Designing Differentially-Private Algorithms for Mobile User Trajectories

Xinxin Li*, Ziming Yuan*, Mehtap Yercel*, Viktoriia Zakharova* and Vasanta Chaganti*
* Computer Science Department
Swarthmore College, Swarthmore, PA 19081
Email: privacy@cs.swarthmore.edu

***Abstract*—In this paper we present a quantitative evaluation of state-of-the-art differentially private algorithms aimed at privatizing users' mobile trajectories. We provide guidelines on the design of differentially private (DP) algorithms for trajectory release based on rigorous evaluations of three classes of DP algorithms; (a) Markov-model based, (b) Grid-or clustering based, and (c) Graph-based DP algorithms; using real-world wireless trace data. We identify properties of a mobile trajectory, that when preserved, increase the utility of privatized trajectories for downstream analysis tasks. We also identify deficiencies in current state-of-the-art DP algorithms in preserving spatio-temporal user-privacy. We show using real-world wireless trace data, that most well-known DP algorithms fall short in the following aspects (a) preserving first and second-order statistical measures of utility (b) lack of an appropriate threat model when evaluating DP algorithms for varying privacy budgets, and finally, (c) the lack of error metrics, and downstream use-cases using privatized trajectories when applied to representative, real-world user mobility data.**

I. Introduction

Mobile user trajectories provide a rich source of data, including peoples' daily activities, travel history, health and activity data, and even co-traveller information [1]. Most trajectory data includes Personally Identifiable Information (PII), that if exposed, can cause significant harm to the end-user [1]. Differential Privacy (DP) [2], offers a formal guarantee of privacy by injecting a tunable amount of random "noise" to a query over a sensitive dataset, such that a precise statistical trade-off is met between data-utility and user-privacy.

While there have been many efforts at privatizing mobile user trajectories [1], [3]–[5], we lack both *standardized error metrics* and comparison of their utility for *downstream analysis tasks* in order to compare and evaluate thee DP algorithms.

More critically, there are no *recent*, or *representative* user mobility datasets that have been used to measure the utility (or the privacy afforded) of most DP algorithms. Real-world user mobility exhibits a power-law distribution [6] – users are often present at a small subset of locations, and have a long tail of hardly visited (infrequent) locations.

We apply state-of-the-art DP algorithms to the KTH wireless trace dataset [7] – a dataset that captures fine-grained mobility of users across the KTH campus across one year. We show that most state-of-the-art DP algorithms exhibit poor utility compared to their reported utility across popularly used datasets including the NYC Taxi Dataset, the Geolife dataset and Brinkoff simulator [1].

In the following sections, we provide insights into performance of different classes of DP algorithms for private trajectory release, and guidelines on generating synthetic privatized trajectories that are both representative of real-world user mobility, while also preserving end-user privacy. We discuss a set of on the downstream utility measures and error analysis needed to *benchmark and evaluate* DP trajectory publication algorithms, to select an algorithm best suited for a given downstream analysis task.

In the rest of this paper we provide examples using a representative set of DP algorithms when applied to the KTH wireless trace dataset. We refer the reader to our report for a more extensive set of error analysis describing the privacy-utility trade-offs across a wide-range of DP algorithms [8].

II. Guidelines for Privacy-preserving trajectory publication

Intuitively, DP algorithms applied to privatize mobile user trajectories, aim to publish a privatized and synthetic set of trajectories, that when compared to the sensitive dataset across their first and second-order distributions, are "close" while protecting any one user from re-identification in the dataset. We now present the definition of $\epsilon-$DP in the context of privatized trajectories.

A randomized differentially private algorithm $\mathcal{A}$ provides ϵ-differential privacy ($\epsilon \in \mathrm{R}, \epsilon > 0$) if, for all neighboring datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ of a set of user trajectories which differ by one row (or one trajectory) and for all possible outputs O of $\mathcal{A}$, we have

$$Pr[\mathcal{A}(\mathcal{D}_1) = O] \leq e^{\epsilon} \times Pr[\mathcal{A}(\mathcal{D}_2) = O] \tag{1}$$

Several variations of DP have been proposed, that relax the privacy guarantees of ϵ-DP. Two in-particular that satisfy the composition theorems of ϵ-DP include (ϵ, δ)-DP and zero-concentrated-DP. We refer the reader to [2] and references therein, for these definitions.

A. Trajectory Characterization

In this section, we provide insights on designing DP algorithms that preserve three critical properties of a mobile trajectory – (a) visit counts at each location (b) transition probabilities amongst visited locations and (c) path-length distributions. Below, we elaborate on the trade-offs, across the classes of DP algorithms, across these metrics of interest.

Visit Counts: Providing privatized visit counts can often be sufficient metric for capacity planning, measuring disease spread, and hotspot monitoring [1]. We note that, DP trajectory publication algorithms are a *poor fit* for this task. DP trajectory publication algorithms aim to preserve sequences of visits to locations, using either Markov-based estimates, or grid and clustering based approximations or using graph-based methods [1]. These methods often require nuanced noise-apportioning mechanisms to prevent a single user being re-identified across a sequence of visited locations, and often introduce significantly more noise. As we have shown in previous work, privatized range and count query DP algorithms offer high utility for stringent privacy budgets ($\epsilon = \{1, 5\}$) [9]. *DP Design Guidelines: We make the observation, that splitting the privacy budget(ϵ), to release visit counts at each node, independently of trajectory publication can significantly improve utility.*

Transition Probabilities: As we show in our poster, most DP algorithms to-date, have significantly struggled with providing accurate transition probabilities using real-world traces. Markov-based DP algorithms suffer from a state-space explosion for large L. Grid and clustering based DP algorithms aim to reduce the state-space by grouping or discretizing geographic regions, but as we show using a representative DP algorithm, AdaTrace [4], this often results in the generation of spurious locations that are uniformly distributed across the discretized state-space, and as we show in our poster, often resulting in trivial filtering attacks [1]. Graph-based DP algorithms have orthogonal drawbacks –trajectories often don't easily translate to graphs – time information, and directionality is lost, and it's unclear how a noisy edge weight between two vertices in a graph, can translate to separate distinguishable trajectories.

We additionally provide the following key insight in our poster – even across extremely relaxed privacy budgets $\epsilon = 20$, DP-algorithms that are Markov-based and Grid-based still introduce significant noise or error in the resulting synthetic trajectories. We make the observation that the noise or error introduced is *stochastic modelling error* rather than *tunable DP noise* further increasing the complexity of evaluating these DP algorithms.

DP Design Guidelines: We note that user mobility modeling is a rich and mature field, and rather than introducing modelling error, that is hard to distinguish from tunable DP noise added, DP algorithm design should incorporateLSTM and RNN based trajectory generation algorithms prior to introducing DP noise. Our preliminary results show that LSTM and RNN based trajectory modeling, along with DP noise added at each layer of the RNN provide higher utility compared to Markov-based and Grid-based approaches.

Trajectory Path Lengths: A related goal in trajectory characterization, is preserving path-length distributions. As we show in our poster, most synthetic paths generated by Markov-based and Graph-based algorithms are significantly smaller (by a factor of 3 or more). Apart from the weaknesses identified in the previous section, we also note that most DP algorithms have been evaluated on a constrained set of open-source datasets [1]. We show that these datasets lack consistent measures of fine-grained mobility over a set of users, across time and similar geographic regions. As we show in our poster, the Brinkoff trace generator generates paths, that lack the power-law distribution that characterizes user mobility, and similar issues exist in the NYC Taxidataset and the geolife dataset. *DP Design Guidelines: By leveraging the power-law distribution exhibited by user mobility, we can both reduce the error introduced by the state-space explosion observed in most trajectory modeling, and can also allow us to use less samples (therefore reducing the privacy budget) when training LSTM and RNN-based DP algorithms.*

B. Evaluation

Finally, we note the lack of a standardized set of error metrics used to evaluate DP algorithms. We provide a preliminary set of standardized results in our poster, and show that the lack of both comparable evaluation, and comparable privacy budgets, increases the ambiguity in evaluating the goodness-of-fit of the DP algorithm for the downstream analysis task.

REFERENCES

[1] À. Miranda-Pascual, P. Guerra-Balboa, J. Parra-Arnau, J. Forné, and T. Strufe, "Sok: Differentially private publication of trajectory data," *Proceedings on Privacy Enhancing Technologies*, 2023.

[2] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[3] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *Proceedings of the 2012 ACM conference on Computer and communications security*, 2012, pp. 638–649.

[4] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, "Utility-aware synthesis of differentially private and attack-resilient location traces," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 196–211.

[5] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, "Dpt: differentially private trajectory synthesis using hierarchical reference systems," *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1154–1165, 2015.

[6] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[7] L. Pajevic, G. Karlsson, and V. Fodor, "CRAWDAD dataset kth/campus (v. 2019-07-01)," Downloaded from https://crawdad.org/kth/campus/ 20190701, Jul. 2019.

[8] V. Chaganti, "Vasanta Chaganti: Teaching and Research," https://www.cs.swarthmore.edu/ chaganti/, 2020 (accessed June 25 2022).

[9] J. Langlieb, G. Lee, and V. Chaganti, "Quantifying the privacy-vs-performance trade-offs for fine-grained wireless network measurement data," in *ACM SIGCOMM 2022 Workshop on Network-Application Integration (NAI '22)*, 2022.

Designing Differentially-Private Algorithms for Mobile User Trajectories

Ziming Yuan, Viktoriia Zakharova, Xinxin Li, Mehtap Yercel, Vasanta Chaganti
Swarthmore College

Introduction

Compare utility-vs-privacy trade-offs in state-of-the-art differentially private algorithms for sequential user mobility data.
Goals:

1) maximize accuracy across metrics in Table 1.
2) mitigate reconstruction attacks against trajectories.
3) preserve individuals' spatio-temporal privacy.

Anonymized Trajectories Leak Sensitive Information

- Anonymization schemes expose unique user mobility patterns.
- As shown in Figure 2, anonymized traces are susceptible to linkage and reconstruction attacks.

Dataset: Wireless Campus Mobile Trajectories

- KTH campus in Sweden, Access Point (AP) connection data from January 2014 to January 2015.

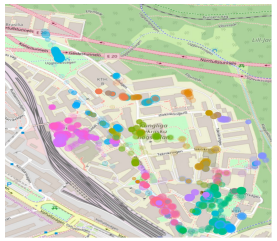

Figure 1: APs on the KTH Campus
(from Crawdad KTH/Campus)

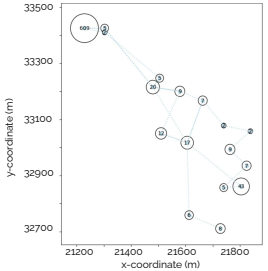

Figure 2: Sample Reconstructed Individual Trajectory

Differential Privacy

- Mathematical guarantee of ε-differential privacy, where ε is the privacy-budget.
- Differential Privacy adds quantifiable "noise" to queries over the data, to mask the presence/absence of an individual row in a database.

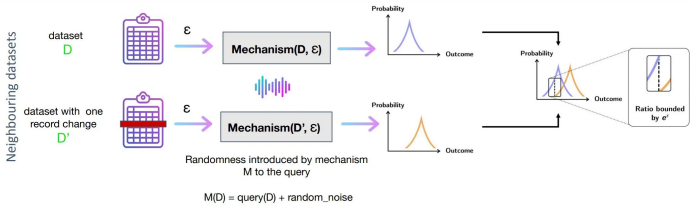

Figure 3. Conceptual Illustration of Differential Privacy
(from Sharing Data with Differential Privacy: a Primer by Anshu Singh)

Overview of Approaches

Markov-based Trajectory Models

Utilizes the Markov assumption and transition probabilities to generate synthetic trajectories.

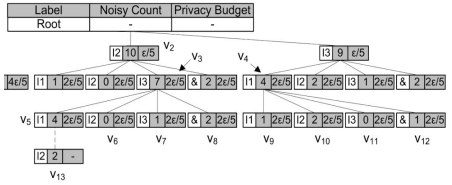

Figure 4. Data Exploration Tree Example used in Markov-based Models [2]

Grid-based Dimensionality Reduction

Transforms geographic regions into multi-scale grids, adding noise to cell counts to privatize trajectories.

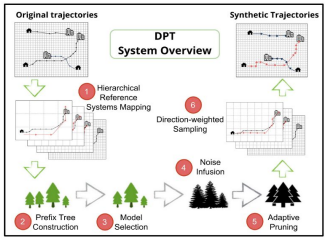

Figure 5. Overview of the DPT Synthetic Generation [4]

Privatizing Points of Interest

Relies on injecting local noise into location input and encrypting that data to share with the data processing algorithm.

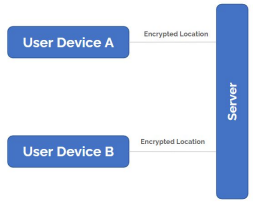

Figure 6. Simplified diagram for zero knowledge range proof using encrypted location information [1].

Comparative Analysis Using Real-world Datasets

DP Algorithm Evaluation on Utility

	NGram Markov-based	AdaTrace Grid-based	Privatizing Points of Interest
Path-length distribution preservation	✖	✖	✖
Accurate visit counts at each location	✖	O	✔
Reconstruction attack	✖	✖	O
Privatized trajectories across multiple instances of the same user	✖	✖	✖

Table 1. Evaluating [2][3][4]

✔ : meets the guideline.
✖ : no attempts and fails to meet the guideline.
O : attempts to meet the guideline but fails.

N-gram Analysis of Trajectory Lengths

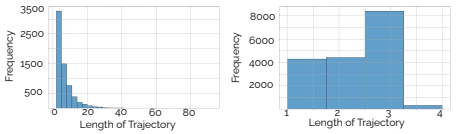

Figure 7. Original Trajectory Lengths

Figure 8. Noisy Trajectory Lengths

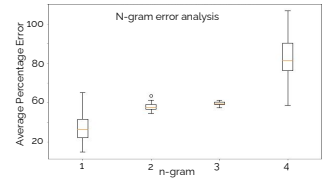

Figure 9. Average Percentage Error per Ngram

N-gram [2] fails to preserve length distribution of KTH trajectories over a single day, with n_max=5, l_max=20, and epsilon=5.

AdaTrace Error Analysis

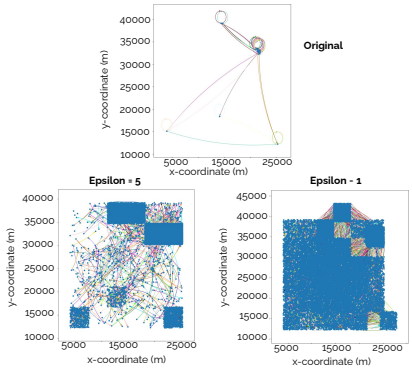

Figure 10. Original (top) and reconstructed (bottom) trajectories of AdaTrace algorithm [3] ran on KTH data for March 3rd, 2014 with epsilons 1 and 5.

Future Work

- Analyze more DP algorithms to refine our utility guidelines.
- Build quantitative evaluation metrics for our guidelines.
- Identify essential algorithmic components that align with our guidelines.
- Use those components to construct a new DP algorithm that satisfies our guidelines.

References

[1] Apple and Google. (2021). Exposure Notification Privacy-preserving Analytics (ENPA) white paper.
[2] Chen, R., Acs, G., & Castelluccia, C. (2012, October). Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security* (pp. 638-649).
[3] Gursoy, M. E., Liu, L., Truex, S., Yu, L., & Wei, W. (2018, October). Utility-aware synthesis of differentially private and attack-resilient location traces. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* (pp. 196-211).
[4] He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C., & Srivastava, D. (2015). DPT: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, *8*(11), 1154-1165.