-EM decreases more slowly with increasing privacy budget compared to the other two algorithms.

Fig. 3 shows the relationship between the area of the privacy protection zone and the MSE in the CFM algorithm. In the experiment, the diameters of the privacy protection zones were set to 1.0, 0.8, and 0.6, with these three parameters

serving as control groups. The experimental results show that as the diameter of the circular protection zone decreases (i.e., the perturbation output range becomes smaller), the MSE of CFM also decreases. This indicates that the size of the perturbation output range affects the frequency estimation error, and appropriately reducing the perturbation output range can improve the accuracy of the frequency estimation.

Fig. 4 shows the relationship between the offset distance between perturbed and real positions and the privacy budget for multiple algorithms. Since the average offset distances for the three algorithms across the three real datasets are almost identical, only the results for one dataset are shown here. The experimental results indicate that the offset distances for all three methods decrease as the privacy budget increases. Among them, LTD-EM performs better than LCM, and CFM significantly reduces the offset distance compared to the other methods. This suggests that CFM effectively reduces the offset between perturbed and real positions, thereby improving the utility of the perturbed data.

5. Conclusion

To address the issues of low data utility and large location perturbation bias caused by using the same perturbation output range for all users, this paper combines personalized perturbation range selection with flexible local differential privacy, introducing an adaptive continuous perturbation method. Users can personalize the perturbation area, reducing unnecessary perturbation and minimizing error. Additionally, perturbation probabilities are adjusted based on the distance from the true location, significantly reducing the bias between the true and perturbed positions. Experimental results show that, compared to existing methods, the proposed approach significantly improves data utility.

References

[1] N. Alikhani, V. Moghtadaiee, A.M. Sazdar, et al. A privacy preserving method for crowdsourcing in indoor fingerprinting localization, 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2018, p. 58-62.

[2] Y. Huang, H. Wang. Frequent Trajectory Mining with Local Differential Privacy, 2023 15th International Conference on Advanced Computational Intelligence (ICACI), IEEE, 2023, p. 1-6.

[3] H. Navidan, V. Moghtadaiee, N. Nazaran, et al. Hide me behind the noise: Local differential privacy for indoor location privacy, 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2022, p. 514-523.

[4] D. Hong, W. Jung, K. Shim. Collecting geospatial data with local differential privacy for personalized services, 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, 2021, p. 2237-2242.

[5] Ú. Erlingsson, V. Pihur, A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, p. 1054-1067.

[6] T. Wang, J. Blocki, N. Li, et al. Locally differentially private protocols for frequency estimation, 26th USENIX Security Symposium (USENIX Security 17), 2017, p. 729-745.

[7] Y. Ye, M. Zhang, D. Feng. Collecting Spatial Data Under Local Differential Privacy, 2021 17th International Conference on Mobility, Sensing and Networking (MSN), IEEE, 2021, p. 120-127.

[8] T. Murakami, Y. Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation, 28th USENIX Security Symposium (USENIX Security 19), 2019, p. 1877-1894.

[9] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis, Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, 2009, p. 19-30.