Towards Guaranteed Privacy in Stream Processing: Differential Privacy for Private Pattern Protection

He Gu
supervised by Vera Goebel and Boris Koldehofe
University of Oslo
Oslo, Norway
heg@ifi.uio.no

ABSTRACT

Sensor data often contain private information that requires proper protection. Most existing privacy-preserving mechanisms (PPMs) for data streams undermine the utility of the entire data stream and limit the performance of data-driven applications. We attempt to break the limitation and establish a new foundation for PPMs by proposing novel pattern-level differential privacy (DP) guarantees and pattern-level PPMs that fulfill pattern-level DP. They operate only on data that correlate with private patterns rather than on the entire data stream, leading to higher data utility. We first describe results for sequence operator based patterns in a centralized system and outline future work to generalize it for other operators and to local solutions.

CCS CONCEPTS

• **Security and privacy** → **Formal security models**; *Database and storage security*.

KEYWORDS

differential privacy, CEP, data streams

ACM Reference Format:
He Gu. 2023. Towards Guaranteed Privacy in Stream Processing: Differential Privacy for Private Pattern Protection. In *The 17th ACM International Conference on Distributed and Event-based Systems (DEBS '23), June 27–30, 2023, Neuchatel, Switzerland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583678.3603284

1 INTRODUCTION

Sensor data analytics and data-driven applications are a force to be reckoned with in the modern world. However, while they become increasingly ubiquitous and reliable, the risk of data misuse also increases. Sensor data in many cases can reveal information that individuals regard as private. Several studies have investigated privacy-preserving mechanisms (PPMs) for sensor data and have proposed reliable privacy guarantees. Most of them protect privacy by operating on raw data tuples in a given data stream [8]. These protections are usually universal for all privacy scenarios and treat all data tuples equally. However, certain data tuples may contain more important or private information than others. Therefore, specific measures targeting them can become more profitable. Consider the example of taxi services. The GPS locations of taxis are required to find nearby waiting passengers and predict nearby traffic jams. However, some taxi passengers want to keep their locations private when traveling to certain sensitive places, such as their homes. Typical state-of-the-art PPMs would add sufficient noise to the GPS records of an entire trip to hide the proximity to these places. It reduces the precision of all GPS records and the utility of all location-based services. Considering that passengers only want to hide information about traveling to sensitive places, it is more efficient to protect only the data that reveal these patterns and to avoid adding noise to all other data. This maintains a higher utility of the entire data stream and enhances the usefulness of all services that use this stream. As such actions only protect specific private patterns, we name them **pattern-level PPMs**.

This PhD project aims to formulate the foundation of pattern-level PPMs by introducing appropriate definitions and pattern-level privacy guarantees based on differential privacy (DP). Given the theoretical foundation, we propose pattern-level PPMs that protect any private pattern against most queries based on data streams. We first design a centralized private architecture in which pattern-level PPMs can be deployed and will also propose a solution for local privacy models. In detail, our work is expected to make the following contributions:

- We will establish the theoretical foundation of pattern-level privacy guarantees, provide complete definitions, and propose a novel pattern-level DP guarantee, which is named pattern-level ϵ-DP (pattern-level DP);
- We will propose a pattern-level PPM that satisfies the pattern-level DP guarantee, which extends the upper performance boundary of privacy-aware data-driven applications from non-pattern-level PPMs;
- We will quantify the advantages of pattern-level PPMs compared to non-pattern-level state-of-the-art PPMs;
- We will design a centralized and a local architecture to deploy our proposed PPMs.

The early-stage work has been presented in the ASTRIDE workshop in conjunction with ICDE 2023 [5]. This paper is organized into seven sections. Section 2 introduces the related work, and in Section 3, we present the problem statement and the research methodology. Section 4 presents the system settings and assumptions. Section 5 formulates the DP guarantee for patterns, and Section 6 proposes the corresponding PPMs. In Section 7, we propose the evaluation, and Section 8 concludes this paper.

DEBS '23, June 27–30, 2023, Neuchatel, Switzerland

ACM ISBN 979-8-4007-0122-1/23/06.
https://doi.org/10.1145/3583678.3603284

2 RELATED WORK

Multiple novel theories and PPMs have been proposed for continuous data observation and data streams. For example, some works [1, 11] focus on adapting traditional PPMs from static databases to infinite dynamic data streams, while others [8] utilize the unique properties of data streams and their applications, e.g., reordering detected events to protect private patterns.

A well-known set of approaches is rebuilding the bases for DP based on data streams[3, 7, 10]. Although these approaches deliver satisfactory results, they rarely emphasize the different characteristics of distinct data streams, which can be utilized to provide dedicated solutions and even superior performance. Landmark privacy[6] makes a move in this direction. It claims that, in reality, not all timestamps and data should be treated equally because some may contain significantly more private or valuable information. The privacy protection of less important information can therefore be reduced to improve data utility. Although it may seem similar to our approach, it does not take into account the connections between different data tuples, which makes it distinct from our work.

3 PROBLEM STATEMENT AND RESEARCH METHODOLOGY

Our research goal is to provide optimal privacy protection under the required data utility. We believe that pattern-level PPMs can reach even superior performance than non-pattern-level PPMs. However, the existing pattern-level PPMs only provide dedicated solutions for limited usage, e.g., limited types of operators, and their theoretical basis is still in its infancy. We, therefore, aim to break the limitation and deliver a universal solution for pattern-level privacy protections.

Our first research objective is to establish a solid foundation for pattern-level PPMs. The main research problem at this stage is the formulation of fundamental definitions, terminologies, and pattern-level privacy guarantees based on DP. We also aim to design optimal pattern-level PPM and quantify its advantage against non-pattern-level solutions. The main challenges during this procedure are the distribution of the privacy budget of pattern-level PPMs, the choice of metrics for evaluation, and the design of practical experiments. We begin with simplified assumptions, i.e., a centralized solution for the sequence operator, and extend the corresponding early-stage outcomes to reach our final objectives and goals. Both formal and empirical approaches are employed during this research procedure.

4 SYSTEM SETTINGS AND ASSUMPTIONS

We propose a centralized privacy model and a local privacy model as our system settings. The centralized system model consists of three components, i.e., data subjects, a trusted Complex Event Processing (CEP) engine, and data consumers. According to the definitions of the General Data Protection Regulation (GDPR) [9], we detail the requirements of these components as follows:

- Data subjects supply data to the CEP engine and expect their privacy to be protected according to their requirements;
- The trusted CEP Engine provides privacy protections to data subjects and delivers the required data to data consumers;
- Data consumers query certain data from the CEP engine. They follow the *honest-but-curious* threat model.

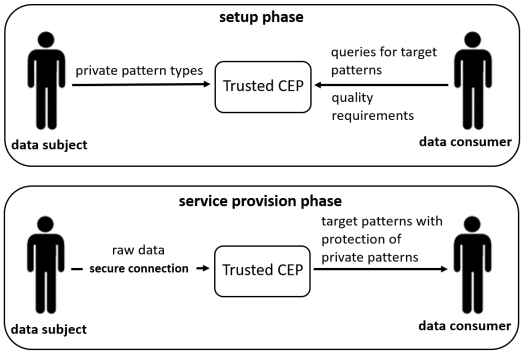

Figure 1: The centralized privacy architecture. The execution consists of a setup phase and a service provision phase.

Figure 1 illustrates the execution procedure of our system model in a centralized privacy model. In the setup phase, data subjects define private patterns, and data consumers define data utility requirements and queries. In the service provision phase, data subjects send raw data to the CEP engine, and data consumers receive responses to their queries with privacy protection from the CEP engine. For the centralized privacy model, we assume that:

- The CEP engine is trusted by data subjects so that it has access to raw data streams, including private data;
- The CEP engine is trusted by data consumers and provides answers to their queries;
- The queries to identify private patterns and target patterns are provided by data subjects and consumers to the CEP engine. The required data utility is defined by the data consumers.

The centralized privacy model requires a trusted CEP engine deployed on an independent server apart from data subjects and consumers. However, providing a separate CEP may not be favorable, considering computing resources and economic revenue. We therefore also aim to propose a local privacy model as a complement in which privacy can be preserved on each data subject's local device. As the centralized CEP engine is removed under this assumption, data subjects will directly respond to data consumers' queries. Most functions of the CEP engine are achieved on data subjects' local devices. However, unlike the CEP engine, data subjects usually cannot be fully trusted by data consumers. For example, consider a mobile phone application. Developers rarely share all details of their queries with application users. We are able to protect private patterns based on DP in such situations. However, without knowing the target patterns, privacy budgets cannot be properly distributed and will harm data utility. Therefore, we aim to investigate some form of supervision from data consumers that neither reveals target patterns nor harms privacy.

5 DP GUARANTEE FOR PATTERNS

We aim to propose a pattern-level DP and a PPM capable of handling most data stream queries and operators. We first present the basic definitions. Given an infinite data stream $S^D = (d_1, d_2, ...)$, any data tuple d of our interest is considered an event e, which

can be either numerical or categorical. A sequence of events can form a pattern $P = opr(e_1, e_2, ..., e_m)$, and a data stream can then be abstracted into a **pattern stream** $S^P = (P_1, P_2, ...)$. Among the detected patterns, some contain private information. These patterns are defined as **private patterns**, while all others are **public patterns**. Data consumers are interested in the patterns that we call **target patterns**.

Pattern-level DP should guarantee that a private and a public pattern are sufficiently indistinguishable concerning the responses after applying the PPM to them. For clarification, we first distinguish **pattern instances** and **pattern types**, and then define **in-pattern neighbors** and **pattern-level neighbors** as follows.

Definition 5.1. A pattern type $\mathcal{P}$ is a group of patterns specified by a given query q. All elements in $\mathcal{P}$ can be identified by q, and any pattern instance P_i identified by q is an element of $\mathcal{P}$, i.e., $P_i \in \mathcal{P}$.

Definition 5.2 (in-pattern neighbors). Two patterns $P = opr(e_1, e_2, ..., e_m)$, and $P' = opr(e'_1, e'_2, ..., e'_m)$, of the same pattern type $\mathcal{P}$ and of the same length are **in-pattern neighbors of** $\mathcal{P}$ if and only if (1) there exists a unique i such that $e_i \neq e'_i$, (2) for all $j \neq i$, $e_j = e'_j$ holds, and (3) if e_i and e'_i are numerical and if there exists e''_i such that $(e''_i - e_i)(e''_i - e'_i) < 0$, then for a pattern $P'' = opr(e_1, e_2, ..., e_{i-1}, e''_i, e_{i+1}, ..., e_m)$ of pattern type $\mathcal{P}$, either $P'' = P$ or $P'' = P'$ holds.

Definition 5.3 (pattern-level neighbors). Given a predefined pattern type $\mathcal{P}$, and two infinite pattern streams $S^P = (P_1, P_2, ...)$ and $S^{P'} = (P'_1, P'_2, ...)$, then S^P and $S^{P'}$ are **pattern-level neighbors** with respect to $\mathcal{P}$ if and only if for any integer i such that $P_i \in \mathcal{P}$, (1) P_i and P'_i are in-pattern neighbors of $\mathcal{P}$, and (2) for $j \neq i$, $P_j = P'_j$ holds.

In-pattern neighboring indicates that two patterns only differ by a basic event e_i, while pattern-level neighboring indicates that two pattern streams only differ by a pattern of a given pattern type $\mathcal{P}$. The pattern-level ϵ-DP can then be defined as follows.

Definition 5.4. Assume that $\mathcal{M}$ is a mechanism that takes a pattern stream D as input and outputs a response R that belongs to the group of all possible responses $\mathcal{R}$. Then we claim that $\mathcal{M}$ satisfies **pattern-level ϵ-DP** of a given type of patterns $\mathcal{P}$ (pattern-level DP of $\mathcal{P}$) if and only if for any pattern-level neighbors S^P and $S^{P'}$ of $\mathcal{P}$ and any sets of response $\mathcal{R}_i \subseteq \mathcal{R}$,

$$\Pr[\mathcal{M}(S^P) \in \mathcal{R}_i] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(S^{P'}) \in \mathcal{R}_i]$$

holds, where Pr denotes the probability function, and ϵ is the privacy budget.

This definition indicates that neighbor patterns should be of the same pattern type and one of the most similar patterns to each other. These requirements set a limit not only to the distance between the events of neighboring patterns but also between these patterns themselves. Our definition controls the upper bound of the intensity of privacy protection to a realistic level while attempting to maximize the output data utility for DP mechanisms.

These definitions are valid for any pattern made by any data stream operators. However, as they have formed a strict privacy guarantee and a complicated framework, the design of corresponding PPMs is also challenging. We therefore verify our methods on a simplified scenario, where we only involve patterns that made by sequence operators. We have presented this early-stage work in [5]. Here, the sequence operator is defined as an ordered set of information items [2]. For sequence operators, the definitions of **pattern-level neighbors** and **pattern-level DP** are still appropriate. Therefore, we are only required to simplify Definition 5.2, i.e., **in-pattern neighbors**. There are three conditions to formulate **in-pattern neighbors**. For sequence operators, the existence of conditions (1) and (2) are still compulsory, since they stipulate the smallest possible difference between in-pattern neighbors. However, condition (3) is redundant for sequence operators. It only controls the largest possible difference between in-pattern neighbors, which for a sequence operator is equivalent to the smallest difference, i.e., a basic event e_i.

6 PRIVACY-PRESERVING MECHANISMS

Given a well-defined pattern-level DP, we can then design pattern-level PPMs. A typical approach is to take existing mechanisms [4] as a basis, e.g., the randomized response, the Laplace mechanism, or the exponential mechanism. We can modify these mechanisms, combine them, and eventually adapt them to our own DP guarantee.

The simplest among them would be the randomized response. It is usually employed for categorical responses to a query, e.g., whether a pattern is detected or to which category the detected pattern belongs. Although there exist algorithms to transform numerical responses, e.g., the age of a person, into categorical responses such that randomized mechanisms can also be applied to numerical responses, these transformations usually lead to huge redundancy in computing resources. Therefore, other mechanisms can be more favorable selections for numerical query responses. We combine the randomized mechanism with other mechanisms to establish our own PPM framework. We assign categorical responses to a modified randomized mechanism while applying an adapted Laplace mechanism to numerical responses. We see the potential to optimize this combination further so that the application of PPMs is determined by both response types and the estimated performance.

For an ϵ-DP mechanism, the most crucial procedure is the distribution of its privacy budget ϵ. A trivial approach is to evenly distribute the privacy budget to each related data tuple. Our previous work [5] attempts to apply an adaptive approach that optimizes its distribution of privacy budgets based on historical data. However, the amount and variety of historical data are usually limited, and the corresponding budget distribution can hardly be fully optimized. In such cases, we may seek help from synthetic data. As data stream applications are usually built upon known scenarios, the corresponding synthetic data also have the potential to precisely simulate real-world cases and strengthen our approach. Combining both types of data, we first utilize synthetic data to build a pretrained budget distribution model and further optimize it using real-world historical data.

However, such an approach can only be conducted under the assumption that the optimized privacy budget distribution based historical and synthetic data is valid for the current data stream, i.e., the ground truth of a data stream is not frequently changed. When such an assumption does not hold, statistical approaches can be an alternative. Given well-defined private patterns and target

patterns, it is possible to calculate their correlations and conditional possibilities of occurrence. We can then attempt to formulate the relations between these statistics and the decrease in data utility when a certain privacy budget is assigned to a certain pattern. However, for different PPMs, statistical methods may vary greatly, as the ground truth and the calculation of probabilities are usually distinct among different PPMs.

7 EVALUATION

Our work needs to be evaluated by comparing it with other state-of the-art PPMs through practical experiments. Under the assumptions introduced, we believe that fairness and variety are the most critical factors in this evaluation.

Fairness indicates that different PPMs should be evaluated by equivalent metrics. Regarding the evaluation of privacy protection, the privacy budget ϵ of DP is the most favorable metric. However, for different DP guarantees, their privacy budgets are not initially equivalent. We must transform all measurements of privacy budgets into the same scale. As we aim to quantify the advantage of our approach, it is more intuitive to transfer the other forms of privacy budgets into ours instead of in a reverse way. Regarding the evaluation of data utility, we prefer to minimize the decrease in utility caused by applying a PPM. Mean Relative Error (MRE) is a typical metric to measure this decrease $MRE_U = \frac{U_{ord} - U_{PPM}}{U_{ord}}$, where U_{ord} denotes the ordinary data utility without applying any PPM, while U_{PPM} is the utility after employing a PPM. Our PPM usually aims to detect as many target patterns as possible, which can be measured by recall, and to reduce false detections, which can be measured by precision. The data utility U is therefore measured by both precision and recall.

Stronger privacy protection leads to additional noise and damages data utility even more. Considering the trade-off between privacy protection and data utility, one of them must be fixed to compare the overall performance among multiple PPMs. As data utility is determined by recall and precision, it is complicated to compare both simultaneously when privacy budgets are fixed. Therefore, we set recall and precision fixed and compare privacy budgets. A lower privacy budget indicates stronger privacy protection under the same data utility and hence an overall superior performance.

Considering the variety of experiments, the principal factors are the variety of datasets, e.g., different types of datasets, and the variety of test examples, e.g., different types of private patterns in the same dataset. There exist sufficient public datasets for PPM evaluations, e.g., the **Taxi** [1] [13, 14] dataset and the **Taobao** [2] [12] dataset. However, since only a few works study privacy protection with respect to patterns, most public datasets cannot be used to generate enough types of patterns for our evaluation, as we know little about the ground truth of these datasets. An alternative solution is to collect our own dataset in need, which costs a significant amount of time and human resources. Another option is to synthesize artificial datasets. They are more flexible but less persuasive than real-world datasets.

[1] https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

[2] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

8 CONCLUSIONS AND REFLECTIONS

For sequence operators, we have established a satisfactory ground for pattern-level approaches. We also propose a pattern-level DP and early-stage pattern-level PPMs with promising performance. Currently, we generalize these approaches for other operators. Furthermore, we will investigate local privacy protection approaches under the assumption that data consumers do not reveal their target patterns. We plan to evaluate the proposed approaches through practical experiments on public real-world datasets, collected datasets, and synthetic datasets.

ACKNOWLEDGMENTS

The author wants to thank Thomas Plagemann and Maik Benndorf for their valuable feedback. This work was funded by the Parrot Project (Research Council of Norway, project number 311197).

REFERENCES

[1] Yan Chen, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. 2017. Pegasus: Data-adaptive differentially private stream processing. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1375–1388.

[2] Gianpaolo Cugola and Alessandro Margara. 2012. Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)* 44, 3 (2012), 1–62.

[3] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. 2010. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*. 715–724.

[4] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[5] He Gu, Thomas Plagemann, Maik Benndorf, Vera Goebel, and Boris Koldehofe. 2023. Differential Privacy for Protecting Private Patterns in Data Streams. arXiv:2305.06105 [cs.DB]

[6] Manos Katsomallos, Katerina Tzompanaki, and Dimitris Kotzinos. 2022. Landmark Privacy: Configurable Differential Privacy Protection for Time Series. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*. 179–190.

[7] Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. 2014. Differentially private event sequences over infinite streams. *Proceedings of the VLDB Endowment* 7, 12 (2014), 1155–1166.

[8] Saravana Murthy Palanisamy, Frank Dürr, Muhammad Adnan Tariq, and Kurt Rothermel. 2018. Preserving privacy and quality of service in complex event processing through event reordering. In *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems*. 40–51.

[9] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.

[10] Xuebin Ren, Liang Shi, Weiren Yu, Shusen Yang, Cong Zhao, and Zongben Xu. 2022. LDP-IDS: Local differential privacy for infinite data streams. In *Proceedings of the 2022 International Conference on Management of Data*. 1064–1077.

[11] Ugur Sopaoglu and Osman Abul. 2021. Classification utility aware data stream anonymization. *Applied Soft Computing* 110 (2021), 107743.

[12] Tianchi. 2018. Taobao Dataset for Click-Through Rate Prediction. https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

[13] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. 2011. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 316–324.

[14] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. 99–108.