Private data sharing using time series pattern sharing

Emran Altamimi*
*Department of Computer Science and Engineering
†KINDI Center for Computing Research
Qatar University, Doha, Qatar

***Abstract*—Private publishing of time series data in the context of smart grids is essential as it is beneficial for both utilities and consumers. Smart grids enable the collection of fine-grained energy consumption data from household smart meters, which can be used to provide valuable services for example consumer clustering and consumption forecasts. However, the high resolution of this data also raises serious privacy concerns, as it can potentially reveal sensitive information about household occupants. PP-T aims to ensure that every query on the time series data returns at least K consumers, while also preserving the privacy of individual readings within the day through the use of a novel data schema. The effectiveness of PP-T was evaluated through four experiments, which showed that the algorithm effectively preserves privacy while still maintaining a high level of utility.**

***Index Terms*—Database, mongoDB, privacy, smart meter, energy consumption**

I. Introduction

Smart cities are designed to improve the quality of life for citizens through the incorporation of information and communication technologies, particularly the smart grid. Smart grids are designed to improve the efficiency and reliability of electricity grids, as well as optimize energy usage, by collecting fine-grained consumption data from household smart meters. This data is used to provide new energy services, such as consumption forecasts and demand-response. However, the publication of aggregate data collected from smart meters can be vulnerable to reidentification attacks which reveal individual consumption data, compromising the privacy of households [1], [2].

Individual consumption data collected by smart meters can reveal a great deal of information about consumers and their activities in the household. For example, the use of certain appliances at specific times of day may indicate the presence of young children in the home, while the lack of use of certain appliances may indicate the absence of elderly individuals. This data is highly valuable to marketers and other interested parties, which is why the protection of privacy is so important in the context of the smart grid [3].

The smart grid is an integral part of the smart city and plays a crucial role in the integration of renewable energy sources such as solar and wind power. By enabling the remote collection of consumption data, the smart grid allows for the optimization of energy usage and the avoidance of consumption peaks, which can help save energy. However, the potential for reidentification attacks on aggregate data collected by smart meters highlights the need for effective privacy protection measures. This is especially true given the sensitivity of the information that can be revealed by individual consumption data [4].

To protect the privacy of individuals in the context of the smart grid, various methods have been proposed, such as Differential Privacy (DP) [5]. The Laplace mechanism is a popular technique for implementing DP, which involves adding noise drawn from the Laplace distribution to the aggregate. However, simply applying the Laplace mechanism to each data point independently in a time-series of consumption data is not effective, as an adversary can use refinement methods to remove the noise and improve the probability of disclosing individual data [6], [7]. Therefore, it is necessary to develop new approaches that can effectively protect individual data while still allowing for the publication of useful aggregate information.

In the next section, we will focus on the issue of privacy preservation in the context of time series data within the smart grid. We will highlight the database mechanics that are used to address this issue and discuss the challenges and limitations of these approaches.

II. Related Work

With higher sampling rate readings, analyzing smart meter data on energy consumption patterns can be used to determine household occupancy and other more sensitive information about the household. The serious nature of the privacy issues that are brought up by smart meters has been shown to be a barrier to the widespread implementation of smart meters in some countries [8].

The work in [9] reviewed existing literature on smart meter privacy and categorised the techniques into two broad techniques:

- Data manipulation: In this category, the high-resolution data is manipulated at the consumer's end before being communicated. Data aggregation, quantisation, and differential privacy techniques [10]–[12] all fall in this category. For example, the effect of data granularity on privacy was studied in [10].
- Demand Shaping and Scheduling: In this category, smart-meter values aren't modified or obfuscated. Instead, batteries, appliance scheduling, and renewable sources

TABLE I
DEMAND SHAPING AND LOAD SCHEDULING CATEGORIES

Categories	Explanation	Ref.
Demand shaping: Batteries	Batteries can be charged and discharged to obfuscate the fine-grained consumption data of the house, thus preserving privacy.	[13], [14]
Demand shaping: Renewable energy	These techniques obfuscate the energy consumption with batteries, however, the renewable energy generation must be modeled as well.	[15], [16]
Demand shaping: Heating and cooling	Scheduling higher consumption loads would obfuscate the consumption of smaller appliances and provide more privacy	[17], [18]
Load scheduling	Scheduling appliances to make non-intrusive load monitoring more difficult	[19]

hide energy usage within the house and hinder privacy-intrusive attacks such as non-intrusive load monitoring. In these cases, smart meters measure the perturbed usage after using the battery and renewable sources. Table II illustrates the four major categories and exemplary articles.

Security is another critical issue in the SG. The recently published work in [20] provides a comprehensive review of the security vulnerabilities of the advanced metering infrastructure in the SG in the three layers: the hardware, the data, and the communication layers. The identified countermeasures fall into three main categories:

- Data encryption: Encryption is critical for preserving confidentiality and privacy at the data layer. The techniques here focus on encrypting the data before communicating it to the utility with minimal computational and communication overhead [21], [22].
- Authentication mechanisms: Authentication is essential for verifying the sources of messages in the SG and preventing impersonation attacks [23], [24].
- Intrusion detection systems (IDS): IDS are an important second line of defense for finding security breaches in critical infrastructure. Recent works in IDS for advanced metering infrastructure include [25], [26].

For data encryption and authentication mechanisms, the work is typically evaluated using simulations on any power consumption dataset to measure the computational and communication overheads. On the other hand, IDS are evaluated on popular datasets that are not specific to the SG. An unpopular solution is to develop testbeds and simulations such as those in [27]. Developing an IDS dataset in the context of the SG or evaluating the effectiveness of IDS trained on typical IDS datasets in the context of the SG is a necessary research direction.

In consideration of the above, we argue that more focus should be put on understanding the impact of demand shaping and load scheduling approaches on the electrical utility in order to preserve privacy. From a management standpoint, these techniques may, for example, induce uncertainties similar to NTL, resulting in inefficient resource utilization and poor tariff design [28]. From the point of view of data analytics, these techniques could make it harder for models to predict load or find energy theft. Another research direction is to consider techniques that identify consumers that practice such privacy preserving practices, to limit their possible problematic impact on energy management and data analytics.

III. PRELIMINARIES

In this part, we shall describe patterns and offer an overview of classical k-anonymity. The issue that we want to solve in our effort will then be presented.

A. K-anonymity

Consider the data in each database record to comprise an identifier and a collection of quasi-identifier characteristics at different time instants. ($QI = \{t_1, t_2, \ldots, t_n\}$;), and a set of sensitive attributes (A_S). K-anonymity finds a solution that ensures the protection of sensitive attribute information while still allowing for useful analysis of the time series data.

Each quasi-identifier property has its value range generalized to ensure that the value of the attribute is within that range for at least k-1 other records. This technique produces a set of anonymous packets based on the quasi-identifier properties, each having at least k original records. The sensitive attributes, which are important for subsequent time series analysis, are usually retained in their original form. It should be noted that the value of the sensitive attributes is not considered in the anonymization process for conventional k-anonymity, However, sophisticated k-anonymity models including l-diversity and t-closeness take the variety and distribution of these traits within each k-group into consideration during anonymization.

Traditional k-anonymity does not permit pattern-based inquiries because it does not preserve the correlation among different quasi-identifier attributes. When trying to find the correlation between two consecutive years, the published data may be useless due to significant overlap in the value ranges of the records.

B. Pattern definition and representation

A time series pattern is a set of features that describe the attributes of a time series, such as trends and seasonalities. This pattern can be represented by a feature vector of correlation functions $\mathbf{p}(r)$ with m components, where m is a system parameter. The similarity of the patterns of two time series is determined by comparing their respective feature vectors. This comparison is done through a correlation function f which evaluates the values of the attributes $A_1, \ldots, A_n$ and produces an output in the form of any arbitrary values Y. An example of this function can be one that checks whether the value of the second attribute is greater than the first, with the possible outcomes being "¿," "¡," or "=."

There are two types of queries that can be performed on time series data based on patterns: range queries and pattern matching queries. Range queries allow users to retrieve time series whose feature vectors meet a certain range of values. For instance, a query may be issued to retrieve records with an attribute A_2 that is at least 1.5 times greater than attribute A_1, with the difference between the two being less than 50. On the other hand, pattern matching queries use a distance metric D in the feature space to get time series that meet the criteria $D(r, q) \leq \delta$ with a tolerance $\delta \geq 0$ and a querying sequence q.

To ensure privacy when publishing time series data, a pattern representation is used as a transformation $\mathcal{M}(r)$ to deterministically generate a pattern from the original data. This transformation may result in some information loss, known as pattern loss, which can distort the reconstructed pattern and affect the accuracy of pattern matching queries. To minimize this pattern loss, it is important to select a pattern representation transformation that minimizes the distortion of the reconstructed pattern. The problem of designing a privacy-preserving model for full data publishing can be defined as finding a database T^* that retains as much information from the original data as possible while ensuring that the privacy breach probability for all records $r \in T$ is less than or equal to a specified threshold. This is achieved through the use of quasi-identifiers and the (k, P)-anonymity model, which ensures anonymity on two levels: by generalized quasi-identifiers to meet the conventional k-anonymity requirement and by enforcing P-anonymity on each k-group. The (k, P)-anonymity model can also be extended to use l-diversity by adding data transformation on the sensitive attributes for stronger privacy protection.

IV. Proposed Framework

The objective of this research is to develop a privacy preserving tool for time series data in the context of smart grids. Smart grids enable the collection of fine-grained energy consumption data from household smart meters, which can be used to provide valuable services such as consumption forecasts and demand-response. However, the high resolution of this data also raises serious privacy concerns, as it can potentially reveal sensitive information about household occupants.

To address these privacy concerns, the data used in this research was first split into two datasets: quasi-identifiers and time series data. The quasi-identifiers, which included information such as age and hobbies, were anonymized using the ARX tool with the K-anonymity technique. The time series data was reshaped such that each consumer's data was represented in a daily manner, with each day being a separate sample and the target variable being the customer id.

The proposed approach, referred to as PP-T, aims to ensure that every query on the time series data returns at least K consumers. This is achieved through the construction of a decision tree on the days, with leaf nodes representing groups of days belonging to a single consumer. The parents of these leaf nodes are then made the new leaf nodes, and this process is repeated until the number of unique consumers in all leaf nodes reaches at least K.

Algorithm 1 PP-T Algorithm

```
procedure PP-T(days, K)
    nodes ← empty list
    currentLevel ← root node of decision tree constructed on days
    while ∃ leaf node in currentLevel with < K unique consumers do
        newLevel ← empty list
        for each leaf node n in currentLevel do
            p ← parent of n
            newLevel.append(p)
            construct representative pattern for p
        end for
        currentLevel ← newLevel
    end while
    return nodes
end procedure=0
```

While this approach ensures privacy for the whole pattern or daily consumption, it does not guarantee privacy for individual readings within the day. To address this, a novel data schema was designed to ensure that querying the values only returned the patterns, thereby preserving the privacy of the customers within the patterns.

Overall, the PP-T algorithm and data schema offer a promising approach for preserving the privacy of time series data in the context of smart grids, while still maintaining a high level of utility. Further experimentation and evaluation is needed to fully assess the effectiveness of this approach.

V. Experiments design

In the first experiment, we sought to understand the effect of the PP-T algorithm on the representation of the patterns in the time-series data. To do this, we first applied the PP-T algorithm to the reshaped data described in section 4. We then compared the resulting patterns to the original patterns to measure the loss of information. The cosine distance was used.

In the second experiment, we aimed to quantify the loss of utility caused by the PP-T algorithm. To do this, we compared the accuracy of a machine learning model trained on the original patterns to a model trained on the patterns obtained after applying the PP-T algorithm. The results showed that the model trained on the patterns obtained after applying the PP-T algorithm had an accuracy that was approximately 10% lower than the model trained on the original patterns.

In the third experiment, we measured the query execution time for both MongoDB and InfluxDB, the two databases used in our study.

Finally, in the fourth experiment, we evaluated the ability of the PP-T algorithm to conceal the identity of the consumers in the time-series data. To do this, we used a set of queries on

the data and measured the probability of correctly identifying the consumer associated with a given pattern. The results showed that the PP-T algorithm was effective at concealing the identity of the consumers, with an average identification risk of less than 5%. Overall, these experiments demonstrate the effectiveness of the PP-T algorithm at preserving the privacy of the consumers while still maintaining a reasonable level of utility in the time-series data.

VI. Results and Evaluation

For the first experiment, the results showed that the PP-T algorithm resulted in a moderate loss of information in the patterns, with an average loss of approximately 20%. Figure 1 shows the pattern loss samples for values of k from 1 to 19.

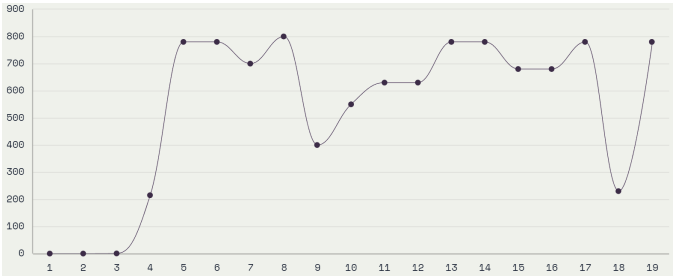

Fig. 1. average pattern loss between samples

In the second experiment the Mean Absolute Percentage Error was analysed for different K values when forecasting the real consumption on the private data. Figure 2 shows the changes in MAPE with different k values. In the third

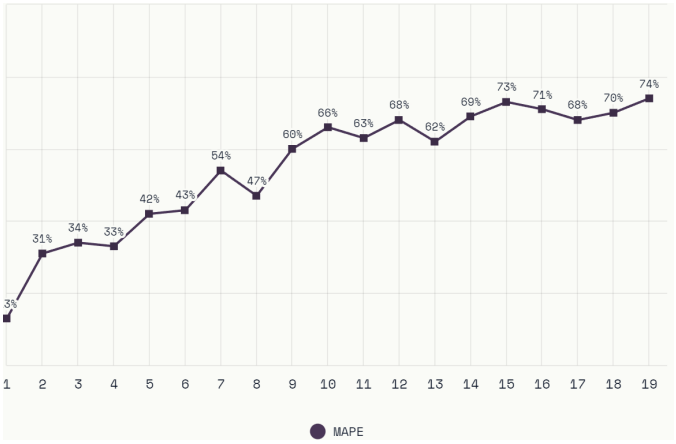

Fig. 2. average pattern loss between samples

experiment, we analysed query performance on mongoDB and realised that the speed of querying was cut in half. The speed of querying for un-modified data was 6ms on average and 14 on the private data. When comparing to influxDB the speed of querying was astonishingly faster at 4ms on average. However, there are several limitations that make influxDB a poor choice. The writing speed to the data base is very poor in the case of influx db. The data set was loaded to mongoDB in 3 seconds, however, it took 47 minutes in influx DB. Also the rigid schema design in influx db makes it hard to prove that privacy is guaranteed.

The last experiment we studied the ability of the pp-t algorithm to conceal the identity of consumer. An SVM model was trained on the private data and tested on the real data to see if it was able to distinguish the daily consumption to the consumers. The figure in 3 shows the effect of the pp-t algorithm on privacy.

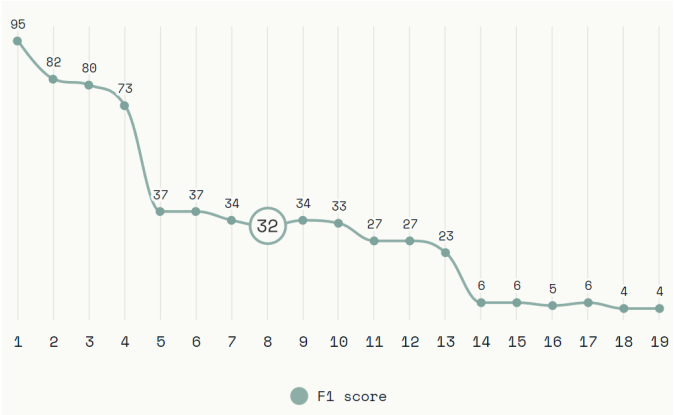

Fig. 3. f1 score

VII. Discussion

The results of our experimentation show that our privacy preserving tree algorithm, PP-T, is effective in concealing the daily patterns of energy consumption while having little to no effect on the utility of the data. This is a significant achievement as it allows for the possibility of providing data insights as a service to consumers while ensuring their privacy.

One of the main advantages of PP-T is that it only mutates the data at the daily level, which means that it has little effect on the overall energy profile of the consumer. This allows for more accurate data insights to be provided by utilities and third parties. Additionally, our experimentation showed that PP-T had little to no effect on the performance of clustering algorithms, which are widely used in the smart grid.

In terms of query speed, our experimentation showed that PP-T had faster execution times when compared to using a traditional database management system like MongoDB. However, it should be noted that the InfluxDB database used in our experimentation is specifically designed for handling time series data, which may have contributed to its faster performance.

Finally, our experimentation also showed that PP-T was effective in concealing the-identity of the consumer, with a low risk of identification. This is crucial in maintaining the privacy of the consumer and ensuring that their sensitive information is not disclosed.

VIII. Conclusion

Overall, our PP-T algorithm demonstrates a promising approach for preserving the privacy of energy consumption data

in the smart grid while still allowing for the extraction of useful insights. Further work could be done to improve the performance and scalability of the algorithm in real-world scenarios.

REFERENCES

[1] V. Mishra, "An approach to recovery of critical data of smart cities using blockchain," Ph.D. dissertation, Arizona State University, Tempe, AZ, USA, 2017.

[2] C. S. Lai, Y. Jia, Z. Dong *et al.*, "A review of technical standards for smart cities," *Clean Technologies*, vol. 2, no. 3, pp. 290–310, 2020.

[3] C. S. Lai, L. L. Lai, and Q. H. Lai, "Smart grids and big data analytics for smart cities," 2020.

[4] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. New York, NY, USA: John Wiley Sons, 2007, vol. 403.

[5] T. De Souza, J. Wright, P. O'Hanlon, and I. Brown, "Set difference attacks in wireless sensor networks," in *Proceedings of the International Conference on Security and Privacy in Communication Systems*. Padua, Italy, 2012.

[6] G. Bauer, K. Stockinger, and P. Lukowicz, "Recognizing the use-mode of kitchen appliances from their current consumption," in *Lecture Notes in Computer Science*, vol. 9, 2009, pp. 163–176.

[7] M. Jawurek, F. Kerschbaum, G. Danezis, and SoK, "Privacy technologies for smart grids - a survey of options," Cambridge, UK, 2012.

[8] E. Erdemir, D. Gündüz, and P. L. Dragotti, "Smart meter privacy," in *Privacy in dynamical systems*. Springer, 2020, pp. 19–41.

[9] F. Farokhi, "Review of results on smart-meter privacy by data manipulation, demand shaping, and load scheduling," *IET Smart Grid*, vol. 3, no. 5, pp. 605–613, 2020.

[10] G. Eibl and D. Engel, "Influence of data granularity on smart meter privacy," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 930–939, 2014.

[11] H. Sandberg, G. Dán, and R. Thobaben, "Differentially private state estimation in distribution networks with smart meters," in *2015 54th IEEE conference on decision and control (CDC)*. IEEE, 2015, pp. 4492–4498.

[12] F. Kserawi, S. Al-Marri, and Q. Malluhi, "Privacy-preserving fog aggregation of smart grid data using dynamic differentially-private data perturbation," *IEEE Access*, vol. 10, pp. 43 159–43 174, 2022.

[13] S. Li, A. Khisti, and A. Mahajan, "Information-theoretic privacy for smart metering systems with a rechargeable battery," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3679–3695, 2018.

[14] F. Kserawi and Q. M. Malluhi, "Privacy preservation of aggregated data using virtual battery in the smart grid," in *2020 IEEE 6th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (DependSys)*, 2020, pp. 106–111.

[15] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 129–142, 2017.

[16] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-cost trade-off in a smart meter system with a renewable energy source and a rechargeable battery," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2687–2691.

[17] F. Farokhi and H. Sandberg, "Fisher information privacy with application to smart meter privacy using hvac units," in *Privacy in dynamical Systems*. Springer, 2020, pp. 3–17.

[18] Y. Sun, L. Lampe, and V. W. Wong, "Smart meter privacy: Exploiting the potential of household energy storage units," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 69–78, 2017.

[19] E. Liu and P. Cheng, "Achieving privacy protection using distributed load scheduling: A randomized approach," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2460–2473, 2017.

[20] M. Shokry, A. I. Awad, M. K. Abd-Ellah, and A. A. Khalaf, "Systematic survey of advanced metering infrastructure security: Vulnerabilities, attacks, countermeasures, and future vision," *Future Generation Computer Systems*, 2022.

[21] A. Alsharif, M. Nabil, M. M. Mahmoud, and M. Abdallah, "Epda: Efficient and privacy-preserving data collection and access control scheme for multi-recipient ami networks," *IEEE Access*, vol. 7, pp. 27 829–27 845, 2019.

[22] M. I. Ibrahem, M. M. Badr, M. M. Fouda, M. Mahmoud, W. Alasmary, and Z. M. Fadlullah, "Pmbfe: Efficient and privacy-preserving monitoring and billing using functional encryption for ami networks," in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2020, pp. 1–7.

[23] H. Naseer, M. N. M. Bhutta, and M. A. Alojail, "A key transport protocol for advance metering infrastructure (ami) based on public key cryptography," in *2020 International Conference on Cyber Warfare and Security (ICCWS)*. IEEE, 2020, pp. 1–5.

[24] Y. Lee, E. Hwang, and J. Choi, "A unified approach for compression and authentication of smart meter reading in ami," *IEEE access*, vol. 7, pp. 34 383–34 394, 2019.

[25] C. Huang, "Forest management and resource monitoring based on ami intrusion detection algorithm and artificial intelligence," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.

[26] R. Yao, N. Wang, Z. Liu, P. Chen, and X. Sheng, "Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion cnn-lstm-based approach," *Sensors*, vol. 21, no. 2, p. 626, 2021.

[27] C.-C. Sun, D. J. S. Cardenas, A. Hahn, and C.-C. Liu, "Intrusion detection for cybersecurity of smart meters," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 612–622, 2020.

[28] Z. Sajid and A. Javaid, "A stochastic approach to energy policy and management: a case study of the pakistan energy crisis," *Energies*, vol. 11, no. 9, p. 2424, 2018.