Practical Differentially Private Modeling of Human Movement Data

Harichandan Roy, Murat Kantarcioglu, Latanya Sweeney

▶ **To cite this version:**

Harichandan Roy, Murat Kantarcioglu, Latanya Sweeney. Practical Differentially Private Modeling of Human Movement Data. 30th IFIP Annual Conference on Data and Applications Security and Privacy (DBSec), Jul 2016, Trento, Italy. pp.170-178, 10.1007/978-3-319-41483-6_13 . hal-01633671

HAL Id: hal-01633671

https://inria.hal.science/hal-01633671

Submitted on 13 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Practical Differentially Private Modeling of Human Movement Data

Harichandan Roy[1] and Murat Kantarcioglu[1,2] and Latanya Sweeney[2]

[1] University of Texas at Dallas, Richardson, USA,
harichandan.roy@utdallas.edu, muratk@utdallas.edu
[2] Harvard University, Cambridge, USA,
kantarcioglu@fas.harvard.edu, latanya@fas.harvard.edu

Abstract. Exciting advances in big data analysis suggest that sharing personal information, such as health and location data, among multiple other parties could have significant societal benefits. However, privacy issues often hinder data sharing. Recently, differential privacy emerged as an important tool to preserve privacy while sharing privacy-sensitive data. The basic idea is simple. Differential privacy guarantees that results learned from shared data do not change much based on the inclusion or exclusion of any single person's data. Despite the promise, existing differential privacy techniques addresses specific utility goals and/or query types (e.g., count queries), so it is not clear whether they can preserve utility for arbitrary types of queries. To better understand possible utility and privacy tradeoffs using differential privacy, we examined uses of human mobility data in a real-world competition. Participants were asked to come up with insightful ideas that leveraged a minimally protected published dataset. An obvious question is whether contest submissions could yield the same results if performed on a dataset protected by differential privacy? To answer this question, we studied synthetic dataset generation models for human mobility data using differential privacy. We discuss utility evaluation and the generality of the models extensively. Finally, we analyzed whether the proposed differential privacy models could be used in practice by examining contest submissions. Our results indicate that most of the competition submissions could be replicated using differentially private data with nearly the same utility and with privacy guarantees. Statistical comparisons with the original dataset demonstrate that differentially private synthetic versions of human mobility data can be widely applicable for data analysis.

Keywords: Differential Privacy, Human Mobility, Utility, Hubway.

1 Introduction

Sharing human related activity data can offer many important benefits to society. For example, mining human mobility data based on cell phone usage can reveal timely information about traffic conditions. "Smart cities" demonstrations show ways human activity data can improve city services. Often, these models require

sharing beyond the person or even the government. A vision is that some of the best possible benefits result from sharing personal activity among organizations. However, the greater the sharing, the greater the risks may be of personal harms. So, privacy concerns may hinder widespread data sharing. Concerns are not unfounded. For example, by correlating location of the individual at a given time of the week, it may be possible to infer someone's religion. Similarly, privacy attacks ranging from stalking to sensitive information disclosure have been widely reported in practice against human mobility data [10], [5].

Of course, personal data can be shared widely if it cannot be personally attributed to a specific person. The idea is that no one can be harmed if his information cannot be isolated in shared data. To address these kinds of privacy challenges, computer scientists have proposed mathematically rigorous techniques in the framework of differential privacy [7]. The main idea in differential privacy is that disclosed results do not change noticeably with the inclusion or exclusion of any given individual's data. Recently, differential privacy has been applied in many different settings ranging from answering basic count queries [2] to building support vector machines [8]. In almost all of these cases, the underlying differential privacy tools are designed for specific use cases and utility is defined and tested for that given use case (e.g., measuring utility for differential private count queries by comparing the euclidean distance between original count query results vs. differentially private results). Usually, it is not clear whether a given approach can support a wide range of uses to which an actual human data scientist may put the data. In this work, we try to understand whether we can provide differentially private synthetic data sets that can be shared instead of an original dataset with confidence that the resulting data will retain utility in different usage scenarios.

One challenge in understanding all the potential uses of a given dataset is that it is impossible to model human imagination. In other words, different data scientists may want to use the data in very different ways. To address this challenge, we look into an existing data set disclosed as a part of the *Hubway Data Visualization Challenge* [1]. *The Hubway* is a public bicycle sharing system with stations throughout Boston, Cambridge, Somerville and Brookline; and it is designed to provide a convenient form of active public transportation by providing access to bicycles. *The Hubway* system stores users' information and generates trip data every day. Hubway data contains users' bike rides history and some personal information, so if it is released publicly or shared with other stakeholders an adversary can take advantage of it and may potentially figure out private information of its target. In 2012, *Hubway* and *Metropolitan Area Planning Council (MAPC)* jointly hosted a challenge named *Hubway Data Visualization Challenge* asking participants to come up with some projects that involve visualizations, animations, artistic representations or interactive data analysis tools. After this challenge, there were reports that some of the disclosed data could be used to identify individuals using location information disclosed on Twitter [9]. Still, submissions to the competition give us a good understanding of what data scientists may want to do with a given human mobility dataset.

To answer the question mentioned above, we propose a model built under differential privacy here. Our model generates differentially private synthetic human mobility dataset from an original dataset that preserves users' privacy. Furthermore, our synthetic dataset also shows a very good accuracy in most important statistical comparisons with the original dataset. Finally, we analyze whether the disclosed differentially private synthetic dataset can adequately provide what data scientists need by analyzing the utility of our disclosed data based on the Hubway challenge submissions. Main **contributions** of this paper are-

- We present a generic *Sanitization Model* built under differential privacy for *resource sharing based human mobility services* to generate differentially private synthetic dataset that preserves users' trip level privacy while sacrificing as little as possible data utility.
- To show the applicability of the generated synthetic data, we compute and compare the most compelling statistics from both synthetic and original datasets. We observe that synthetic data upholds a very impressive accuracy.
- Moreover, a thorough and extensive utility evaluation of synthetic distribution has been done with respect to four different utility metrics.

In section 2, we talk about some preliminaries about differential privacy. Our sanitization model is described in section 3. The experimental evaluation and possible application of it are discussed in section 4 and section 5. Section 6 talks about related works and the conclusion in section 7.

2 Preliminaries

Definition 21. *Differential Privacy [7]*: A privacy mechanism $\mathcal{A}$ gives ϵ-differential privacy if for any database D and $\hat{D}$ differing on at most one record, and for any possible output $O \in Range(\mathcal{A})$,

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

where the probability is taken over the randomness of $\mathcal{A}$.

Definition 22. *Global Sensitivity [7]*: For any function $f : D \rightarrow \mathbb{R}^d$, the $L1$-sensitivity of f is,

$$\nabla f = \max_{D,\hat{D}} \parallel f(D) - f(\hat{D}) \parallel_1 \tag{2}$$

for all D, $\hat{D}$ differing on at most one record.

Theorem 21. Laplace Mechanism [7]*: For any function $f : D \rightarrow \mathbb{R}^d$, and $\epsilon > 0$, the following mechanism $\mathcal{A}$, called the Laplace Mechanism, is ϵ-differentially private: $\mathcal{A}_f(D) = f(D) + \left\langle Lap(\nabla f/\epsilon) \right\rangle^d$.*

3 Sanitization Model

In this section, we propose a *Sanitization Model*, shown in Algorithm 1, which is built under differential privacy. The model takes original dataset, D and privacy budget, ϵ as input and generates ϵ-differentially private synthetic data, $\hat{D}$

as output. *First*, it removes invalid entries and outliers from original dataset applying the statistical $3IQR$ [11] rule to attributes. *Second*, the attributes form some non-disjoint groups based on their associativity. The associativity among the attributes can be examined using well-known *Chi-square Test for Independence/Homogeneity* or *G-Test*. We would like to emphasize that, no statistics are disclosed as a part of this step here. We assume that the grouping of attributes is public information.[3] *Third*, the model builds desired synthetic distributions for each group which is described in section 3.1. *Finally*, the synthetic dataset is generated by taking samples from these distributions and aggregating them together which is discussed in section 3.2.

Algorithm 1 SanitizationModel(D, ϵ)

Input: Original Dataset D, Privacy budget ϵ
Output: Synthetic Dataset $\hat{D}$

1: remove invalid entries and outliers from D
2: $\mathbb{G} = \{<\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m> | (\mathcal{G}_i \cap \mathcal{G}_j) \neq \emptyset; 1 \leq i, j \leq m; \mathcal{G}_i, \mathcal{G}_j \subset \{\mathbb{X} = all_attribute_set\}\}$
3: $\hat{\Phi} = \{<\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_m> | \hat{\phi}_i = diff_private_dist_of(\mathcal{G}_i, \epsilon); 1 \leq i \leq m\}$
4: $\hat{D} = \epsilon$-differentially private synthetic dataset sampling from $\hat{\Phi}$

3.1 Constructing Synthetic Distributions

The first step of constructing a distribution is to create a contingency table (CT) for it. For any particular group $\mathcal{G}$, frequency distribution of all possible distinct combinations of values of all attributes belong to the group represents a CT of that group. Now using the equation mentioned in theorem 21 if we add laplace noise to each frequency, it will become a synthetic CT. In case of negative values, we set them zero since frequency cannot be negative. The frequencies are then normalized to compute respective probability density function.

In our case, hubway dataset has nine attributes- (1) *id* (trip id), I, (2) *start_station_id*, S, (3) *end_station_id*, E, (4) *start_time*, ST, (5) *duration*, L, (6) *end_time*, ET, (7) *zip_code*, Z, (8) *subscription_type*, U and (9) *gender*, G. In step-2 of Algorithm 1, we divide the attributes into five groups: $\mathcal{G}_1$- $\{S, E\}$, $\mathcal{G}_2$- $\{S, ST\}$, $\mathcal{G}_3$- $\{S, E, L\}$, $\mathcal{G}_4$- $\{S, Z\}$, and $\mathcal{G}_5$- $\{Z, U, G\}$. Note that, we do not include ET in any group because it can be calculated from ST and L. For $\mathcal{G}_1$, we first make a CT and after adding laplace noise we convert the resulting synthetic CT finally to CDF which is represented as *Trip CDF*, $\hat{\Phi}_{\mathcal{T}}$. For $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$ and $\mathcal{G}_5$, rather than making a single CDF, we make a number of CDFs for each group instead. More specifically for $\mathcal{G}_2$ and $\mathcal{G}_4$, we build a total of $|S|$ number of CDFs one corresponds to a specific station in S. Likewise, for $\mathcal{G}_5$ we build a total of $|Z|$ number of CDFs one for each zip code. $\mathcal{G}_2$ has an attribute ST which, in essence, is a combination of *year*, *month*, *date* and *hour* sub-attributes (we ignore *min* and *sec* here). Taking these four sub-attributes into account, we build a CDF for each start station in S. The *StartTime CDF* is denoted by $\hat{\Phi}_{\mathcal{ST}}$. For $\mathcal{G}_4$, we build

[3] Since our main focus is slightly different, we skip the discussion about *Chi-square Test for Independence/Homogeneity* and *G-Test* here.

zip distribution for each start station and the *CDF* for this group is denoted by $\hat{\Phi}_{\mathcal{SZ}}$. Similarly, for $\mathcal{G}_5$ we construct *Subscription-Gender* distribution for each zip and it is represented by $\hat{\Phi}_{\mathcal{ZUG}}$. In case of $\mathcal{G}_3$, instead of fitting *duration* into an existing parametric distribution, we build total $||S| \times |S||$ empirical distributions for *Duration* where one corresponds to a particular combination of start and end stations. The reason for building empirical distributions is that they show better results than the fitted parametric *Exponential, Normal and Log-normal* distributions. And we build $||S| \times |S||$ duration distributions rather than only $|S|$ distributions like $\mathcal{G}_2$ and $\mathcal{G}_4$ because duration mainly depends on the distance between two stations. The number of bins is set to 7. These empirical distributions are indeed *CT*s for duration and their corresponding synthetic *CDF* is denoted by $\hat{\Phi}_{\mathcal{L}}$. Note that we add laplace noise to the degree that satisfies the equation stated in Theorem 21 to make synthetic distributions ϵ-differentially private. In all cases, global sensitivity ∇f is 2 since adding or removing or changing an entry can change the function value at most 2.

3.2 Differentially Private Synthetic Data Generation

In this section we will describe how to generate differentially private synthetic data for *The Hubway* from the distributions constructed in section 3.1. Among nine attributes, *id I* is unique in the original dataset. Thus, we assign an unique id for each newly generated trip entry. The steps to generate other attributes of a particular trip, i, as follows: *First*, select a trip (s_i, e_i) by a random sampling from trip *CDF*, $\hat{\Phi}_{\mathcal{T}}$ for i. *Second*, start time for trip i, st_i is randomly sampled from $\hat{\Phi}_{\mathcal{ST}}(s_i)$. Here, $\hat{\Phi}_{\mathcal{ST}}(s_i)$ returns sample from the start time *CDF* of station, s_i. Since it gives *year*, *month*, *date* and *hour* only, we add *min* and *sec* by taking samples from $Uniform(0, 59)$ distribution. *Third*, to get *duration* l_i for trip i, we need to take a random sample from duration distribution of trip (s_i, e_i), $\hat{\Phi}_{\mathcal{L}}(s_i, e_i)$. In this case, each sample taken from $\hat{\Phi}_{\mathcal{L}}(s_i, e_i)$ returns a bin with its start value, a and end value, b. To get the exact value of duration for trip, i, we take a random sample from $Uniform(a, b)$ distribution. By adding l_i to st_i, et_i is calculated accordingly. *Fourth*, we get z_i from a random sample taking from station-zip *CDF* of s_i, $\hat{\Phi}_{\mathcal{SZ}}(s_i)$. *Finally*, a sample (z_i, u_i, g_i) is taken from $\hat{\Phi}_{\mathcal{ZUG}}$ where z_i is restricted to the value computed in step *Fourth*.

4 Experimental Evaluation

In our experiment, we use hubway trip history data released in [1] in February 2014. It contains total of 1029739 entries. For simplicity and without loss of generality, we work on *nine* attributes among them. We will show some aspects of applicability and effective use of synthetic data in this section. Our experiments show that 0.9 is the lowest value of ϵ where we get maximum utility. We run the experiment 20 times and compute the following statistics in each run. In Fig 1, we show the average and standard deviation σ of 20 runs.

For each station, there are two types of trips: *outgoing* and *incoming*. A trip is considered as *outgoing* to its starting station and as *incoming* to its destination. Both are statistics are important in practice and so we study both

Org.		Syn. (0.9)	
Pop. st.	Trips(%)	Avg. Trips(%)	σ
67	2.98	2.98	0.019
22	2.89	2.89	0.020
53	2.21	2.21	0.017
113	2.19	2.19	0.013
36	2.07	2.07	0.017

(a)

Org.		Syn. (0.9)	
Pop. st.	Trips(%)	Avg. Trips(%)	σ
67	3.04	3.04	0.015
22	2.89	2.89	0.020
74	2.22	2.22	0.020
36	2.14	2.14	0.014
113	2.11	2.11	0.012

(b)

Org.		Syn. (0.9)	
Pop. day	Trips(%)	Avg. Trips(%)	σ
2013-10-03	0.56	0.54	0.013
2013-10-02	0.55	0.52	0.011
2013-09-20	0.55	0.52	0.013
2013-09-17	0.54	0.53	0.012
2013-10-18	0.54	0.52	0.013

(c)

Org.		Syn. (0.9)	
Trip	Trips(%)	Avg. Trips(%)	σ
53 - 67	0.47	0.47	0.007
67 - 53	0.46	0.46	0.007
33 - 67	0.25	0.25	0.006
67 - 33	0.24	0.24	0.004
40 - 22	0.20	0.20	0.005

(d)

Fig. 1. Statistical Analysis: (a) Top 5 popular stations (outgoing trips), (b) Top 5 popular stations (incoming trips), (c) Top 5 popular days and (d) Popular Top 5 trip-routes in original and synthetic ($\epsilon = 0.9$) data [20 runs].

cases here. Fig 1(a) shows the *outgoing* trips percentage of top 5 popular stations in original data and their corresponding percentage in synthetic data with σ. As we observe, the percentages of trips shown in the table are identical in both datasets with very low deviation. Similar statistics considering *incoming* trips are shown in Fig 1(b) and it holds similar observation. Besides popular stations, finding popular days is an another essential statistics needed for planning purposes. Fig 1(c) shows the top 5 popular days in original dataset with trips percentage and the corresponding percentage in synthetic data along with their standard deviation. As we see, the percentage of each of the popular days in original and synthetic data is almost same and the corresponding σ is very low as well. Finding popular trip routes is also another statistics that carries important information. In Fig 1(d), we show the top 5 popular trip routes in original data and the corresponding statistics in synthetic datasets. Popularity is measured based on their percentages in entire dataset. The result shows that all top 5 popular trip routes in original data have same percentage of trips in synthetic data as well with very low σ.

We also compute some other statistics but due to space constraint the figures are not shown in the paper. We briefly discuss these statistics here. Comparing the trip percentages in different time periods between two datasets is another important measure for understanding utility. Results show that the noise impact is negligible and in all cases, $Morning$, $Afternoon$ and $Night$, synthetic data preserves the original statistics almost precisely. For example, the difference in morning trip percentage is 0.15 with σ 0.103 only. Gender distribution for each station may be useful in some practical applications (e.g., targeting adds for given stations). We pick few stations randomly to see their gender distributions. According to the results, synthetic data shows promising results in this case as

well. For example, station 67 has gender distribution: Male- 62.68%, Female- 14.04%, *X*- 23.28% in original data and in synthetic data it is: 57.74%, 19.00%, 23.26% with σ 0.25, 0.22 and 0.20 respectively. The subscription distribution per station seems another relevant statistics that has also a significant impact in resource optimizations (e.g., for *Smart City*). The subscription distribution of (*Registered*, *Casual*) for station 67 in original data is (76.72%, 23.28%) and in synthetic data it is (76.74%, 23.26%) with σ 0.20 which is to some extent identical with original statistics. Result is very much alike for other stations as well. However, the comparison between original and synthetic trip duration shows that synthetic data almost accurately measures overall average duration but failed to measure maximum and minimum durations precisely. The result is even worse if we use parametric distribution for duration. We notice in the empirical distribution that a significant number of cells have very low frequency. Due to this fact, a notable noise impact is reflected in the synthetic results.

Furthermore, we study four utility metrics (*Average Relative Error (Avg. RLE)*, *Earth Mover's Distance (EMD)*, *True Positive (TP)* and *Utility Loss (UL)*) to compare synthetic distributions with their original distributions. The results show that the range of *Avg. RLE* is $[0.06 - 0.30]$ with σ range $[0.003 - 0.015]$ for ϵ 0.1 to 1. *EMD*, *TP* and *UL* are $[0.06 - 0.99]$, $[95.25 - 97.5]$ and $[4.41-10.1]$ with σ range $[0.021-0.124]$, $[2.22-3.08]$ and $[2.22-3.03]$ respectively. Due to space constraints, figures are not shown here.

5 Discussion

In 2012, *Hubway* and *Metropolitan Area Planning Council (MAPC)* jointly hosted a challenge named *Hubway Data Visualization Challenge* [1] asking participants to come up with projects that involve visualizations, animations, artistic representations or interactive data analysis tool. It had received total 67 projects that used original data provided by the host. We went through short description and/or little demo provided with each of these projects to find out the statistics that were computed by most of the participants. The comparison of top 7 of these statistics are shown in section 4 and it seems that releasing differentially private data preserves utility in each case.

Due to the the way original data is released, we do not provide user level privacy, our synthetic data provides trip level instead (e.g., sensitivity is computed based on adding or removing one trip, not on adding or removing individual). The synthetic data is also provides more protection than original dataset w.r.t. *Intimate Stalker Threat* [10]. First, unlike the original data set, synthetic data does not release real time visit information. Moreover, it is ϵ-differentially private which means it hides a particular trip information with ϵ privacy. As a result, identity as well as location resolution would be more harder for an intimate stalker using the synthetic data compared to original data.

6 Related Work

Few works [4], [3], [6], [2] have been done on publishing and characterizing human mobility based on cellular network and other spatio-temporal data. All these

papers built their model under Differential privacy. Chen et al. [4] study the problem of publishing trajectory data of commuters in Montreal. In paper [3], authors make use of the variable-length n-gram model. Mir et al. [6] models the human mobility based on Call Detail Records from a cellular telephone network. Acs et al. [2] presents a new anonymization scheme to release the spatio-temporal density of Paris in France. All these papers addressed the specific utility goal. This inspires us to study the possible utility for arbitrary queries.

7 Conclusion

In this paper, we propose a sanitization model for hubway dataset built under differential privacy that preserves users' trip level privacy. To show the applicability and utility of the generated synthetic data for arbitrary range of queries, we compare the most essential and compelling statistics derived from both synthetic and original datasets. Based on the comparison results, we conclude that most of the information required by human analysts can be provided accurately by differentially private synthetic data. We also discuss that the synthetic data release could be used to reduce threats due attacks such as *Intimate Stalker* compared to original data release.

Acknowledgement- The research reported herein was supported in part by NIH awards 1R0-1LM009989 & 1R01HG006844, NSF CNS-1111529, CNS-1228198, CNS-1237235 & CICI-1547324.

References

1. Hubway data visualization challenge. `http://hubwaydatachallenge.org/`, 2012
2. Acs, G., Castelluccia, C.: A case study: Privacy preserving release of spatio-temporal density in paris. KDD (August 2014)
3. Chen, R., Acs, G., Castelluccia, C.: 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. ACM (2012)
4. Chen, R., Fung, B., Desai, B.C., Sossou, N.M.: Differentially private transit data publication: a case study on the montreal transportation system. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 213–221. ACM (2012)
5. Chittaranjan, G., Blom, J., Gatica-Perez, D.: Mining large-scale smartphone data for personality studies. Personal and Ubiquitous Computing 17(3), 433–450 (2013)
6. D. J. Mir, S. Isaacman, R.C.M.M., Wright, R.N.: Dp-where: Differentially private modeling of human mobility. BigData Conference p. 580 to 588 (2013)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography, pp. 265–284. Springer (2006)
8. Li, H., Xiong, L., Ohno-Machado, L., Jiang, X.: Privacy preserving rbf kernel support vector machine. BioMed research international 2014 (2014)
9. Mahmud, J., Nichols, J., Drews, C.: Where is this tweet from? inferring home locations of twitter users. ICWSM 12, 511–514 (2012)
10. Sweeney, L.: Risk assessments of personal identification technologies for domestic violence homeless shelters (2005)
11. WIKIPEDIA: Interquartile range. `http://en.wikipedia.org/wiki/Interquartile_range/`, interquartile range and outliers, 3IQR