Mobile Crowdsensing Location Aggregation Data Release with Differential Privacy Protection

Liuqiaoyu Mo[1], Xiaofang Deng[1(✉)], Xingshan Zeng[2], Lina Gao[2], and Lin Zheng[1]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guangxi, China
xfdeng@guet.edu.cn

[2] Army Infantry College Shijiazhuang Campus, Shijiazhuang, China

Abstract. In the scenario of mobile crowdsensing, the release of location aggregation data has supported the development of many domains, however, personal sensitive privacy information of mobile users implicit in location aggregated data may be disclosed as a result, which greatly discourages people from sharing their own location data. In this paper, we propose a differential privacy-based mobile crowdsensing location aggregation data release scheme Re-LDCR. Specifically, to make the privacy budget application adaptable to data changes while avoiding excessive budget consumption, we use the defined data change rate as the basis for budget allocation, and combine the recycle factor to limit the allocation proportion. Then, to improve the noise resistance of individual data, we comprehensively consider the data change characteristics and privacy protection ability, using BIRCH clustering to group the data with similar features. Finally, the combination of prediction, sampling, perturbation, and filtering mechanisms ensures the data utility of privacy protection results. Experimental results show that the proposed Re-LDCR outperforms the existing scheme and achieves a balance between privacy protection effectiveness and data utility.

Keywords: Differential Privacy · Mobile Crowdsensing · Location Data Release

1 Introduction

The current boom in Internet-accessible mobile devices has driven society into the big data era. To make better use of the implied value of big data, mobile crowdsensing (MCS) has emerged, combining mobile sensing and crowdsourcing ideas to achieve complex, dynamic, large-scale sensing task data collection and processing at a lower cost by motivating a wide range of mobile users to participate in sensing tasks, effectively improving social benefits and the quality of life of the general public. At present, MCS provides solid data support for the development of urban planning, traffic management, environmental monitoring, social networking, and other fields, and has replaced traditional static sensing devices as the new paradigm for IoT sensing data acquisition.

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 340–347, 2025.
https://doi.org/10.1007/978-3-031-60347-1_28

In location data application scenarios, aggregated location data collected and released by the MCS platform can provide strong support for various geo-information applications and decision-making processes. However, untrustworthy data users can easily infer the trajectory of mobile users by observing aggregated location data published by the MCS platform, resulting in a constant risk of leakage of sensitive personal information of participating mobile users, which greatly affects the enthusiasm of mobile users to participate in sensing tasks.

Differential privacy [1] introduces random noise into the data and sacrifices some of the data utility to achieve strong privacy protection, which is now widely used for privacy protection in the data publishing process. Kellaris et al. [2] proposed w - event differential privacy, which can guarantee the privacy of individuals during the continuous release of infinite data sequences. Wang et al. [3] proposed a RescueDP scheme based on w - event differential privacy, which designed a data grouping mechanism to merge the small-value data, improve the anti-noise ability of data and effectively enhance the final release utility, and further, Wang et al. [4] used a dynamic programming method to optimize the flexibility of the intelligent grouping mechanism. Huo et al. [5] adopted K-means algorithms to improve data grouping efficiency. Existing research demonstrates that the utility of differential privacy protection data can be optimized by combining small-value data. However, the situation that large-value data can further improve the overall anti-noise performance of a group is ignored. Meanwhile, the spatiotemporal characteristics of location aggregation data pose privacy protection challenges.

To this end, we propose a differential privacy protection scheme called Re-LDCR, aiming at providing w - event level differential privacy protection for the release of location-aggregated data in MCS. Specifically, to ensure the rationality of privacy budget allocation in the privacy protection process, we use data change rate as the main basis for privacy budget allocation, while the recycle factor works for preventing budget wastage. For assurance of final data availability, based on the BIRCH clustering algorithm, a data grouping scheme is designed, which groups the mobile aggregated data according to the feature of data changes and privacy protection capabilities. Finally, Laplace noise is added to the groups and the Kalman filter is applied to optimize the utility. We conducted experiments on a real-world dataset and compared Re-LDCR with the existing scheme to verify its performance.

The remaining work is organized as follows: Sect. 2 introduces the theories of differential privacy. Section 3 describes the overall architecture and details of Re-LDCR. Section 4 verifies the performance of Re-LDCR. Section 5 is the summary.

2 Differential Privacy

Definition 1 (ε-differential privacy [1]): If any output $\mathcal{O}(\mathcal{O} \subseteq Range(\mathcal{K}))$ of a randomized algorithm $\mathcal{K}$ on any two data sets D_1 and D_2 that are identical except for one record satisfies Eq. (1):

$$\Pr[\mathcal{K}(D_1) \in \mathcal{O}] \leq e^{\varepsilon} \cdot \Pr[\mathcal{K}(D_2) \in \mathcal{O}] \tag{1}$$

then algorithm $\mathcal{K}$ satisfies ε-differential privacy. Where ε is the privacy budget, the smaller ε is, the higher the strength of privacy protection provided, and vice versa.

Definition 2 (w - event differential privacy [2]): A randomized mechanism $\mathcal{K}$ satisfies w - event privacy if for any pair of w - neighboring data stream prefixes S_t, S_t' and all set $S \subseteq Range(\mathcal{K})$ and all t, it holds that,

$$\Pr[\mathcal{K}(S_t) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{K}(S_t\prime) \in S] \tag{2}$$

in order for mechanism $\mathcal{K}$ to satisfies w− event privacy, it requires that the sum of the privacy budget within the sliding window of length w at any timestamp is not larger than the total privacy budget ε. That is, $\forall i \in [t]$, $\sum_{v=i-\mathrm{w}+1}^{i} \varepsilon_v \leq \varepsilon$.

3 Differential Privacy Protection Mechanism for Mobile Crowdsensing Location Aggregation Data Release

3.1 Re-LDCR Overall Architecture

Re-LDCR's composition is shown in Fig. 1. The arrows indicate the data transfer process between the corresponding modules, which means that the output of the current mechanism has an effect on the mechanism indicated by the arrow.

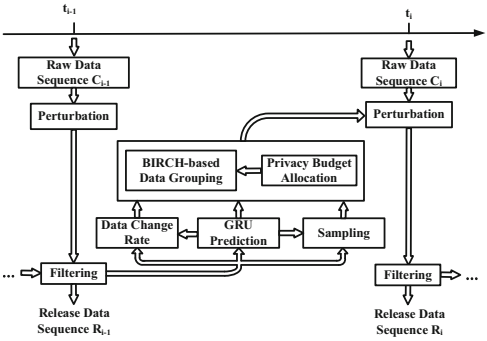


Fig. 1. The overview of Re-LDCR.

After the release of privacy protection data at t_{i-1}, to perceive the subsequent data changes at timestamp t in advance, the GRU mechanism predicts the regional aggregation values at t based on historical released data. Then, for each geographic area j, the data change rate can be calculated by Eq. (3),

$$v_i^j = \begin{cases} \frac{|\arctan k_2|}{90^\circ}, k_1 = 0 \\ \frac{|\arctan k_2|}{|\arctan k_1| + |\arctan k_2|}, k_1 \neq 0 \end{cases} \tag{3}$$

where $k_1 = \left(r_{i-1}^j - r_{i-2}^j\right) \Big/ [(i-1) - (i-2)]$,$k_2 = \left(p_i^j - r_{i-1}^j\right) \Big/ [i - (i-1)]$,which are the slopes of historical release data r_{i-1}^j, r_{i-2}^j and GRU predicted output data p_i^j between two adjacent time periods $[i-2, i-1]$ and $[i-1, i]$, respectively.

Next, the sampling mechanism obtains the set of aggregated data that needs to be released at timestamp t by adaptively adjusting the sampling interval using a proportional integral differential (PID) controller. The PID controller uses the difference between the predicted value and the previous moment's output as a feedback error to adjust the PID error, and after that, the sampling interval is jointly adjusted according to the error and the remaining privacy budget. Non-sampled data are approximate with the released value at the previous sampling timestamp, which can reduce the introduction of noise. The sampling process is as follows:

$$E_{k_l} = \left| p_i^j - r_{k_l}^j \right| \tag{4}$$

$$\delta^j = G_P \cdot E_{k_l}^j + G_I \cdot \left(\sum\nolimits_{o=l-\Phi-1}^{l} E_{k_o}^j \Big/ \Phi \right) + G_D \cdot \left(E_{k_l}^j \Big/ k_l - k_{l-1} \right) \tag{5}$$

$$T = \max\left\{ 1, T_l + \theta\left(1 - \left(\varepsilon_{ri}^j \cdot \delta_i^j \right)^2 \right) \right\} \tag{6}$$

For a specific geographic area, allocate the corresponding privacy budget based on its data change rate and the defined recycling factor. After that, group the sampled data with similar properties by BIRCH clustering. The perturbation mechanism will perform differential privacy on the sum of data according to the minimum budget in the group, and assign noise perturbation values to each sampled data based on the proportion of predicted data to the total predicted data within the group. Finally, since any operations on the post-perturbation data without involving the original data conditions do not affect the degree of differential privacy protection, thus, refer to [6], to improve the utility of the final released data, the noisy data is optimized using the Kalman filter.

Cause the GRU prediction mechanism, the change rate calculation mechanism, the sampling mechanism, and the perturbation filtering mechanism of Re-LDCR have already been described in detail in our previous work [7], in the following, we focus on the privacy budget allocation mechanism and the grouping mechanism.

3.2 Privacy Budget Allocation Mechanism

For measuring the difficulty of recycling privacy budgets in the sliding window to raise the level of caution in the application of privacy budget, we defined the recycle factor *recycle* as the number of time intervals between the start of the sliding window at the current timestamp t_{ws} and the first sampling point in the window t_{fs} then plus one, which means that the privacy budget can be recycled after *recycle* timestamps.

$$recycle = t_{fs} - t_{ws} + 1 \tag{7}$$

Algorithm 1 shows the privacy budget allocation process for sampling geographic area j. First, the current remaining budget ε_{ri}^j is calculated and then obtain *recycle*. In line 2, v_i^j is the expected allocation ratio based on the data change rate, and we limit it based on *recycle* and sampling interval to prevent serious damage to the subsequent data utility. Afterward, assign the remaining budget based on the calculated ratio, where $\tau_{\min}$ and $\tau_{\max}$ are used as lower and upper limits for the allocation ratio, $\tau_{\min}$ is used to

avoid severe distortion caused by too little privacy budget, and $\tau_{\max}$ is used to reserve sufficient available privacy budget for subsequent releases. Finally, the allocation budget ε_i^j will not be greater than $\varepsilon_{\max}$, cause the data utility will not be significantly improved even has the privacy budget exceeds $\varepsilon_{\max}$.

Algorithm 1 Privacy Budget Allocation

Input: window size w, total privacy budget ε, data change rate shreshold v, Minimum allocation ratio $\tau_{\min}$, maximum allocation ratio $\tau_{\max}$, maximum allocation budget value $\varepsilon_{\max}$, change rate of sampled j at i, i.e. v_i^j, new sampling interval T_i^j, and the privacy budget at prior timestamp $\varepsilon_j = \{\varepsilon_1^j, \varepsilon_2^j, ..., \varepsilon_{i-1}^j\}$

Output: pivacy budget for sampling j at timestamp i

1 Calculating the remaining budget $\varepsilon_{ri}^j = \varepsilon - \sum_{l=i-w+1}^{i-1} \varepsilon_l^j$, and obtain *recycle*

2 Calculate the allocation proportion:

if $v_i^j \le v$, then $\tau_{allocated} = v_i^j$

else, $\tau_{allocated} = v_i^j \cdot \left(1 - \frac{recycle}{\alpha T_i^j + recycle}\right)$

3 Remaining privacy budget allocation: $\varepsilon_{allocation}^j = \min(\max(\tau_{\min}, \tau_{allocation}), \tau_{\max}) \cdot \varepsilon_{ri}^j$

4 Privacy budget for sample j: $\varepsilon_i^j = \min\{\varepsilon_{allocation}^j, \varepsilon_{\max}\}$

3.3 BIRCH-Based Data Grouping Mechanism

Unlike existing solutions that only combine sampled data with values smaller than a preset anti-noise threshold, the grouping mechanism we have designed groups all sampled data to make use of large values to improve the overall noise resistance of the grouped data. Meanwhile, the grouping takes into account the data change rate and the remaining budget to achieve a similar level of privacy protection for data with similar characteristics. Since the MCS platform needs to process large-scale data, the sparsity and density of data will vary greatly due to different factors such as collection scenarios, ranges, and times. Therefore, a data clustering method that can adapt to large-scale data processing is needed to achieve adaptive grouping of sampled data.

Among existing clustering algorithms, the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm [8] has advantages such as memory saving, fast speed, noise point recognition, and low computational complexity, moreover, the number of clustering categories does not need to be specified during the clustering process, which is consistent with the requirements of grouping sampled data in MCS.

The grouping process of Re-LDCR is mainly according to the data change rate and remaining privacy budget, using the BIRCH clustering algorithm. Firstly, at i, the sampled data set $\left[..., c_i^j, ...\right]$ is obtained through the sampling mechanism. At timestamp i, for each data point, get its data change rate v_i^j and the privacy budget remaining at the next time stamp after privacy budget allocation $\varepsilon_r^j = \varepsilon - \sum_{l=i-w+2}^{i-1} \varepsilon_l - \varepsilon_i^j$, forming the set of sample points $\left\{..., \left(v_i^j, \varepsilon_r^j\right), ...\right\}$. Then, BIRCH clustering is applied to this set to obtain the corresponding grouping strategy. Re-LDCR's privacy budget allocation is mainly based on data change rate, and more privacy budget is allocated to data with a

faster change rate to protect the data change trend. At this time, if the remaining privacy budget at the next moment is too small, then the current release occupies more privacy budget which will lead to excessive noise introduced by the perturbation process at subsequent time points, causing the data utility to deteriorate. By grouping data with similar rates of change and similar remaining budgets, and taking into account both the current privacy protection needs and the utility of subsequent releases, the grouped data can achieve improved noise immunity while balancing data availability and privacy protection.

4 Experiments

4.1 Setting

We conducted simulation experiments on Beijing taxi trajectory dataset T-driver[9]. Each location record contains four pieces of information: taxi ID, sampling time, longitude, and latitude. The areas in the dataset with longitude in $[116.26°E, 116.51°E]$ and latitude in $[39.79°N, 40.04°N]$ were divided into three maps, each map then is evenly divided into geographic areas with different ranges, and we release taxi aggregation data for each geographical area at 15-min intervals. Table 1 shows details of the map divisions.

Table 1. Map information

T-driver	Single geographical area range(km^2)
Map1	1.09×0.85
Map2	2.73×2.13
Map3	3.89×3.04

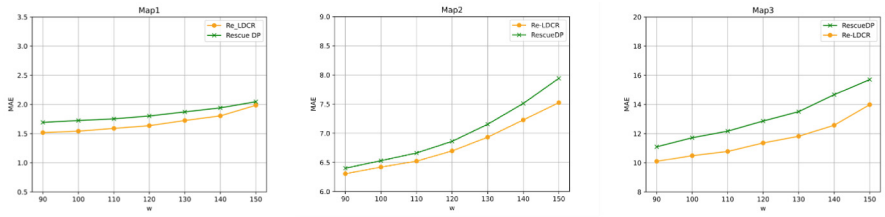
In experiments, we measure privacy-preserving data release utility in terms of mean absolute error (MAE) and compare Re-LDCR with the existing differential privacy data release scheme RescueDP. The MAE between the original aggregated data series c for m geographical regions and the privacy-preserving release data series r is defined as:

$$MAE(c, r) = \frac{1}{m} \sum_{j=1}^{m} |c_j - r_j| \tag{8}$$

4.2 Results and Analysis

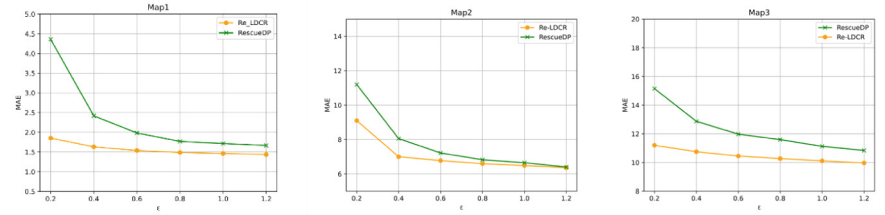
In w-event Differential privacy, the longer the sliding window is, the better the privacy protection effect is when the overall privacy budget is fixed. At first, we fix the total privacy budget to 1 and compare Re-LDCR with RescueDP on different maps as the sliding window length increases from 90 to 150, the results are shown in Fig. 2. It can be observed that as the sliding window length increases, the MAE of both schemes shows an

upward trend. The reason is that the larger the window, the more release points it contains, and the less privacy budget is allocated to each sampled point, thus causing an increase in error. In addition, Re-LDCR outperforms RescueDP, indicating that Re-LDCR is less sensitive to changes than RescueDP. The reason is that the privacy budget allocation mechanism we designed restricts the proportion of the current budget allocated when it is difficult to recycle the budget, so that the available privacy budget for releasing data at a later point in time is not too small, thus ensuring data utility.



(a)Map1 (b)Map2 (c)Map3

Fig. 2. Comparison of MAE as w changes



(a)Map1 (b)Map2 (c)Map3

Fig. 3. Comparison of MAE as ε changes

In Differential privacy, a smaller privacy budget means a better privacy protection effect. Figure 3 shows the comparison results of data utility between Re-LDCR and RescueDP under changes in the privacy budget. As can be seen from Fig. 3, the average error of both methods decreases as the total privacy budget increases, mainly because smaller privacy budgets introduce more noise, which affects data utility more. Compared with RescueDP, Re-LDCR has a smaller average error, which proves that Re-LDCR can maintain better data availability while achieving privacy protection. The reason for this is that the Re-LDCR's data grouping process makes full use of the large value data to improve the overall noise resistance of the group, and merges all the sampled data in groups according to the changing characteristics of the data itself and the subsequent privacy protection capability, so as to have better noise resistance during the data perturbation step.

5 Conclusion

In this paper, we researched the privacy-preserving location aggregation data publishing problem in MCS scenarios and proposed Re-LDCR, which could provide w - event level differential privacy protection. In Re-LDCR, we leverage the data change rate and the defined recycle factor to allocate the privacy budget, thereby improving the rationality of the privacy budget application. We then apply BIRCH clustering to group data according to their own change characteristics and remaining budget situation, thus improving the flexibility of the grouping mechanism. In addition, we verify the effectiveness of Re-LDCR through simulation experiments as well as comparisons with existing solutions. The results show that Re-LDCR outperforms existing solutions and can maintain high data usability during the privacy protection process.

Acknowledgments. This work was partially supported by the Innovation Project of GUET Graduate Education (No. 2022YCXS048), Project (No. LJ20212C03128), National Natural Science Foundation of China (Nos. 62161006, 61861003, 61662018), Guangxi Natural Science Foundation of China (No. 2018GXNSFAA050028), Director Fund project of Key Laboratory of Cognitive Radio and Information Processing of Ministry of Education (Nos. CRKL190102, CRKL210205), State Key Laboratory of Integrated Services Networks (No. ISN22–10), Gaming Strategies in Privacy-Preserving Data Aggregation Optimization (No. C22KYS00RX08).

References

1. C. Dwork. Differential privacy. In: 33rd International Colloquium on Automata, Languages and Programming, pp. 1–12. (2006)
2. Georgios, K., Stavros, P., Xiaokui, X., et al.: Differentially private event sequences over infinite streams. Proc. VLDB Endow. **7**(12), 1155–1166 (2014)
3. Q. Wang, Y. Zhang, X. Lu, et al.: RescueDP: Real-time Spatio-temporal Crowd-sourced Data Publishing with Differential Privacy. In: IEEE 35th Annu. IEEE Int. Conf. Comput. Commun. pp. 1–9. (2016)
4. Wang, Q., Zhang, Y., Lu, X., et al.: Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. IEEE Trans. Dependable Secure Comput. **15**(4), 591–606 (2018)
5. Huo, Y., Yong, C., Lu, Y.: Re-ADP: real-time data aggregation with adaptive ω-event differential privacy for fog computing. Wirel. Commun. Mob. Comput. **2018**, 1–13 (2018)
6. Fan, L., Xiong, L.: An adaptive approach to real-time aggregate monitoring with differential privacy. IEEE Trans. Knowl. Data Eng. **26**(9), 2094–2106 (2014)
7. L. Mo, X. Deng, M. Ye, et al.: Continuous Release of Location Data Based on Differential Privacy. In: 2022 21st International Symposium on Communications and Information Technologies (ISCIT), pp. 1–6. (2022)
8. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec. **25**(2), 103–114 (1996)
9. J. Yuan, Y. Zheng, C. Zhang, et al.: T-Drive: Driving directions based on taxi trajectories. In: 18th Annual ACM International Conference on Advances inGeographic Information Systems, pp. 99–108. (2010)