DATA PUBLICATION BASED ON DIFFERENTIAL PRIVACY IN V2G NETWORK

Xiong LIU and Haiqing LIU

ABSTRACT

Nowadays, the massive amount of data from V2G networks makes many important data mining applications possible and improves the quality of our daily lives. However, data privacy protection in the V2G network has also attracted great attention when data is released to a third party for analysis. In response to this problem, this paper proposes a dynamic data publishing algorithm based on differential privacy. Combined with the idea of data prediction, we propose a combined prediction model to predict the data, adaptively sample according to the prediction error, and then dynamically allocate a privacy budget to the sampling points. , And finally release streaming data that meets w-event-level privacy. We experiment on real data sets. Compared with traditional input data set noise injection methods, the proposed algorithm improves data availability and reduces publishing errors

Index Terms: *differential privacy, privacy protection, Data release, V2G network.*

Reference to this paper should be made as follows:

Xiong LIU and Haiqing LIU, (2021), "DATA PUBLICATION BASED ON DIFFERENTIAL PRIVACY IN V2G NETWORK" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 45-53, DOI 10.30696/IJEEA.IX.I.2021.45-53.

Biographical notes:

***Xiong Liu**, male, born in Xingtai, Hebei in 1997. Now he is a master's student in the School of Control and Computer, North China Electric Power University. The main research direction is data privacy protection. E-mail:1164436838@qq.com;*

***Haiqing Liu**, currently associate professor at the School of Control and Computer Engineering, North China Electric Power University*

1. INTRODUCTION

With the rapid development of new energy technologies, electric vehicles have gradually become popular. Due to the limitations of on-board battery technology, EVs need to frequently visit charging piles for charging. A large amount of data stream information with user privacy is continuously collected and aggregated. This data can be used by third-party scientific research institutions and enterprises to provide data services (charging service providers provide users with personalized services and guide electric vehicle load forecasting. Power grid dispatching work, etc.), data owners need to publish and share part of the data they own to third parties. If the data owner directly releases the data, it may seriously damage the privacy of the user, because the data information involves a lot of sensitive information [1].

Differential privacy, as a de facto privacy paradigm, requires the output of data query results to be roughly the same, so that the adversary cannot judge whether a single tuple in the database has been modified. It has a strong theoretical proof that it has been widely used to protect data stream release User privacy in [2]. Specifically, these studies can be further divided into two paradigms: event-level privacy and user-level privacy. The former protects any single time at a specific time stamp, while the latter protects all events of any individual at the entire time stamp. Literature [3] combines event-level differential privacy and user-level differential privacy, and proposes a new concept of w-event privacy on infinite streams. w-event privacy, which protects any sequence of events that occur within any window of the w timestamp. For w=1, w-event privacy is equivalent to event-level privacy, and when w is infinite, it is equivalent to user-level privacy.

Based on the above research, literature [4] proposes that the protection of discharge privacy is an important work that needs to be studied in the future. At the same time, how to ensure the online disclosure of the charging and discharging data of electric vehicle users, electronic bills, etc. is also a problem to be solved. However, no specific solutions and methods have been given. Literature [5] proposed a new differential privacy real-time aggregation statistical framework FAST based on filtering and adaptive sampling. The framework uses differential privacy technology to publish sampling point data, and predicts and publishes non-sampling point data, which reduces the cost of privacy publishing. Wang et al. [6] proposed the RescueDP scheme. By dynamically grouping related dimensions, while realizing w-event privacy, it greatly improves the utility of the published multi-dimensional data. Zhang et al. [7] proposed a streaming histogram publishing algorithm SHP based on sliding window segmentation. In the adaptive sampling mechanism, the SHP algorithm is based on the current sliding window and relies on an adaptive sampling method for the next moment. Counting can save the overall privacy budget. Li et al. [8] combined autoregressive integrated moving average model and dynamic sliding window counting, and proposed a data stream differential privacy histogram publishing algorithm ASDP-HPA, which improves memory utilization. Literature [9] proposed a data self-adaptive multi-dimensional stream real-time publishing framework, AdaPub, which can realize w event differential privacy, and at the same time effectively improve data practicability by using the spatial and temporal correlations in multi-dimensional data streams. Literature [10] uses dynamic sampling to process data for vehicle trajectory information, uses integrated Kalman filtering of transition probability matrix in prediction calculations, and constructs a privacy budget allocation method based on regional privacy weights to ensure data privacy and practicability. Literature [11] defines a new concept to measure the utility of published data, and proposes an adaptive w-event differential privacy model based on this, but it does not consider non-sampling point errors. Based on the existing research work, we propose a dynamic data publishing algorithm based on differential privacy for the data publishing problem in the v2g network. Specifically, we propose a data prediction model to improve the accuracy of data prediction, and sample the data according to the prediction results. And considering the actual situation such as the peak period and trough period of the charging within a day, the data is adaptively allocated to the budget, which saves the overall privacy budget and improves the availability of data.

2. PRELIMINARIES AND PROBLEM STATEMENT

2.1 PRELIMINARIES

Differential privacy

Differential privacy given adjacent data sets D and D', they differ by at most one record from each other. Given a privacy algorithm A, Rang(A) is the value range of A, if algorithm A outputs the result range S on the data sets D and D' arbitrarily, where S belongs to Rang(A) and satisfies the following inequality, then algorithm A satisfies ε-differential privacy, and its privacy can be measured by the parameter ε:

$$\Pr[\mathrm{A}(\mathrm{D}) = \mathrm{O}] \leq \exp\,(\varepsilon) \times \Pr\,[\mathrm{A}(\mathrm{D}') = \mathrm{O}] \quad (1)$$

In the inequality, Pr[] controls the randomness of algorithm A, and the parameter ε is used to control the degree of privacy protection. From the inequality: the smaller the value of the parameter ε, the probability of A(D)=O and A(D')=O The closer the value is, the higher the degree of privacy protection of Algorithm A is.

The global sensitivity defines the adjacent data sets D and D'. For any query function$f: D \to R^d$, the global sensitivity of the query function is:

$$\mathrm{GS} = \max\,(D, D')\|f(D) - f(D')\|_1 \quad (2)$$

Among them, R represents the real number space mapped by the function, and d represents the query dimension of the function f.

Laplace mechanism, for any function$f: D \to R^d$, if the output result of algorithm A satisfies the following equation, then A satisfies ε-differential privacy:

$$\hat{f}(D) = f(D) + Laplace(\frac{GS}{\varepsilon}) \quad (3)$$

Among them, $Laplace(\frac{GS}{\varepsilon})$ represents the added Laplace noise.

2.2 PROBLEM STATEMENT

V2G network is an important application of the smart grid power distribution terminal. When users exchange information with the power grid in both directions, their private information may be exposed to unauthorized persons. Therefore, the privacy protection of user data is very important. As shown in the [Figure-1], the data publishing model in the V2G network mainly includes two parts: The green box is the collection and storage stage of electric vehicle charging and discharging information, and the red box is the data release stage. In the data collection and storage stage, it starts from BVs and ends with the central server. First, BVs charge and discharge at the corresponding charging station to generate information interaction, which contains user privacy information (account ID, charging data, etc.), and then the charging station will respond accordingly. The data information transmission center server report. According to the charging information, the server creates a database D over time. The goal is to continuously release user charging data under the circumstance of w event privacy guarantee. In the data release stage, the data is released to a third party to provide some value-added services. Considering that the data contains user privacy, the trusted server applies an appropriate privacy protection mechanism instead of the true value of the data. The focus of this paper is to design a data release program to protect the data set while preventing the leakage of sensitive information while ensuring the availability of the data set.

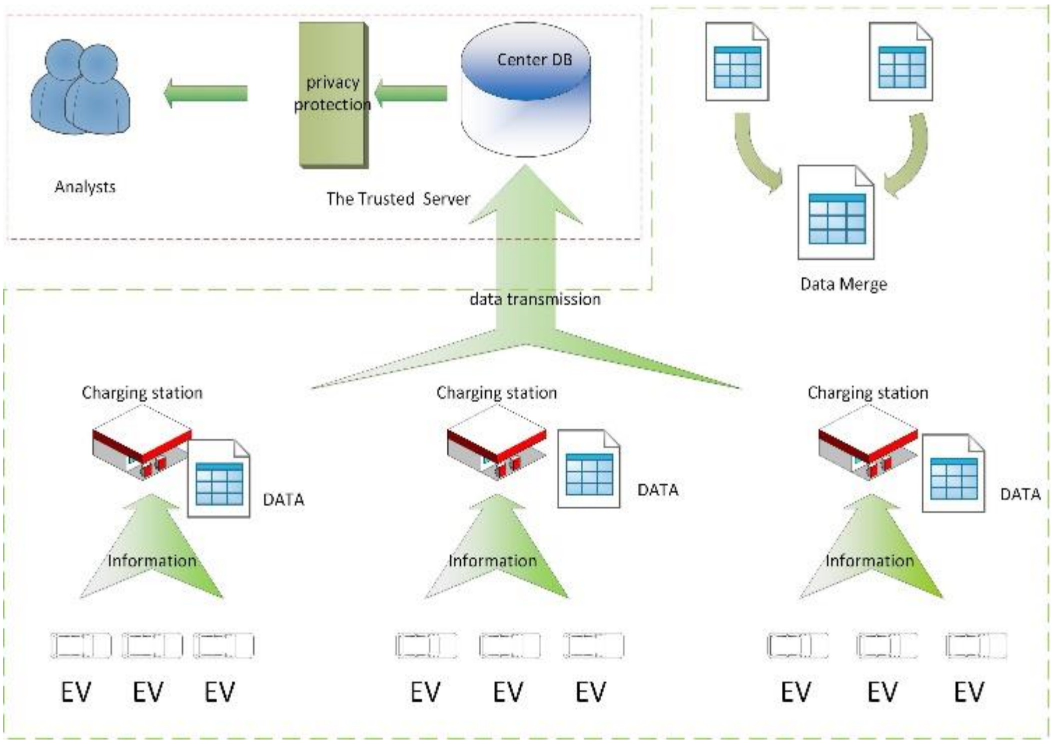

Figure: 1. System model

3 METHOD FRAMEWORK

On the basis of the above research, we propose a charging data release method based on differential privacy. For the data flowing into the window, we use the combined prediction model we proposed to predict the data value, and perform adaptive sampling according to the prediction error. In the window, for a reasonable allocation of privacy budget, we set the window to be variable according to the data dynamics and then use the window size Allocate a privacy budget, and output the released data after adding noise.

In this paper, an adaptive sampling algorithm is proposed by combining Kalman filter and ARIMA combined prediction model. Kalman filter is a linear filtering and prediction method, it is mainly divided into two steps, prediction and correction. Prediction is to estimate the current state based on the state of the previous moment, while correction is to integrate the estimated state and predicted state at the current moment to estimate the optimal state. The process of prediction and correction is as follows:

$$x_k^1 = x_{k-1} \ (4)$$

Where x_k^1 is the predicted value, and x_{k-1} is the corrected value at the previous moment. According to the literature [12], the observed value can be approximated by Gaussian noise:

$$z_k = \tilde{x}_k + v \ (5)$$

$$P_k = E[(\tilde{x}_k - x_k^1)(\tilde{x}_k - x_k^1)^T] \ (6)$$

$$K_k = P_k(P_k + R)^{-1} \ (7)$$

$$x_k = x_k^1 + K_k(z_k - x_k) \ (8)$$

$$P_{k+1} = (1 - K_k)P_k + Q \ (9)$$

Formula 7 is the Kalman gain calculation, Formula 8 is the state correction, that is, the final result of Kalman output, and Formula 9 is the error matrix update.

The differential autoregressive moving average model is a commonly used time series analysis model, also called ARIMA. First judge the smoothness of the numerical sequence to determine the difference d, and then determine the autoregressive coefficient p and the moving average coefficient q through the autocorrelation coefficient ACF and part of the autocorrelation coefficient PACF.

$$\hat{y}_t = u + \phi_1 * y_{t-1} + \cdots + \phi_p * y_{t-p} + \theta_1 * e_{t-1} + \cdots + \theta_q * e_{t-q} \ (10)$$

Among them, ϕ represents the coefficient of AR, and θ represents the coefficient of MA.

The concept of dynamic weight is introduced here. The so-called dynamic weight refers to the continuous adjustment of the combined model weight with the change of the prediction error, so that the prediction output with the best accuracy can dominate the total result, and the error formula is defined:

$$w_i(t) = 1 - \frac{|\tilde{x}_k - \hat{y}_i(\mathrm{t})|}{\sum_i^n e_{ar,i}(t)} \ (11)$$

The final prediction result:

$$S_t = w_1(t)\hat{y}_1(\mathrm{t}) + w_2(t)\hat{y}_2(\mathrm{t}) \ (12)$$

The combined prediction model is used to sample the data stream, and the time stamp of whether t is sampled or not is calculated according to the difference with the previous publishing node. If it is not the sampling point, directly publish the forecast data. If the sampling point allocates a privacy budget based on differential privacy requirements, the conditions are as follows:

$$|S_i - d_t| < 1/\epsilon_i \ (13)$$

If the formula is satisfied, it is a non-sampling point, otherwise the sampling interval is updated to i-t, and a privacy budget is allocated.

In order to reflect this data change and reduce the release error, we use an adaptive window to allocate a privacy budget. When the data is flat, there are relatively few sampling points.

Specifically, we adjust the window size according to the PID mechanism:

$$E_i = \left|\frac{n_i}{M} - \frac{n}{N}\right| \ (14)$$

Among them, $\frac{n}{N}$ is the defined optimal sampling frequency, M is the number of timestamps in the current time period, and n_i is the number of sampling points in the M timestamp. For the convenience of calculation, we calculate M according to the number of sampling points to obtain the current Sampling frequency.

$$\delta_i = C_p \times E_i + C_i \times \sum_{o=n-\pi-1}^{n} E_o + C_d \times \frac{E_i - E_j}{t_i - t_j} \ (15)$$

Use PID error to define the window:

$$W_t = \mathrm{N} + \mathrm{sign}\left(\frac{n}{N} - \frac{n_i}{M}\right) \times \theta \times \delta_i \ (16)$$

Here sign(.) is a sign function, which controls whether the window is enlarged or reduced, and the parameter θ determines the degree of influence of the PID error. Below is our budget allocation algorithm.

```
Algorithm 1.The proposed algorithm
Input: data stream D, Sliding window $w_i$,  Privacy budget ε,  the release data at last
        sampling time stamp $d_t$
Output: The released charging data $\widetilde{D}$
    Predict the data at each timestamp  $S_i \leftarrow pre(D)$
    Determine whether the current timestamp is a sampling point based on the release
    data at last sampling time stamp $d_t$
    if $|S_i - d_t| > 1/\varepsilon_i$  then
      Set i as the sampling point, update sampling interval I = i − t
      if $t_i$ not included in the window  $W_t$
          Calculate the number of sampling points $n_i$
          Calculate the current sampling frequency $\frac{n_i}{M}$ and update $W_t$
      Calculate the remaining privacy budget $\varepsilon_r = \varepsilon - \sum_{k=i-w+1}^{i-1} \varepsilon_k$
      Calculate the allocated privacy budget $\varepsilon_i = \varepsilon_r \times (1 - \frac{I}{w}) \cdot \ln(I+1)$
      $\tilde{y}_i = \tilde{x}_i + Lap(\frac{\Delta f}{\varepsilon_i})$
      Release its Perturbation value $\hat{y}_i$
    Else i is not a sampling point, release predictive value $S_i$
    Return $\hat{y}_i \cup S_i = \widetilde{D}$
```

4 EXPERIMENT

The hardware environment of the experiment is A10-7300 Radeon R6, 10 Compute Cores 4C+6G 1.90GHz, 8GB memory, and the software environment is Windows 10 operating system, implemented by java. This article uses the average relative error and the average absolute error:

$$\mathrm{MAE} = \sum_{r \in R} \frac{|D(r) - D'(r)|}{|R|} \quad (17)$$

$$\mathrm{MRE} = \sum_{r \in R} \frac{|D(r) - D'(r)|}{|R| * D(r)} \quad (18)$$

Among them, $D(r)$ is the original data stream, $D'(r)$ is the data stream after interference, and R is the set of all timestamps.

Because the method similar to ours is ADP [11], we compare it with the ADP method in the experiment. In the experiment, we set a different privacy budget ε and compare the average relative error and absolute error of the two methods. As shown in the figure below, privacy The budget ε ranges from 0.1 to 1. From the results, we can see that as the privacy budget increases, the average error and absolute error gradually become smaller. This is because as the privacy budget increases, the amount of noise added according to the Laplace mechanism decreases and the corresponding error decreases, and the usability increases.

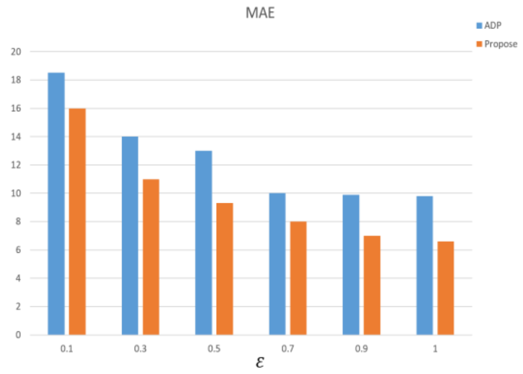

Figure: 2. MAE under different privacy budgets

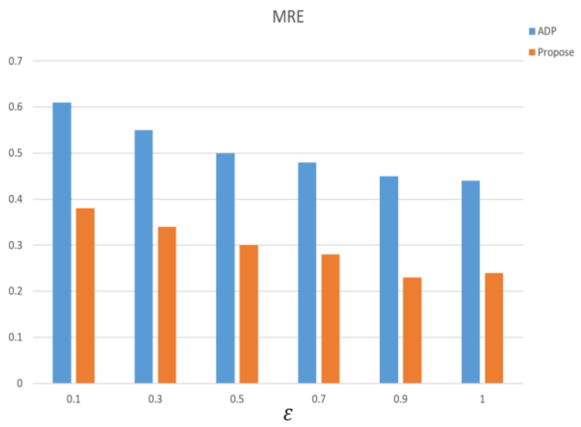

Figure:3 . MRE under different privacy budgets

Comparison of sampling algorithms. In order to compare the sampling performance of our proposed prediction model, we replaced the PID sampling method in [11] with our data prediction sampling, and compared the published results error.

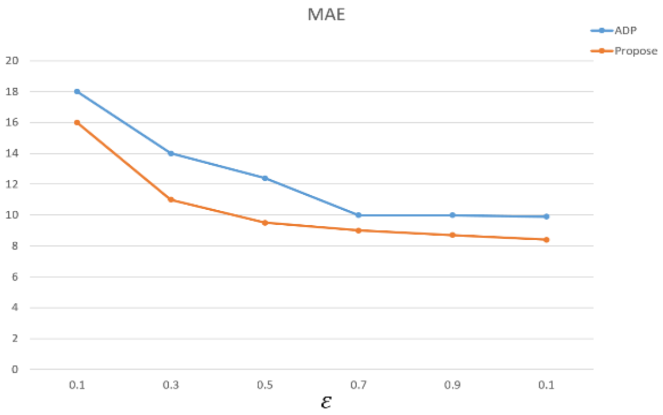

Figure: 4. MAE under sampling algorithm

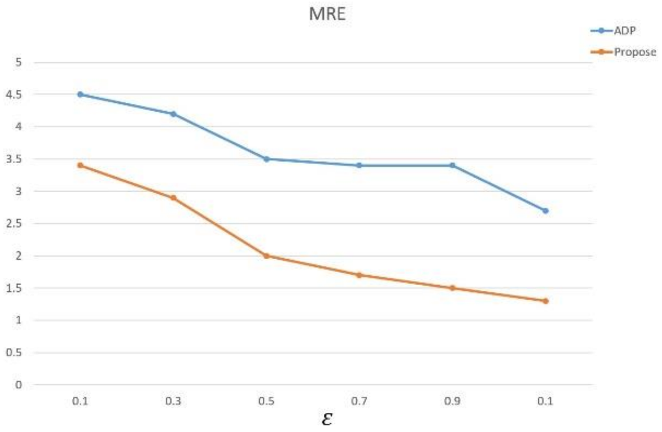

Figure:5 . MRE under sampling algorithm

Because the sampling result of the sampling algorithm directly affects the budget allocation, it can be seen from the utility of the data release that the performance of our sampling algorithm is better than ADP.

5 CONCLUSION

In view of the large amount of sensitive data that needs to be released in the V2G network, we have studied the differential privacy data release algorithm. In this paper, we propose a combined data prediction model, and then sample the data based on the data prediction, and consider the dynamic changes of the data to be more adaptable to reality Scenes. Finally, the experiment shows that our method is superior to the existing methods, and it improves the utility of the data under the protection of privacy.

REFERENCES

[1] Langer L , Skopik F , Kienesberger G , et al. Privacy issues of smart e-mobility[C]// Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE. IEEE, 2013.

[2] Dwork C , Mcsherry F , Nissim K , et al. Calibrating Noise to Sensitivity in Private Data Analysis[J]. Proceedings of the VLDB Endowment, 2006.

[3] Kellaris G , Papadopoulos S , Xiao X , et al. Differentially Private Event Sequences over Infinite Streams[J]. Proceedings of the VLDB Endowment, 2014.

[4] Han W , Xiao Y . Privacy preservation for V2G networks in smart grid: A survey[J]. Computer Communications, 2016, 91-92(oct.1):17-28.

[5] Fan L , Xiong L . An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(9):1-1.

[6] Wang Q , Zhang Y , Lu X , et al. Real-Time and Spatio-Temporal Crowd-Sourced Social Network Data Publishing with Differential Privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4):591-606.

[7] Zhang Xiaojian, Meng Xiaofeng. A streaming histogram publishing method based on differential privacy[J]. Journal of Software, 2016, 27(02): 381-393.

[8] Li Y , Li S . Research on Differential Private Streaming Histogram Publication Algorithm[C]// 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2018

[9] Wang T , Yang X , Ren X , et al. Adaptive Differentially Private Data Stream Publishing in Spatio-temporal Monitoring of IoT[C]// 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE, 2019.

[10] Ma Z , Zhang T , Liu X , et al. Real-Time Privacy-Preserving Data Release Over Vehicle Trajectory[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8):8091-8102.

[11] Yong C , Huo Y , Hu C , et al. A real-time aggregate data publishing scheme with adaptive ω-event differential privacy[J]. Mathematical Foundations of Computing, 2018, 1(3):295-309.

[12] Fan L , Li X . Real-time aggregate monitoring with differential privacy[C]// Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

[13] G. Sahoo, R. K. Tiwari, “Designing an Embedded Algorithm for Data Hiding using Steganographic Technique by File Hybridization”, IJCSNS, Vol. 8, No. 1, pp. 228-233, January 2008.

[14] Kumar Gunjan, R. K. Tiwari and G. Sahoo, “Towards Securing APIs in Cloud Computing” International Journal of Computer Engineering and Applications, Volume 2, Issue 2, 2014.

[15] Abu Salim, R. K. Tiwari and S. Tripathi “ Secure Cloud Environment a Novel Approach”, International Journal of Computer Engineering and Applications Volume XI, Issue IX.

[16] Rajesh Kumar Tiwari “Hybrid database A Steganographic Approach”, International Journal of Computer Engineering and Applications, Volume II, Issue I, 2013.