Landmark Privacy: Configurable Differential Privacy Protection for Time Series

Manos Katsomallos, Katerina Tzompanaki, Dimitris Kotzinos

Landmark Privacy: Configurable Differential Privacy Protection for Time Series

Manos Katsomallos, Katerina Tzompanaki, and Dimitris Kotzinos
ETIS UMR 8051, CY Cergy Paris University, ENSEA, CNRS
{emmanouil.katsomallos,aikaterini.tzompanaki,dimitrios.kotzinos}@cyu.fr

ABSTRACT

Several application domains, including healthcare, smart building, and traffic monitoring, require the continuous publishing of data, also known as time series. In many cases, time series are geotagged data containing sensitive personal details, and thus their processing entails privacy concerns. Several definitions have been proposed that allow for privacy preservation while processing and publishing such data, with *differential privacy* being the most prominent one. Most existing differential privacy schemes protect either a single timestamp (event-level), or all the data per user (user-level), or per window (w-event-level) in the time series, considering however all timestamps as equally significant. In this work, we define a novel configurable privacy notion, *landmark privacy*, which differentiates events into significant (*landmarks*) and regular, achieving to provide better data utility while preserving adequately the privacy of each event. We propose three schemes that guarantee landmark privacy, and design an appropriate dummy landmark selection module to better protect the actual temporal position of the landmarks. Finally, we provide a thorough experimental study where (i) we study the behavior of our framework on real and synthetic data, with and without temporal correlation, and (ii) demonstrate that landmark privacy achieves generally better data utility in the presence of landmarks than user-level privacy.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

differential privacy, privacy-preserving data publishing, time series

ACM Reference Format:
Manos Katsomallos, Katerina Tzompanaki, and Dimitris Kotzinos . 2022. *Landmark Privacy*: Configurable Differential Privacy Protection for Time Series . In *Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy (CODASPY '22), April 24–27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3508398.3511501

1 INTRODUCTION

The plethora of sensors currently embedded in personal devices and numerous *crowdsensing services* (e.g., Ring [1], TousAntiCovid [2], Waze [3], etc.) based on the collected personal, and usually geotagged and timestamped data. User–service interactions gather personal event-like data, which are tuples of an identifying attribute of an individual and the—possibly sensitive—information with a timestamp e.g., (*'Quackmore', 'dining', 'Canal Saint-Martin',* 17:00). When the interactions are performed in a continuous manner, we obtain *time series* of events. Depending on the duration, we distinguish the interaction/observation into *finite*, when taking place during a predefined time interval, and *infinite*, when taking place in an uninterrupted fashion. Example 1.1 demonstrates a user–service interaction that results in retrieving location-based information or reporting user-state at various locations.

Example 1.1. *Figure 1 shows a finite sequence of spatiotemporal data, generated by Quackmore, during an interval of* 8 *timestamps. Events in gray correspond to significant events, which Quackmore has defined beforehand, because they are related to his home (around Élysée), his workplace (around the Louvre), and his hangout (around Canal Saint-Martin).*

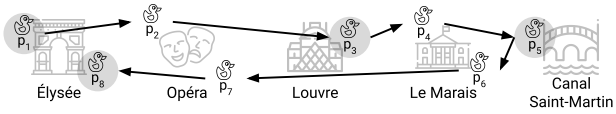


Figure 1: A time series with landmarks (highlighted in gray).

The regulation regarding the processing of user-generated data sets [34] requires the provision of privacy guarantees to the users. To accomplish this, various privacy techniques perturb the original data or their statistical output at the expense of the overall utility of the final output. Meanwhile, it is essential to provide data of high utility to the final consumers of the privacy-preserving process. A widely recognized method that introduces probabilistic randomness to the original data, while quantifying with a parameter ε ('privacy budget' [28]) the privacy/utility ratio, is *ε-differential privacy* [13]. *Event*, *user* [14], and *w-event* [22] comprise the possible levels of privacy protection. Event-level limits the protection to *any single event*, user-level protects *all the events* of any user, and w-event provides protection to *any sequence of w events*. In every case, privacy protection boils down to allocating to events an overall privacy budget that does not exceed ε.

The privacy mechanisms for the aforementioned levels assume that in a time series any single event, or any sequence of events, or the entire series of events respectively is equally privacy-significant for the users. In reality, this is an assumption that deteriorates unnecessarily the utility of the released data. The significance of an

event is related to certain user-defined privacy criteria, or to its adjacent events, as well as to the entire time series. We term significant events as *landmark events* or simply *landmarks*. Identifying landmarks can be done in an automatic or manual way; this is an orthogonal problem to the one presented here and thus it is out of scope of this work. For example, in spatiotemporal data, *places where an individual spent some time* denote *points of interest* (POIs) also known as stay points [25]. Such events, and more particularly their spatial attribute values, can be less privacy-sensitive, e.g., parks, theaters, etc., if the user visits them few times, but if individuals frequent them, they can reveal supplementary information, e.g., residences (home addresses), places of worship (religious beliefs) [30], etc. Another example is the detection of user interactions by *contact tracing* applications. This can be helpful in epidemic control, similar to the recent outbreak of the Coronavirus disease 2019 (COVID-19) epidemic [4]; however, the user may distinguish among contacts that are more privacy-sensitive than others. Last but not least, landmarks in *smart grid* electricity usage patterns may not only reveal the energy consumption of a user, but also information regarding activities, e.g., 'at work', 'sleeping', etc., or types of appliances already installed or recently purchased [23].

Example 1.2. *Continuing Example 1.1, Quackmore cares about protecting his landmarks (p_1, p_3, p_5, p_8) along with every release that he makes, however he is not equally interested for the other regular events in his trajectory. More technically, he cares about allocating a total budget of ε on any set of timestamps containing the landmarks and one regular event. Event-level protection is not suitable for this case, since it can only protect one event at a time. So, let us assume that we apply user-level privacy*[1]*, by distributing equal portions of ε to all the events, i.e., $\frac{\varepsilon}{8}$ to each one (see Figure 2). Indeed, we have protected the landmark points plus one regular event at any release as expected; we have allocated a total of $\frac{5\varepsilon}{8} < \varepsilon$ to these* 5 *events.*

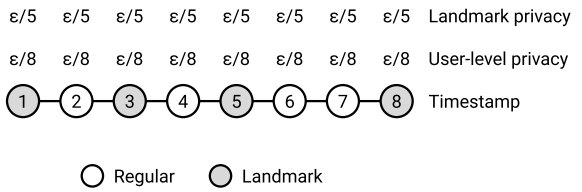


Figure 2: User-level and landmark ε-differential privacy protection for the time series of Figure 1.

However, perturbing by $\frac{\varepsilon}{8}$ each one of the regular points deteriorates the data utility unnecessarily; any budget less than or equal to $\frac{4\varepsilon}{8}$ would be sufficient for covering the user privacy requirements. On the other hand, our proposed privacy model, landmark privacy, directly considers only the 5 *events of interest (*4 *landmarks* +1 *current event) in every release, and thus changing the scope from all the time series to a significant subset of events. Subsequently, it allocates $\frac{\varepsilon}{5}$ to each one of these events. Consequently, we still achieve to protect all the significant events, while the utility of a perturbed event is greater than in the case of user-level privacy ($\frac{\varepsilon}{5} > \frac{\varepsilon}{8}$).*

[1]In this scenario, in order to protect all the landmarks from timestamp 1 to 8, w must be set to 8, which makes w-event privacy equivalent to user-level.

Motivation. We argue that protecting only landmark events along with any regular event is sufficient for the user privacy protection, while it improves data utility with respect to the conventional user-level privacy. Considering landmarks can prevent over-perturbing the data in the benefit of their final utility. Revisiting the scenario in Figure 2, if we want to protect the landmark points, we have to allocate at most a budget of ε to the landmarks, while saving some for the release of regular events. Essentially, the more budget we allocate to an event the less we protect it, but at the same time the more we maintain its utility. With landmark privacy we propose to distribute the budget by accounting only for the landmarks when we release an event of the time series, i.e., allocating $\frac{\varepsilon}{5}$ (4 landmarks + 1 regular point) to each event (see Figure 2). This way, we still guarantee that the landmarks are adequately protected, as they receive a total budget of $\frac{4\varepsilon}{5} < \varepsilon$. At the same time, we avoid over-perturbing the regular events, as we allocate to them a higher total budget ($\frac{4\varepsilon}{5}$) than in user-level ($\frac{\varepsilon}{2}$), and thus less noise. Hence, at any timestamp we achieve an overall privacy protection bounded by ε in the event set consisting of the released event and the landmarks.

Contributions. In this work, we formally define a novel privacy notion that we call *landmark privacy*. We apply this privacy notion to finite time series consisting of *landmarks* and regular events, and we design and implement three landmark privacy schemes. We further enhance our proposal by protecting the temporal position of the landmarks in the time series. We investigate landmark privacy under temporal correlation, which is inherent in time series publishing, and discuss how landmarks can affect the propagation of temporal privacy loss. Finally, we evaluate landmark privacy with real and synthetic data sets, in settings with or without temporal correlation, showcasing the validity of our proposal.

2 BACKGROUND AND RELATED WORK

This section provides the background knowledge that is essential to the presentation of this paper and discusses the related work. Landmark privacy is based on ε-differential privacy, and hence we revisit its definition and important properties in Section 2.1 before moving on to the main ideas of this paper. Although, its local variant [12] is more compatible with microdata, which is our use case, for the sake of simplicity in presentation we use its standard version. Still, all discussed notions apply for local differential privacy. Section 2.2 surveys implementations of differential privacy in scenarios of continuous data publishing, and Section 2.3 describes how it behaves under the presence of temporal correlation.

2.1 Differential privacy

Differential privacy [13] is a property of a privacy mechanism $\mathcal{M}$ processing a set of *privacy-sensitive* personal data D, while providing quantifiable privacy and utility guarantees.

Definition 1. [13] A privacy mechanism $\mathcal{M}$, with domain $\mathcal{D}$ and range O, satisfies ε-differential privacy, for a given privacy budget ε, if for every pair of neighboring data sets $D, D' \in \mathcal{D}$ and all sets $O \subseteq O$

$$\Pr[\mathcal{M}(D) \in O] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in O]$$

The *privacy budget* ε is a positive real number that represents the user-defined privacy goal [28]. $\mathcal{M}$ achieves stronger privacy protection for lower values of ε. Indeed, for lower values of ε, the *neighboring* data sets D and D' (i.e., they differ by one tuple) have greater chances to produce the same output. $\mathcal{M}$ is chosen based on the range and sensitivity of the query function f, the results of which it perturbs. We define the sensitivity of a query function f for all neighboring data sets D, $D' \in \mathcal{D}$ as $\Delta f = \max_{D,D' \in \mathcal{D}} \|f(D) - f(D')\|_1$.

2.1.1 Popular privacy mechanisms. A typical example of a differential privacy mechanism, for any function with range the set of real numbers, is the *Laplace mechanism* [15]. It draws randomly a value from the probability distribution of Laplace(μ, b), where μ stands for the location parameter and $b > 0$ is the scale parameter. In our case, μ is the original output value of f, and b is $\frac{\Delta f}{\varepsilon}$. A specialization of this mechanism for location data, based on a multivariate Laplace distribution, is the *Planar Laplace mechanism* [6].

For query functions that do not return a real number, e.g., 'What is the most visited country this year?', or in cases where perturbing the value of the output will completely destroy its utility, e.g., 'How many patients in the ICU?', most works use the *Exponential mechanism* [27]. Initially, a utility function u, with sensitivity Δu, maps pairs of the input value x and output value r to utility scores. Thereafter, the mechanism $\mathcal{M}$ selects an output value r from a set of possible outputs R with probability proportional to $\exp(\frac{\varepsilon u(x,r)}{2\Delta u})$.

Another technique for differential privacy mechanisms is the *randomized response* [39]. It is a privacy-preserving survey method that introduces probabilistic noise to the statistics of a research by randomly instructing respondents to answer truthfully or 'Yes' to a sensitive, binary question. Based on this methodology, the *Random response* mechanism [36] returns the true or flipped answer value with a probability proportional to the privacy budget ε.

2.1.2 Composition. Any combination of a set of independent differential privacy mechanisms satisfying a corresponding set of privacy guarantees shall satisfy differential privacy as well, i.e., provide a differentially private output. When we apply a series of independent (i.e., in the way that they inject noise) differential privacy mechanisms on independent data, we can quantify the privacy of the resulting output by summing up the privacy loss of each individual mechanism [28]. However when the data sets are disjoint, the final output's privacy loss equals the maximum privacy loss of the independent mechanisms.

2.1.3 Post-processing. Every time a data publisher interacts with (any part of) the original data set, it is mandatory to consume some of the available privacy budget according to the composition property of differential privacy (Section 2.1.2). However, the *post-processing* of a perturbed data set can be done without using any additional privacy budget as outlined in Theorem 1.

Theorem 1. [28] The post-processing of the output of a differential privacy mechanism does not change its privacy guarantee.

2.2 Differential privacy in time series

In privacy-preserving continuous data publishing [21], we consider the protection level with respect to not only the users, but also to the *events* occurring in the data. An event is a tuple of an identifying attribute of an individual and the sensitive data (including contextual information), and we can see it as corresponding to a record in a database where each individual may participate once. Data publishers typically release events in the form of sequences of data items, usually indexed in time order (time series) and geotagged, e.g., ('Daisy', 'at home at Montmartre at t_1'), ..., ('Donald', 'dining at Opera at t_1'). We use the term 'users' to refer to the *individuals*, also known as *participants*, who are referenced in the processed and published data. Therefore, they should not be confused with the consumers of the released data sets. We further define the three levels of privacy provided to users, and highlight some works implementing them.

Event-level [14] limits the privacy protection to *any single event* in a time series, providing high data utility. Wang and Zu [35] defined Correlated Time Series Differential Privacy, which guarantees that the correlation between the perturbation that is introduced by a Correlated Laplace Mechanism (CLM), and the original time series is indistinguishable. Chen et al. [11] developed *PeGaSus*, an algorithm for event-level differentially private stream processing that supports different categories of stream queries (counts, sliding window, and event monitoring) over multiple stream resolutions. Al-Dhubhani and Cazalas [5] proposed an adaptive privacy-preserving technique based on geo-indistinguishability, which adjusts the amount of noise required to obfuscate an individual's location based on its correlation level with the previously published locations.

User-level [14] protects *all the events* in a time series, providing high privacy protection. Fan et al. designed *FAST* [16], an adaptive system that allows the release of real-time aggregate time series by implementing sampling, perturbation, and filtering. Chen et al. [10] and Hua et al. [20] exploited a text-processing technique, the *n-gram* model, i.e., a contiguous sequence of n items from a given data sample, to release sequential data without releasing the noisy statistics (counts) of all of the possible sequences. Contrary to this approach, Li et al. [24] focus on publishing a set of trajectories, where each one is considered as a single entry in the data set. Farokhi [17], based on the discounted utility theory in economics, proposed *temporally discounted differential privacy*, a relaxation of the user-level protection that assigns different weights to the privacy budgets that have been invested in previous timestamps.

w-event-level [22] provides privacy protection to *any sequence of w events* in a time series. Based on the notion of decayed privacy [7], Kellaris et al. [22] proposed two mechanisms (Budget Distribution and Budget Absorption) following a sliding window methodology, which effectively distribute the privacy budget (exponentially fading and uniformly) to sub-mechanisms applied on the data of a window of the stream. Cao et al. [8] developed a framework that achieves *l-trajectory* privacy protection by dynamically adding noise at each timestamp, which exponentially fades over time. Wang et al. [37] presented *DP-PSP* which segments trajectories by taking into account points of interest in road networks and publishes privacy-preserving statistics.

Contrary to event-level, which provides privacy guarantees for a single event, user- and w-event-level offer stronger privacy protection by protecting a series of events. Event- and w-event-level better fit to scenarios of infinite data observation, whereas user-level is more appropriate when the span of data observation is finite. w-event- is narrower than user-level protection due to its

sliding window processing approach. In the extreme cases where w is equal either to 1 or to the length of the time series, w-event matches event- or user-level protection, respectively.

All of the aforementioned privacy protection levels consider all events as equally significant in terms of privacy, and hence cannot be applied to scenarios that require configurability. Furthermore, they cannot be easily modified to differentiate among events and adapt to the notion of landmarks. To fill this gap, we propose the novel privacy notion of landmark privacy, which guarantees differential privacy while accounting for landmarks.

2.3 Privacy loss under temporal correlation

Cao et al. [9] proposed a method for computing the temporal privacy loss (TPL) of a differential privacy mechanism in the presence of temporal correlation and background knowledge. The goal of their technique is to guarantee privacy protection and to bound the overall privacy loss at every timestamp under the assumption of independent data releases. It calculates TPL as the sum of the backward and forward temporal privacy loss, α_t^B and α_t^F, minus the default privacy loss ε of the mechanism (because it is counted twice in the aforementioned entities). This calculation is done for each individual included in the original data set and the overall TPL is equal to the maximum calculated value at every timestamp. α_t^B (or α_t^F) at any timestamp depends on the α_t^B (or α_t^F) at the previous/next timestamp, the backward/forward temporal correlation, and ε as described in Definition 2.

Definition 2. [9] The potential privacy loss of a privacy mechanism at a timestamp $t \in T$ due to a series of outputs $(\boldsymbol{o}_i)_{i\in T}$ and temporal correlation in its input D_t with respect to any adversary, targeting an individual with potential data items x_t (or x_t') and having knowledge $\mathbb{D}_t$ equal to $D_t - \{x_t\}$ (or $D_t' - \{x_t'\}$), is

$$\alpha_t = \sup_{x_t, x_t', (\boldsymbol{o}_i)_{i\in T}} \ln \frac{\Pr[(\boldsymbol{o}_i)_{i\in T} | x_t, \mathbb{D}_t]}{\Pr[(\boldsymbol{o}_i)_{i\in T} | x_t', \mathbb{D}_t]} \tag{1}$$

By analyzing Equation 1 we get the following

$$\begin{aligned}(1) = &\underbrace{\sup_{x_t, x_t', (\boldsymbol{o}_i)_{i\in[\min(T),t]}} \ln \frac{\Pr[(\boldsymbol{o}_i)_{i\in[\min(T),t]} | x_t, \mathbb{D}_t]}{\Pr[(\boldsymbol{o}_i)_{i\in[\min(T),t]} | x_t', \mathbb{D}_t]}}_{\text{Backward privacy loss } (\alpha_t^B)} \\ &+ \underbrace{\sup_{x_t, x_t', (\boldsymbol{o}_i)_{i\in[t,\max(T)]}} \ln \frac{\Pr[(\boldsymbol{o}_i)_{i\in[t,\max(T)]} | x_t, \mathbb{D}_t]}{\Pr[(\boldsymbol{o}_i)_{i\in[t,\max(T)]} | x_t', \mathbb{D}_t]}}_{\text{Forward privacy loss } (\alpha_t^F)} \\ &- \underbrace{\sup_{x_t, x_t', \boldsymbol{o}_t} \ln \frac{\Pr[\boldsymbol{o}_t | x_t, \mathbb{D}_t]}{\Pr[\boldsymbol{o}_t | x_t', \mathbb{D}_t]}}_{\text{Present privacy loss } (\varepsilon_t)}\end{aligned} \tag{2}$$

The intuition behind [9] is that a stronger temporal correlation results in higher privacy loss. When the transition matrix becomes sufficiently large, then the loss decreases due to the fact that larger matrices result in more uniform distributions. The authors investigate briefly all of the possible privacy levels; however, the solutions that they propose are applied only on the event-level.

3 LANDMARK PRIVACY

In this section, we introduce a new privacy definition, motivate its importance, and propose the methodology for achieving it.

3.1 Problem description and definition

Users generate a finite series of sensitive data over time, which are processed in batch mode in a secure and private way locally (or by a trusted curator) and are later published in order to be consumed by potentially adversarial data analysts. Data are produced as a series of events, which we call time series.

We argue that in continuous user-generated data publishing, events are not equally significant in terms of privacy. We term a significant event—according to user- or data-related criteria—as a *landmark* event. The identification of landmark events can be performed manually or automatically, and is an orthogonal problem to ours. First, we consider the landmark timestamps, i.e., their position in time, non-sensitive and provided by the user as input along with the privacy budget ε. For example, events p_1, p_3, p_5, p_8 in Figure 1 are landmark events. In Section 3.4, we extend our privacy framework to protect landmark timestamps when they are considered privacy-sensitive.

Definition 3 extends the notion of neighboring data sets (see Section 2.1) to the context of landmarks.

Definition 3. *Two time series S_T, S_T' of the same length $|T|$, with common starting and ending timestamps, are* L-landmark neighboring *with respect to a set of timestamps $L \subseteq T$ when*

(i) for each $S_T[i]$, $S_T'[i]$ with $i \in T$ and $S_T[i] \neq S_T'[i]$, it holds that $S_T[i]$, $S_T'[i]$ are neighboring, and
(ii) for each $i \in L \cup \{t\}$ such that $t \in T$, we have that $S_T[i]$, $S_T'[i]$ are neighboring.

Intuitively, S_T, S_T' are pairwise, i.e., at the same timestamps, equal or neighboring and their neighboring elements are on common landmarks and/or at most on one regular event. Definition 3 allows us to guarantee at any timestamp t that, given a set of landmark timestamps L in a time series, all data sets corresponding to $L \cup \{t\}$ are protected. Thus, we proceed to define *landmark privacy* (Definition 4), a configurable variation of differential privacy for time series with significant events.

Definition 4 (Landmark privacy). *Let $\mathcal{M}$ be a privacy mechanism with range O and domain $\mathcal{S}_T$ being the set of all time series with length $|T|$, where T is a sequence of timestamps, and $L \subseteq T$ be a set of landmark timestamps. $\mathcal{M}$ satisfies* (ε, L)-landmark privacy *if for all sets $O \subseteq O$, and for every pair of* L-landmark neighboring *time series S_T, S_T', it holds that*

$$Pr[\mathcal{M}(S_T) \in O] \leq e^{\varepsilon} Pr[\mathcal{M}(S_T') \in O]$$

User-level privacy can achieve landmark privacy, but it overperturbs the final data by not distinguishing between landmark and regular events. Theorem 2 states how to achieve the desired privacy goal for the landmarks and any event, i.e., a total budget less than ε, and at the same time provide better utility overall.

Theorem 2. *Let $\mathcal{M}$ be a mechanism with input a time series S_T, where T is the set of the involved timestamps, and $L \subseteq T$ be the set of landmark timestamps. $\mathcal{M}$ is decomposed to ε-differential private*

sub-mechanisms $\mathcal{M}_t$, for every $t \in T$, which apply independent randomness to the event at t. Then, given a privacy budget ε, $\mathcal{M}$ satisfies (ε, L)-landmark privacy *if for any t it holds that*

$$\sum_{i \in L \cup \{t\}} \varepsilon_i \leq \varepsilon$$

Proof. All mechanisms use independent randomness, and therefore for a time series $S_T = (D_i)_{i \in T}$ and outputs $(\boldsymbol{o}_i)_{i \in T} \in O \subseteq \mathcal{O}$ it holds that

$$Pr[\mathcal{M}(S_T) = (\boldsymbol{o}_i)_{i \in T}] = \prod_{i \in T} Pr[\mathcal{M}_i(D_i) = \boldsymbol{o}_i]$$

Likewise, for any landmark-neighboring time series S'_T of S_T with the same outputs $(\boldsymbol{o}_i)_{i \in T} \in O \subseteq \mathcal{O}$

$$Pr[\mathcal{M}(S'_T) = (\boldsymbol{o}_i)_{i \in T}] = \prod_{i \in T} Pr[\mathcal{M}_i(D'_i) = \boldsymbol{o}_i]$$

According to Definition 3, there exists $L \cup \{t\} \subseteq T$ such that $D_i = D'_i$ for $i \in L \cup \{t\}$. Thus, we get

$$\frac{Pr[\mathcal{M}(S_T) = (\boldsymbol{o}_i)_{i \in T}]}{Pr[\mathcal{M}(S'_T) = (\boldsymbol{o}_i)_{i \in T}]} = \prod_{i \in L \cup \{t\}} \frac{Pr[\mathcal{M}_i(D_i) = \boldsymbol{o}_i]}{Pr[\mathcal{M}_i(D'_i) = \boldsymbol{o}_i]}$$

D_i and D'_i are neighboring for $i \in L \cup \{t\}$. $\mathcal{M}_i$ is differential private and from Definition 1 we get that $\frac{Pr[\mathcal{M}_i(D_i) = \boldsymbol{o}_i]}{Pr[\mathcal{M}_i(D'_i) = \boldsymbol{o}_i]} \leq e^{\varepsilon_i}$. Hence, we can write

$$\frac{Pr[\mathcal{M}(S_T) = (\boldsymbol{o}_i)_{i \in T}]}{Pr[\mathcal{M}(S'_T) = (\boldsymbol{o}_i)_{i \in T}]} \leq \prod_{i \in L \cup \{t\}} e^{\varepsilon_i} = e^{\sum_{i \in L \cup \{t\}} \varepsilon_i}$$

For any $O \in \mathcal{O}$ we get $\frac{Pr[\mathcal{M}(S_T) \in O]}{Pr[\mathcal{M}(S'_T) \in O]} \leq e^{\sum_{i \in L \cup \{t\}} \varepsilon_i}$. If the formula of Theorem 2 holds, then we get $\frac{Pr[\mathcal{M}(S_T) \in O]}{Pr[\mathcal{M}(S'_T) \in O]} \leq e^{\varepsilon}$. Due to Definition 4 this concludes our proof. □

3.2 Achieving landmark privacy

In this section, we present schemes to achieve landmark privacy.

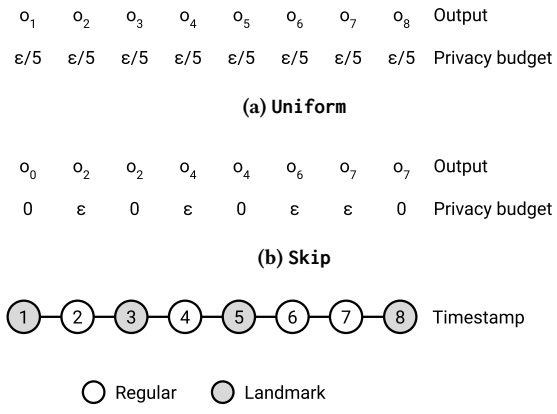


Figure 3: Application scenario of the (a) `Uniform` and (b) `Skip` landmark privacy schemes on a time series.

`Uniform`. Figure 3a shows the implementation of the baseline landmark privacy scheme for Example 1.2 which distributes uniformly the available privacy budget ε. In this case, it is enough to distribute at each timestamp the total privacy budget divided by the number of timestamps corresponding to landmarks, plus one, i.e., $\frac{\varepsilon}{|L|+1}$. Consequently, at each timestamp we protect every landmark, while reserving a part of ε for the current timestamp.

`Skip`. One might argue that we could skip the landmark data releases as we demonstrate in Figure 3b, by republishing previous, regular event releases. This would result in preserving all of the available privacy budget for regular events, equivalently to event-level protection, i.e., $\varepsilon_i = \varepsilon, \forall i \in T \setminus L$.

In practice, however, this approach can eventually pose arbitrary privacy risks, especially when dealing with geotagged data. Particularly, sporadic location data publishing or misapplying location cloaking could result in areas with sparse data points, indicating privacy-sensitive locations [19, 31]. We study this problem and investigate possible solutions in Section 3.4.

`Adaptive`. Next, we propose an adaptive privacy scheme (Figure 4) that accounts for changes in the sequence of sensitive data sets $(D_i)_{i \in T}$ by analyzing the respective private data releases $(\boldsymbol{o}_i)_{i \in T}$, and thus exploiting the post-processing property of differential privacy (Theorem 1).

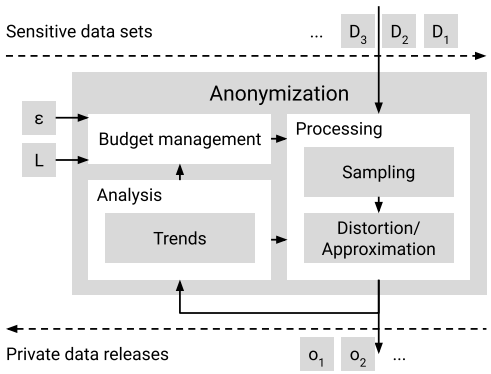


Figure 4: Concept of `Adaptive` landmark privacy.

Initially, the budget management component reserves uniformly the available privacy budget ε for each future release $\mathbf{o}$. At each timestamp, the processing component decides to either sample from the time series the current input and publish it with noise or release an approximation based on previous releases. In the case when it publishes with noise the original data, the analysis component estimates the data trends by calculating the difference between the current and the previous releases and compares the difference with the scale of the perturbation, i.e., $\frac{\Delta f}{\varepsilon}$ [22]. The outcome of this comparison determines the adaptation of the sampling rate of the processing component for the next events: if the difference is greater it means that the data trends are evolving, and therefore it must increase the sampling rate. When the mechanism approximates a landmark (but not a regular timestamp), the budget management component distributes the reserved privacy budget to the next

timestamps. Due to Theorem 1, the analysis component does not consume any privacy budget allowing for better data utility.

3.3 Landmark privacy under temporal correlation

From the discussion so far, it is evident that for the budget distribution it is not the positions, but rather the number of the landmarks that matters. However, this is not the case under the presence of temporal correlation.

The Hidden Markov Model scheme (as used in [9]) stipulates two important independence properties: (i) the future (or past) depends on the past (or future) via the present, and (ii) the current observation is independent of the rest given the current state. Hence, there is independence between an observation at a specific timestamp and previous/next data sets under the presence of the current input data set. Intuitively, knowing the data set at timestamp t stops the propagation of the Markov chain towards the next or previous timestamps in the time series.

In Section 2.3 we showed that the temporal privacy loss α_t at a timestamp t is calculated as the sum of the backward and forward privacy loss, α_t^B and α_t^F, minus the privacy budget ε_t, to account for the extra privacy loss due to previous and next releases $\boldsymbol{o}$ of $\mathcal{M}$ under temporal correlation. By Theorem 2, at every timestamp t we consider the data at t and at the landmark timestamps L. When sequentially composing the data releases for each timestamp i in $L \cup \{t\}$ we consider the previous releases in the whole time series until the timestamp i^- that is exactly before i in the ordered $L \cup \{t\}$, and the next data releases in the whole time series until the timestamp i^+ that is exactly after i in the ordered $L \cup \{t\}$. Figure 5 illustrates i^- and i^+ in Example 1.1, while we formalize the landmark temporal privacy loss in Definition 5.

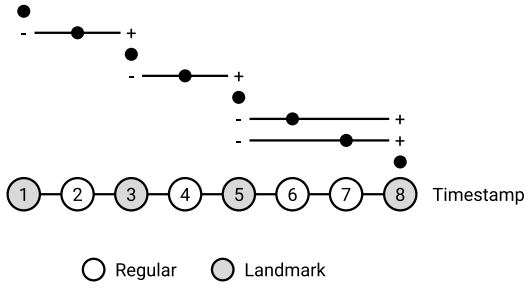


Figure 5: The timestamps exactly before (−) and after (+) every timestamp, where that is applicable, for the calculation of the temporal privacy loss.

Definition 5. *Given a landmark set L in a set of timestamps T, the potential overall temporal privacy loss of a privacy mechanism $\mathcal{M}$ at any timestamp in $L \cup \{t\}$ is*

$$\sum_{i \in L \cup \{t\}} \alpha_i$$

where for $i^-, i^+ \in L \cup \{t\}$ being the timestamps exactly before and after i, α_i is equal to

$$\underbrace{\ln \frac{\Pr[(\boldsymbol{o})_{i \in [i^-+1,i]} | D_i]}{\Pr[(\boldsymbol{o})_{i \in [i^-+1,i]} | D_i']}}_{\alpha_i^B} + \underbrace{\ln \frac{\Pr[(\boldsymbol{o})_{i \in [i,i^+-1]} | D_i]}{\Pr[(\boldsymbol{o})_{i \in [i,i^+-1]} | D_i']}}_{\alpha_i^F} - \underbrace{\ln \frac{\Pr[\boldsymbol{o}_i | D_i]}{\Pr[\boldsymbol{o}_i | D_i']}}_{\varepsilon_i} \quad (3)$$

As presented in [9], the temporal privacy loss of a time series (without landmarks) can be bounded by a given privacy budget ε. Intuitively, by Equation 3 the temporal privacy loss incurred when considering landmarks is less than the temporal loss in the case without the knowledge of the landmarks. Thus, the temporal privacy loss in landmark privacy can be also bounded by ε.

3.4 Protecting landmarks

So far, we assumed that the timestamps in the landmark set L are not privacy-sensitive, and therefore we used them in our schemes as they were. This may pose a direct or indirect privacy risk to the users. For the former, we consider the case where we desire to publish L as complimentary information to the release of the event values. For the latter, a potentially adversarial data analyst may infer L by observing the values of the privacy budget, which is usually an inseparable attribute of the data release as an indicator of the privacy guarantee to the users and as an estimate of the data utility to the analysts. Hence, in both cases, a user-defined L, which is supposed to facilitate the configurable privacy protection of the user, could end up posing a privacy risk to them.

In Example 3.1, we demonstrate the extreme case of the application of the `Skip` landmark privacy scheme from Figure 3b, where we approximate landmarks with the latest data release and invest all of the available privacy budget to regular events.

Example 3.1. *Figure 6 shows the privacy risk that the application of a landmark privacy scheme that nullifies or approximates outputs, similar to* `Skip`*, might cause. We point out in red the details that might cause indirect information inference. In this extreme case, the minimization of the privacy budget in combination with nullifying the output (either by not publishing or by adding a lot of noise) or approximating the current output with previously released outputs might hint to any adversary that the current event is a landmark.*

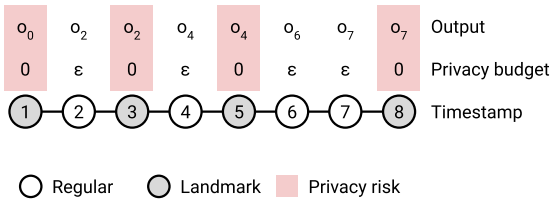


Figure 6: The privacy risk (highlighted in red) that the application of the landmark privacy `Skip` scheme might pose.

Apart from the privacy budget that we invested at landmarks, we can observe a pattern for the budgets at regular events as well. Therefore, an adversary who observes the values of the privacy budget can easily infer not only the number but also the exact temporal position of the landmarks.

The main idea of the privacy-preserving dummy landmark selection module is to privately select extra landmark event timestamps,

i.e., dummy landmarks, from the set of timestamps $T \setminus L$ of the time series S_T and add them to the original landmark set L. Selecting extra events, on top of the actual landmarks, as dummy landmarks, can render the actual ones indistinguishable. The goal is to create a new set L' such that $L \subset L' \subseteq T$.

First, we generate a set of dummy landmark set options by adding regular event timestamps from $T \setminus L$ to L (Section 3.4.1). Then, we utilize the exponential mechanism, with a utility function that calculates an indicator for each of the options in the set, based on how much it differs from the original landmark set L, and randomly select one of the options (Section 3.4.2). This process provides an extra layer of privacy protection to landmarks, and thus allows the processing, and thereafter releasing, of landmark timestamps.

3.4.1 *Dummy landmark selection.*

`Optimal`. The `Optimal` algorithm generates every possible combination (options) of landmark sets L' containing one set from every possible size, i.e, $|L|+1, |L|+2, \ldots, |T|$. Each L' contains the original landmarks along with timestamps of regular events from $T \setminus L$ (dummy landmarks). Then, it evaluates each option by comparing each of its sets with the original landmark set L and estimating an overall similarity score for each option. We discuss possible utility score functions later on in Section 3.4.2. The goal of this process is to select the option that contains the combination of dummy landmark sets that achieve the best score.

This procedure guarantees to return the optimal option with regard to the original set L. However, it is rather costly in terms of complexity. In more detail, given $|T \setminus L|$ regular events and a combination of size r, it requires $O(C(|T \setminus L|, r) + 2^{C(|T \setminus L|, r)})$ time and $O(r * C(|T \setminus L|, r))$ space. Next, we present a `Heuristic` solution with improved time and space requirements.

`Heuristic`. The `Heuristic` algorithm follows an incremental methodology and at each step it selects a new timestamp, corresponding to a regular event from $T \setminus L'$. In this case, the elements of L' at each step differ by one from the one that the algorithm selected in the previous step. Similar to the `Optimal`, it selects a new set based on a predefined similarity metric until it selects a set that is equal to the size of the series of events, i.e., $L' = T$.

In terms of complexity, given $|T \setminus L|$ regular events, the `Heuristic` requires $O(|T \setminus L|^2)$ time and space. Note that the reverse process, i.e., starting with T landmarks and removing until $|L'| = |L| + 1$, performs similarly.

`Partitioned`. We improve the complexity of the `Heuristic` algorithm by partitioning the landmark timestamp sequence L. The novelty of this algorithm lies in the fact that it deals with the event series as a histogram which allows it to take advantage of its relevant features and methodology. Particularly, it uses the Freedman-Diaconis rule, which is resilient to outliers and takes into account the data variability and data size [29], and generates a histogram from the landmark set L. This way, it achieves an improved complexity, compared to the `Heuristic`, that is dependent on the histogram's bin size. Algorithm 1 demonstrates the overall process.

Function `getHist` generates a histogram with bins of size h for a given time series timestamps T and landmark set L. For every new histogram version, the `getDiff` function (Line 11) finds the difference from the original histogram; for this operation it utilizes the Euclidean distance (see Section 4.3.1 for more details). In Lines 7-14, the algorithm checks every histogram version by incrementing each bin by 1 and comparing it to the original (Line 12). In the end, it returns opts which contains all the versions of hist that are closest to the original hist for all possible bin sizes of hist.

Algorithm 1: `Partitioned` landmark set options generation

Data: the time series timestamps T, the landmark set L
Output: the selected landmark set options opts

```
1  hist, h ← getHist(T, L)
2  histCur ← hist
3  opts ← []
4  while sum(histCur) ≠ len(T) do
5      diffMin ← ∞
6      opt ← histCur
7      foreach h_i in histCur do
8          if h_i + 1 ≤ h then
9              histTmp ← histCur
10             histTmp[i] ← histTmp[i] + 1
11             diffCur ← getDiff(hist, histTmp)
12             if diffCur < diffMin then
13                 diffMin ← diffCur
14                 opt ← histTmp
15     histCur ← opt
16     opts ← opt
17 return opts
```

3.4.2 *Privacy-preserving option selection.*

The algorithms that we presented in Section 3.4.1 return a set of possible versions of the original landmark set L by adding extra timestamps in it from the series of events at timestamps $T \setminus L$. In the next step, we randomly select a set by utilizing the exponential mechanism (Section 2.1.1). For this procedure, we allocate a small fraction of the available privacy budget, i.e., 1% or even less (see Section 4.3.2 for more details), which adds up to that of the publishing scheme according to the sequential composition theorem.

Utility score function. Prior to selecting a landmark timestamp set including the original along with dummy landmarks, the exponential mechanism evaluates each set using a utility score function. We present here two ways of doing so.

One way to evaluate each set is by taking into account the temporal position of the events in the sequence. Events that occur at recent timestamps are more likely to reveal sensitive information regarding the users involved [22]. Hence, indicating the existence of dummy landmarks nearby actual landmarks can increase the adversarial confidence regarding the location of the latter within a series of events. In other words, sets with dummy landmarks with less average temporal distance from actual landmarks achieve better utility scores.

Another approach for the utility score function is to consider the number of events in each set. Sets with more dummy landmarks may render actual landmarks more indistinguishable, and therefore provide less utility. Consequently, more dummy landmarks lead

to distributing the privacy budget to more events, and therefore leading to more robust overall privacy protection.

Option release. In the last step, the privacy-preserving dummy landmark selection module releases a new landmark set (including the original landmarks along with the dummy ones) from the options that were generated in the previous step, by utilizing the exponential mechanism.

The options generated by the `Optimal` and `Heuristic` algorithms contain actual timestamps that can be utilized directly by the landmark privacy schemes that we presented in Section 3.2. However, the `Partitioned` algorithm returns histograms instead of timestamps. Therefore, we need to process the result of the exponential mechanism further by sampling without replacement from the set $T \setminus L$ according to the selected histogram's probability density function.

4 EXPERIMENTAL EVALUATION

In this section we present the experiments that we performed in order to evaluate landmark privacy on real and synthetic data sets. Section 4.1 contains all the details regarding the data sets the we used for our experiments along with the system configurations. Section 4.2 evaluates the data utility of the landmark privacy schemes that we designed in Section 3.2 in comparison to user and event level, and investigates the behavior of the privacy loss under temporal correlation for different distributions of landmarks. Section 4.3 justifies our decisions while designing the privacy-preserving dummy landmark selection module in Section 3.4 and the data utility impact of the latter.

4.1 Configurations and data sets

We implemented our experiments[2] in Python 3.9.7 and executed them on a machine with an Intel i7-6700HQ at 3.5GHz CPU and 16GB RAM, running Manjaro Linux 21.1.5. We repeated each experiment 100 times and we report the mean over these iterations.

4.1.1 *Data sets.*

Real data sets. For consistency, we sample from each of the following data sets the first 1,000 entries that satisfy the configuration criteria that we discuss in detail in Section 4.1.2.

Copenhagen [32] data set was collected via the smartphone devices of 851 university students over a period of 4 week as part of the Copenhagen Networks Study. Each device was configured to be discoverable by and to discover nearby Bluetooth devices every 5 minutes. Upon discovery, each device registers (i) the timestamp in seconds, (ii) the device's unique identifier, (iii) the unique identifier of the device that it discovered (-1 when no device was found or -2 for any non-participating device), and (iv) the Received Signal Strength Indicator (RSSI) in dBm. Half of the devices have registered data at at least 81% of the possible timestamps. 3 devices (449, 550, 689) satisfy our configuration criteria (Section 4.1.2) within their first 1,000 entries. From those 3 devices, we picked the first one, i.e., device with identifier '449', and utilized its 1,000 first entries out of 12,167 unique valid contacts.

[2]Source code available at https://gitlab.com/crabbysalmon/codaspy22

HUE [26] contains the hourly energy consumption data of 22 residential customers of BCHydro, a provincial power utility in British Columbia. The measurements for each residence are saved individually and each measurement contains (i) the date (YYYY-MM-DD), (ii) the hour, and (iii) the energy consumption in kWh. In our experiments, we used the first residence, i.e., residence with identifier '1', that satisfies our configuration criteria (Section 4.1.2) within its first 1,000 entries. In those entries, out of a total of 29,231 measurements, we estimated an average energy consumption equal to 0.88kWh and a value range within $[0.28, 4.45]$.

T-drive [40] consists of 15 million GPS data points of the trajectories of 10,357 taxis in Beijing, spanning a period of 1 week and a total distance of 9 million kilometers. The taxis reported their location data on average every 177 seconds and 623 meters approximately. Each vehicle registers (i) the taxi unique identifier, (ii) the timestamp (YYYY-MM-DD HH:MM:SS), (iii) longitude, and (iv) latitude. These measurements are stored individually per vehicle. We sampled the first 1000 data items of the taxi with identifier '2', which satisfied our configuration criteria (Section 4.1.2).

Synthetic. We generated synthetic time series of length equal to 100 timestamps, for which we varied the number and distribution of landmarks. In this way, we have a controlled data set that we can use to study the behavior of our proposal. We take into account only the temporal order of the points and the position of regular and landmark events within the time series. In Section 4.1.2, we explain in more detail our configuration criteria.

4.1.2 *Configurations.*

In this section we discuss how we tune relevant parameters of the problem for the experiments. First, we vary the landmark percentage, i.e., the ratio of timestamps that we attribute to landmarks and regular events, in order to explore the behavior of our methodology in all possible scenarios. Second, for each data set, we implement a privacy mechanism that injects noise related to the type of its attribute values and we tune the parameters of each scheme accordingly. Third, we explain how we generate synthetic data sets with various degrees of temporal correlation so as to observe the impact on the temporal privacy loss.

Landmark percentage. In the Copenhagen data set, a landmark represents a timestamp at which a specific contact device is registered. After identifying the unique contacts within the sample, we achieve each desired landmarks to regular events ratio by considering a list that contains a part of these contact devices. In more detail, we achieve 0% landmarks by considering an empty list of contact devices, 20% by extending the list with [3, 6, 11, 12, 25, 29, 36, 39, 41, 46, 47, 50, 52, 56, 57, 61, 63, 78, 80], 40% with [81, 88, 90, 97, 101, 128, 130, 131, 137, 145, 146, 148, 151, 158, 166, 175, 176], 60% with [181, 182, 192, 195, 196, 201, 203, 207, 221, 230, 235, 237, 239, 241, 254], 80% with [260, 282, 287, 289, 290, 291, 308, 311, 318, 323, 324, 330, 334, 335, 344, 350, 353, 355, 357, 358, 361, 363], and 100% by including all of the possible contacts.

In HUE, we consider as landmarks the events that have energy consumption values below a certain threshold. More precisely, we get 0%, 20% 40%, 60%, 80%, and 100% landmarks by setting the energy consumption threshold at 0.28kWh, 1.12kWh, 0.88kWh, 0.68kWh, 0.54kWh, and 4.45kWh respectively.

In T-drive, a landmark represents a location where a vehicle stopped for some time. We achieved the desired landmark percentages by utilizing the method of Li et al. [25] for detecting stay points in trajectory data. In more detail, the algorithm checks for each data item if each subsequent item is within a given distance threshold Δl and measures the time period Δt between the present point and the last subsequent point. After analyzing the data and experimenting with different pairs of distance and time period, we achieve 0%, 20% 40%, 60%, 80%, and 100% landmarks by setting the (Δl in meters, Δt in minutes) pairs input to the stay point discovery method as [(0, 1000), (2095, 30), (2790, 30), (3590, 30), (4825, 30), (10350, 30)].

We generated synthetic data with *skewed* (the landmarks are distributed towards the beginning/end of the series), *symmetric* (in the middle), *bimodal* (both end and beginning), and *uniform* (all over the time series) landmark distributions. In order to get landmark sets with the above distribution features, we generate probability distributions with restricted domain to the beginning and end of the time series, and sample from them, without replacement, the desired number of points. For each case, we place the location, i.e., centre, of the distribution accordingly. More precisely, for symmetric we put the location in the middle of the time series and for left/right skewed to the right/left. For bimodal we combine two mirrored skewed distributions. Finally, for the uniform distribution we distribute the landmarks randomly throughout the time series. For consistency, we calculate the scale parameter of the corresponding distribution depending on the length of the time series by setting it equal to the series' length over a constant.

Privacy parameters. For all the real data sets, we implement ε-differential privacy by selecting a mechanism from those that we described in Section 2.1.1 that is better suited for the type of its sensitive attributes. To perturb the contact tracing data of the Copenhagen data set, we utilize the *random response* technique [38], and at each timestamp we report truthfully, with probability $p = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}$, whether the current contact is a landmark or not. We randomize the energy consumption in HUE with the Laplace mechanism [15]. For T-drive, we perturb the location data with noise that we sample from the Planar Laplace mechanism [6].

We set the privacy budget $\varepsilon = 1$ for all of our experiments and, for simplicity, we assume that for every query sensitivity it holds that $\Delta f = 1$. Note that changing the value of ε would not affect the conclusions we drive from the experiments (even though the range of the results would change). For the experiments that we performed on the synthetic data sets, the original values to be released are not relevant to what we want to observe, and thus we ignore them.

Temporal correlation. Despite the inherent presence of temporal correlation in time series, it is challenging to correctly discover and quantify it. For this reason, and in order to create a more controlled environment for our experiments, we chose to conduct tests relevant to temporal correlation using synthetic data sets. We model the temporal correlation in the synthetic data as a *stochastic matrix* P, using a *Markov Chain* [18]. P is an $n \times n$ matrix, where the element P_{ij} represents the transition probability from a state i to another state $j, \forall i, j \leq n$. It holds that the elements of every row j of P sum up to 1. We follow the *Laplacian smoothing* technique [33], as utilized in [9], to generate the matrix P with a degree of temporal correlation $s > 0$ equal to

$$\frac{(I_n)_{ij} + s}{\sum_{k=1}^{n}((I_n)_{jk} + s)}$$

where I_n is an *identity matrix* of size n. The value of s is comparable only for stochastic matrices of the same size and dictates the strength of the correlation; the lower its value, the stronger the correlation degree. In our experiments, for simplicity, we set $n = 2$ and we investigate the effect of *weak* ($s = 1$), *moderate* ($s = 0.1$), and *strong* ($s = 0.01$) temporal correlation degree on the temporal privacy loss.

4.2 Landmark events

In this section, we present the experiments that we performed, to test the methodology that we presented in Section 3.2, on real and synthetic data sets.

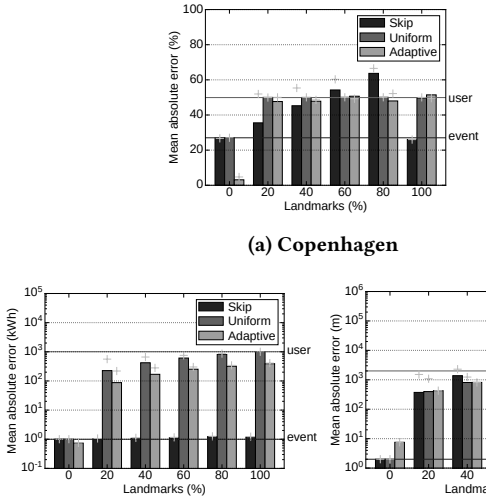
With the experiments on the real data sets (Section 4.2.1), we show the performance in terms of data utility of our three landmark privacy schemes: `Skip`, `Uniform` and `Adaptive`. We define data utility as the mean absolute error introduced by the privacy mechanism. We compare with the event- and user-level differential privacy protection levels, and show that, in the general case, landmark privacy allows for better data utility than user-level differential privacy while balancing between the two protection levels.

With the experiments on the synthetic data sets (Section 4.2.2) we show how the temporal privacy loss, i.e., the privacy budget ε with the extra privacy loss because of the temporal correlation, changes when tuning the size and statistical characteristics of the input landmark set L. We observe that a greater average landmark–regular event distance in a time series can result into greater temporal privacy loss under moderate and strong temporal correlation.

4.2.1 Landmark privacy schemes. Figure 7 exhibits the performance (bars) of the three schemes, `Skip`, `Uniform`, and `Adaptive` applied on the three data sets that we study. Notice that, in the cases when we have 0% and 100% of the events being landmarks, we get the same behavior as in event- and user-level privacy respectively. This happens due the fact that at each timestamp we take into account only the data items at the current timestamp and ignore the rest of the time series (event-level) when there are no landmarks. Whereas, when each timestamp corresponds to a landmark we consider and protect all the events throughout the entire series (user-level).

For the Copenhagen data set (Figure 7a), `Adaptive` has an overall consistent performance and works best for 60% and 80% landmarks. We notice that for 0% landmarks, it achieves better utility than the event-level protection due to the combination of more available privacy budget per timestamp (due to the absence of landmarks) and its adaptive sampling methodology. `Skip` excels, compared to the others, at cases where it needs to approximate 20%, 40%, or 100% of the times. In general, we notice that, for this data set and due to the application of the random response technique, it is more beneficial to either invest more privacy budget per event or prefer approximation over introducing randomization.

The combination of the small range of measurements ([0.28, 4.45] with an average of 0.88kWh) in HUE (Figure 7b) and the



(a) Copenhagen

(b) HUE **(c) T-drive**

Figure 7: The mean absolute error (a) as a percentage, (b) in kWh, and (c) in meters of the released data, for different landmark percentages. The markers indicate the corresponding measurements with the incorporation of the privacy-preserving dummy landmark selection module.

large scale in the Laplace mechanism, allows for schemes that favor approximation over noise injection to achieve a better performance in terms of data utility. Hence, `Skip` achieves a constant low mean absolute error. Regardless, the `Adaptive` scheme performs by far better than `Uniform` and balances between event- and user-level protection for all landmark percentages.

In T-drive (Figure 7c), `Adaptive` outperforms `Uniform` by 10%–20% for all landmark percentages greater than 40% and `Skip` by more than 20%. The lower density (average distance of 623m) of the T-drive data set has a negative impact on the performance of `Skip` because republishing a previous perturbed value is now less accurate than perturbing the current location.

Principally, we can claim that the `Adaptive` is the most reliable and best performing scheme, if we consider the drawbacks of `Skip`, particularly in spatiotemporal data, e.g., sporadic location data publishing or misapplying location cloaking, that could lead to the indication of privacy-sensitive attribute values [19, 31]. Moreover, implementing a more data-dependent sampling method that accounts for changes in the trends of the input data and adapts its rate accordingly, would result in a more effective budget allocation that would improve the performance of `Adaptive` in terms of data utility. We defer this study for the future.

4.2.2 Temporal distance and correlation. Figure 8 shows a comparison of the average temporal distance of the events from the previous/next landmark or the start/end of the time series for various distributions in our synthetic data. More specifically, we model the distance of an event as the count of the total number of events between itself and the nearest landmark or the time series edge.

We observe that the uniform and bimodal distributions tend to limit the regular event–landmark distance. This is due to the fact

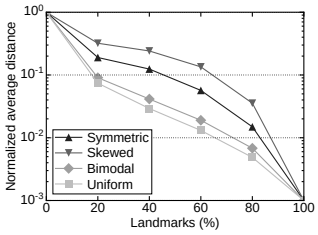
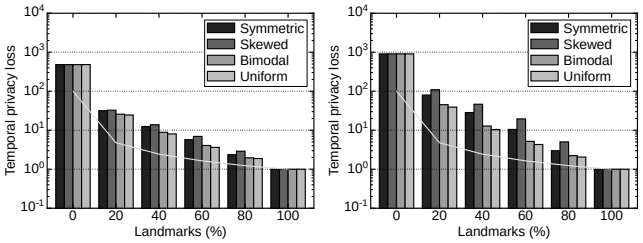


Figure 8: Average temporal distance of regular events from the landmarks for different landmarks percentages within a time series in various landmark distributions.

that the former scatters the landmarks, while the latter distributes them on both edges, leaving a shorter space uninterrupted by landmarks. On the contrary, distributing the landmarks at one part of the sequence, as in skewed or symmetric, creates a wider space without landmarks. This study provides us with different distance settings, to be used in the subsequent temporal privacy loss study.

Figure 9 illustrates a comparison among the aforementioned distributions regarding the temporal privacy loss under (a) moderate, and (b) strong temporal correlation degrees. The line shows the overall privacy loss—for all cases of landmark distribution—without temporal correlation. The privacy loss under a weak correlation degree converges with all possible distributions for all landmark percentages, and thus we omit its presentation.



(a) Moderate correlation **(b) Strong correlation**

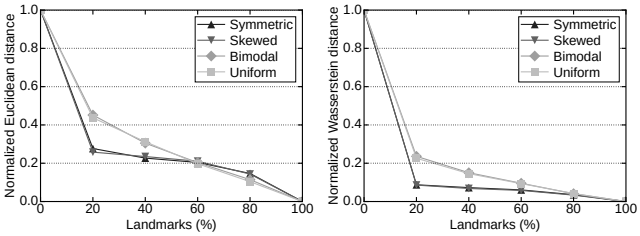
Figure 9: The temporal privacy loss for different landmark percentages and distributions under (a) moderate, and (b) strong degrees of temporal correlation. The line shows the overall privacy loss without temporal correlation.

In combination with Figure 8, we conclude that a greater average landmark–regular event distance in a distribution can result into greater temporal privacy loss under moderate and strong temporal correlation. This is due to the fact that the backward/forward privacy loss accumulates more over time in wider spaces without landmarks (see Section 2.3). Furthermore, the behavior of the privacy loss is as expected regarding the temporal correlation degree: a stronger correlation degree generates higher privacy loss while widening the gap between the different distribution cases. On the contrary, a weaker correlation degree makes it harder to differentiate among the landmark distributions.

4.3 Selection of dummy landmarks

In this section, we present the experiments on the methodology for the dummy landmark selection presented in Section 3.4, on the real and synthetic data sets. Due to the high complexity of the `Optimal` and `Heuristic` algorithms, we choose to evaluate only the `Partitioned`, which is the optimized solution that we designed. With the experiments on the synthetic data sets (Section 4.3.1) we show the normalized Euclidean and Wasserstein distance metrics (not to be confused with the temporal distances in Figure 8) of the time series histograms for various distributions and landmark percentages. This allows us to justify our design decisions for our concept that we showcased in Section 3.4. With the experiments on the real data sets (Section 4.3.3), we show the performance in terms of utility of our three landmark schemes in combination with the privacy-preserving dummy landmark selection module, which enhances the privacy protection that our concept provides.

4.3.1 Dummy landmark selection utility metrics. Figure 10 demonstrates the normalized distance that we obtain when we utilize either (a) the Euclidean or (b) the Wasserstein distance metric to obtain a set of landmarks including regular events.

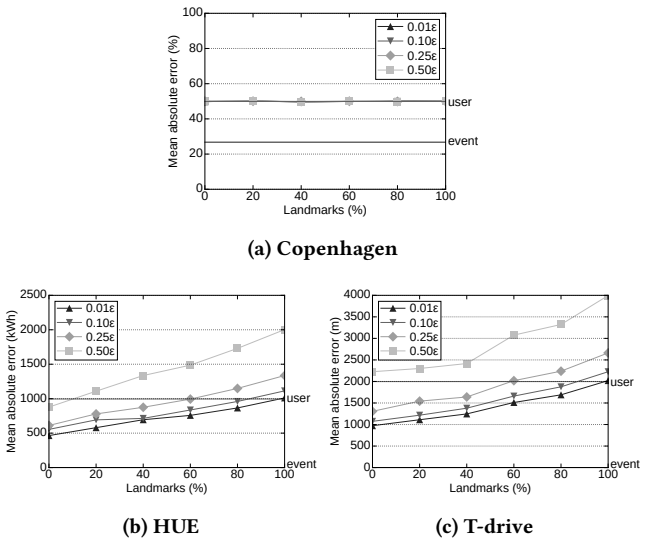


(a) Euclidean **(b) Wasserstein**

Figure 10: The normalized (a) Euclidean, and (b) Wasserstein distance of the generated landmark sets for different landmark percentages.

Comparing the results of the Euclidean distance in Figure 10a with those of the Wasserstein in Figure 10b we conclude that the Euclidean distance provides more consistent results for all possible distributions. The maximum difference per landmark percentage is approximately 0.2 for the former and 0.15 for the latter between the bimodal and skewed landmark distributions. Overall, the Euclidean distance achieves a mean normalized distance of 0.3, while the Wasserstein distance a mean normalized distance that is equal to 0.2. Therefore, and by observing Figure 10, Wasserstein demonstrates a less consistent performance and less linear behavior among all possible landmark distributions. Thus, we choose to utilize the Euclidean distance metric for the implementation of the privacy-preserving dummy landmark selection module in Section 3.4.

4.3.2 Privacy budget tuning. In Figure 11, we test the `Uniform` mechanism in real data by investing different ratios (1%, 10%, 25%, and 50%) of the available privacy budget ε in the dummy landmark selection module and the remaining to perturbing the original data values, in order to figure out the optimal ratio value. `Uniform` is our baseline implementation, and hence allows us to derive more accurate conclusions in this case. In general, we are expecting to observe that greater ratios will result in more accurate, i.e., smaller, landmark sets and less accurate values in the released data.



(a) Copenhagen

(b) HUE **(c) T-drive**

Figure 11: The mean absolute error (a) as a percentage, (b) in kWh, and (c) in meters of the released data for different landmark percentages. We apply the `Uniform` landmark privacy mechanism and vary the ratio of the privacy budget ε that we allocate to the dummy landmark selection module.

The randomized response mechanism is tolerant to the fluctuations of the privacy budget, and hence its application to the Copenhagen data set (Figure 11a) results in a constant performance in terms of data utility. For HUE (Figure 11b) and T-drive (Figure 11c), we observe that our implementation performs better for lower ratios, e.g., 0.01, where we end up allocating the majority of the available privacy budget to the data release process instead of the dummy landmark selection module. The results of this experiment indicate that we can safely allocate the majority of ε to the data publishing process, and therefore achieve better data utility, while guaranteeing more robust privacy protection.

4.3.3 Privacy schemes and dummy landmark selection. Figure 7 exhibits the performance of `Skip`, `Uniform`, and `Adaptive` schemes (presented in detail in Section 3.2) in combination with the dummy landmark selection module (Section 3.4).

In comparison with the utility performance without the dummy landmark selection module (solid bars), we notice a slight deterioration for all three schemes (markers). This is natural since we allocated part of the available privacy budget to the privacy-preserving dummy landmark selection module, which in turn increased the number of landmarks, except for the case of 100% landmarks. Therefore, there is less privacy budget available for data publishing throughout the time series. `Skip` performs best in our experiments with HUE (Figure 7b), due to the low range in the energy consumption and the high scale of the Laplace noise that it avoids due to the employed approximation. However, for the Copenhagen data set (Figure 7a) and T-drive (Figure 7c), `Skip` attains high

mean absolute error, which exposes no benefit with respect to user-level protection. Overall, Adaptive has a consistent performance in terms of utility for all of the data sets that we experimented with, and almost always outperforms the user-level privacy protection. Thus, Adaptive is selected as the best scheme to use in general.

5 CONCLUSION

We presented *landmark privacy* for privacy-preserving time series publishing, which allows for the protection of significant events while improving the utility of the final result compared to user-level differential privacy. We proposed three schemes for landmark privacy, and quantified the privacy loss under temporal correlation. Furthermore, we designed a module to enhance our privacy notion by protecting the actual timestamps of the landmarks. The contribution of the landmark privacy, and its best performing scheme, Adaptive, is experimentally demonstrated on real and synthetic data sets. The experiments showed that our dummy landmark selection module introduces a reasonable data utility decline to all of our schemes, which is minimal for the Adaptive. In terms of temporal correlation, we observe that under moderate and strong correlation, greater average regular–landmark event distance causes greater overall privacy loss. In the future, we aim to work on automatically learning the initial landmark set by analyzing the input data sets, semantics, and user preferences. We also plan to introduce learning for the tuning of our Adaptive scheme parameters.

REFERENCES

[1] Accessed on October 11, 2021. *Ring*. https://ring.com

[2] Accessed on October 11, 2021. *TousAntiCovid*. https://bonjour.tousanticovid.gouv.fr

[3] Accessed on October 11, 2021. *Waze*. https://waze.com

[4] Nadeem Ahmed, Regio A Michelin, Wanli Xue, Sushmita Ruj, Robert Malaney, Salil S Kanhere, Aruna Seneviratne, Wen Hu, Helge Janicke, and Sanjay K Jha. 2020. A survey of COVID-19 contact tracing apps. *IEEE access* 8 (2020), 134577–134601.

[5] Raed Al-Dhubhani and Jonathan M Cazalas. 2018. An adaptive geo-indistinguishability mechanism for continuous LBS queries. *Wireless Networks* 24, 8 (2018), 3221–3239.

[6] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 901–914.

[7] Jean Bolot, Nadia Fawaz, S Muthukrishnan, Aleksandar Nikolov, and Nina Taft. 2013. Private decayed predicate sums on streams. In *Proceedings of the 16th International Conference on Database Theory*. ACM, 284–295.

[8] Yang Cao and Masatoshi Yoshikawa. 2015. Differentially private real-time data release over infinite trajectory streams. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, Vol. 2. IEEE, 68–73.

[9] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. 2018. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Transactions on Knowledge and Data Engineering* 31, 7 (2018), 1281–1295.

[10] Rui Chen, Gergely Acs, and Claude Castelluccia. 2012. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 638–649.

[11] 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*. ACM, 1375–1388.

[12] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 429–438.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[14] 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*. ACM, 715–724.

[15] 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.

[16] Liyue Fan and Li Xiong. 2014. An adaptive approach to real-time aggregate monitoring with differential privacy. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2014), 2094–2106.

[17] Farhad Farokhi. 2020. Temporally Discounted Differential Privacy for Evolving Datasets on an Infinite Horizon. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 1–8.

[18] Paul A Gagniuc. 2017. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

[19] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. 34–41.

[20] Jingyu Hua, Yue Gao, and Sheng Zhong. 2015. Differentially private publication of general time-serial trajectory data. In *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 549–557.

[21] Manos Katsomallos, Katerina Tzompanaki, and Dimitris Kotzinos. 2019. Privacy, Space and Time: a Survey on Privacy-Preserving Continuous Data Publishing. *Journal of Spatial Information Science* 2019, 19 (2019), 57–103.

[22] 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.

[23] Himanshu Khurana, Mark Hadley, Ning Lu, and Deborah A Frincke. 2010. Smart-grid security issues. *IEEE Security & Privacy* 8, 1 (2010), 81–85.

[24] Meng Li, Liehuang Zhu, Zijian Zhang, and Rixin Xu. 2017. Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences* 400 (2017), 1–13.

[25] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 1–10.

[26] Stephen Makonin. 2018. *HUE: The hourly usage of energy dataset for buildings in British Columbia*. Technical Report.

[27] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, 94–103.

[28] Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 19–30.

[29] Kourosh Meshgi and Shin Ishii. 2015. Expanding histogram of colors with gridding to improve tracking accuracy. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 475–479.

[30] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 2018. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys & Tutorials* (2018).

[31] Jon Russell. 2018. Fitness app Strava exposes the location of military bases. https://techcrunch.com/2018/01/28/strava-exposes-military-bases. Accessed on October 11, 2021.

[32] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. 2019. Interaction data from the copenhagen networks study. *Scientific Data* 6, 1 (2019), 1–10.

[33] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. 2004. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 175–184.

[34] Colin Tankard. 2016. What the GDPR means for businesses. *Network Security* 2016, 6 (2016), 5–8.

[35] Hao Wang and Zhengquan Xu. 2017. CTS-DP: publishing correlated time-series data via differential privacy. *Knowledge-Based Systems* 122 (2017), 167–179.

[36] K Wang, R Chen, BC Fung, and PS Yu. 2010. Privacy-preserving data publishing: A survey on recent developments. *Comput. Surveys* (2010).

[37] Shuo Wang, Richard Sinnott, and Surya Nepal. 2018. Privacy-protected statistics publication over social media user trajectory streams. *Future Generation Computer Systems* 87 (2018), 792–802.

[38] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally differentially private protocols for frequency estimation. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 729–745.

[39] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.

[40] 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.