Locality Sensitive Hashing of Trajectories Under Local Differential Privacy

Fabrizio Boninsegna[1]

[1]*University of Padova, Department of Information Engineering.*

Abstract

Searching for similar GPS trajectories is a fundamental problem for several applications like frequent routes and ride-sharing recommendations. To face the challenges of large-volume data and the complexity of trajectories various approaches have been developed which all rely on the technique of locality sensitive hashing (LSH), a probabilistic algorithm that maps similar items into the same bucket. However, releasing trajectory data may compromise the privacy of the involved user and LSH does not satisfy the strong requirements of differential privacy (DP). In this proposal paper, we will focus on releasing differential private sketches of trajectories with the scope of finding nearest-neighbor trajectories with theoretical guarantees. We will adapt a recently proposed LSH-DP algorithm to our setting, showing that it has poor utility when the LSH family of functions maps secret items in a discrete space with high cardinality.

Keywords

Differential privacy, Locality sensitive hashing, Trajectories, Nearest neighbors search

1. Introduction

Privacy and Trajectories. Trajectory data can be represented as a tuple (identity, position, time), including the user identity, the spatial information, and the temporal information. Hashing the identity information or completely deleting it does not guarantee privacy, indeed various reconstruction attacks that use side information such as public datasets can be used to infer user identity or home address [1]. Cynthia Dwork in 2006 developed the concept of *differential privacy* [2] (DP), a formal mathematical definition of how an algorithm must be designed in order to guarantee that an adversary cannot distinguish single users in a dataset, regardless of how much external knowledge he/she could have. In other words, a DP mechanism returns outputs such that any individual in the dataset is *indistinguishable* from one another. This theory has been first developed in the context of statistical databases with a trusted data curator, where the problem was to answer queries without revealing private data. When the data curator is untrusted we need to refer to the concept of *local differential privacy*. In this scenario, a random mechanism is applied to user data before they are sent to a data curator. Figure 1 and figure 2 represent the two scenarios. The downside of the local approach is that it is usually necessary to add too much noise, jeopardizing the utility of the released data. Furthermore, in the scenario where a single user is involved, such as in the case of finding the closest trajectories or locations in a public dataset, requiring indistinguishability for each possible user would mean requiring a

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy
fabrizio.boninsegna@phd.unipd.it (F. Boninsegna)

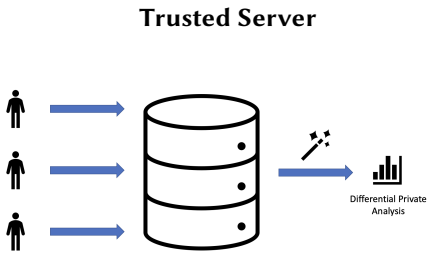


Figure 1: Central differential private model

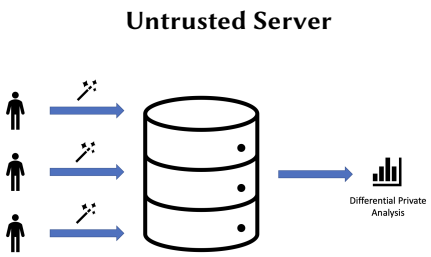


Figure 2: Local differential private model

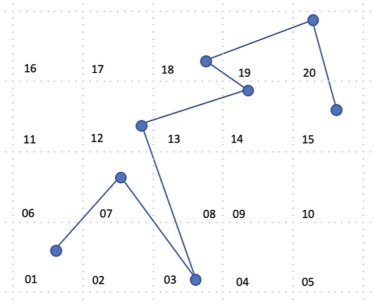
negligible effect on the output for each possible input. With this approach, it would therefore be impossible to provide any useful information.

A step forward in this direction was the introduction of *metric differential privacy* [3], where the notion of indistinguishability was changed from the user to the user's secret inputs. In this setting, a randomized mechanism is designed such that secrets inputs have different level of indistinguishability according to some distance.
Following this definition, our goal is to create a randomized mechanism that processes a trajectory to release a sketch with metric differential privacy guarantees according to the Fréchet distance of curves.

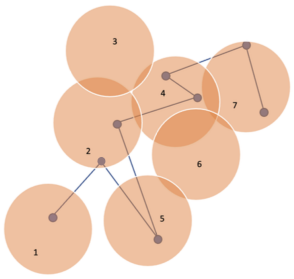
Locality sensitive hashing and curves. Exact nearest neighbor searching of points suffers the curse of dimensionality, hence either linear query time or space exponential to the dimension of the vector, which increases quickly for the trajectories case. The approximate nearest neighbor search problem, which returns points within a certain distance from the query, has been solved using locality sensitive hashing in sublinear query time and polynomial in the dimension [4]. The idea is to construct a family of functions that probabilistically map similar items to the same hash value.

Recently, a LSH family for trajectories was developed in [5]. The method works by snapping vertices of a trajectory into a random grid, ensuring that trajectories with Fréchet distance smaller than the grid size are likely to snap to the same cells. The ordered set of grid cells touched by the trajectory gives the hash values (see figure 3a). Another similar approach was developed in [6] where a LSH family was constructed by deploying a random set of disks with fixed radius on the plane. The hash value (or trajectory sketch) is computed as the sequence of indices of the disks that a trajectory enters and exits (see figure 3b). The trajectory sketches can be used as an index to construct a dataset to efficiently solve approximate nearest neighbor, with similar results of [5], but also efficiently compute distances, clusters, and sub-trajectory similarities.

LSH and DP Even though these sketches reduce the information of a trajectory, still they do not provide strong differential private guarantees [7]. To overcome this problem a randomized mechanism on sketches for the Jaccard similarity has been studied [8]. This idea was further developed in [9] for the random hyperplane projection LSH algorithm proposed by Charikar [10], where the authors provide metric DP guarantees using randomized response on the hashes.



(a) Hashing technique in [5]



(b) Hashing technique in [6]

Figure 3: Two possible ways to create a hash of a trajectory. In the left figure (a) we have the method developed in [5] where the vertices of a trajectory are snapped in a random grid (hash: 0107031214182015). On the right figure (b) we have the method developed in [6] where the hash function returns enters and exists indices of randomly selected disks with fixed radius (hash: 1225524247).

Contributions We extend the analysis in [9] to any LSH algorithms, with the scope to study DP guarantees for trajectories LSH constructions. We also study the quality of the sketches produced by this mechanism, showing that it is necessary to add noise that scales with the logarithm of the cardinality of the sketches. Due to this results, we did not still provide experiments for the nearest neighbor search.

Structure of the Paper In section 2.1 we will give the mathematical definitions of local differential privacy, metric differential privacy, and we will introduce the randomized response mechanism. In section 2.2 we will define locality sensitive hashing. In section 3 we will give our results about the privacy guarantees, while in section 4 we will present an utility analysis. In the conclusion, we will individuate also future research directions.

2. Preliminaries

2.1. Privacy

For the purpose of this paper, it is sufficient to give the definition of *probabilistic* differential privacy, which is more intuitive and implies the standard definition of differential privacy [11].

Definition 1 (Probabilistic Local Differential Privacy)**.** *Given $\varepsilon > 0$ and $\delta \in [0, 1)$, a randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to be (ε, δ) - probabilistic local differential private if, for any two* ***neighboring*** *inputs $x, x' \in \mathcal{X}$ and any range $S \subseteq \mathcal{Y}$:*

$$Pr\left[\frac{Pr[\mathcal{M}(x) \in S]}{Pr[\mathcal{M}(x') \in S]} \leq e^{\varepsilon}\right] \geq 1 - \delta.$$

The parameter ε is called privacy budget.

The privacy budget ε quantifies the level of privacy, the smaller it is the higher the privacy. When *any* two inputs are neighbors, this definition ensures indistinguishability among the users. A mechanism that satisfies $(\varepsilon, 0)$ - local differential privacy is the randomized response.

Definition 2 (Randomized Response). *A randomized response mechanism RR* : $\mathcal{X} \to \mathcal{X}$ *is defined with the following probabilities for all* $x, y \in \mathcal{X}$

$$Pr[RR(x) = y] = \begin{cases} \frac{e^{\varepsilon}}{|\mathcal{X}| - 1 + e^{\varepsilon}} & \text{if } x = y \\ \frac{1}{|\mathcal{X}| - 1 + e^{\varepsilon}} & \text{if } x \neq y \end{cases}$$

where $|\mathcal{X}|$ *is the cardinality of the set.*

It is important to notice that the smaller ε the less informative the above mechanism is. This is true for any DP mechanism, the more privacy we want the more noise we need to inject on the outputs. We now provide the definition of probabilistic extended DP, which is a more general formulation of metric DP, that it allows different level of indistinguishably according to the metric space $(\mathcal{X}, d)$.

Definition 3 (Probabilistic Extended Differential Privacy [3]). *Given two functions* $\xi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ *and* $\delta : \mathcal{X} \times \mathcal{X} \to [0, 1]$*, a randomized mechanism* $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ *provides* (ξ, δ) *- probabilistic extended DP if* $\forall x, x' \in \mathcal{X}$ *and any range* $S \subseteq \mathcal{Y}$

$$Pr\left[\frac{Pr[\mathcal{M}(x) \in S]}{Pr[\mathcal{M}(x') \in S]} \leq e^{\varepsilon^* \xi(x,x')}\right] \geq 1 - \delta(x, x').$$

If $\xi(x, x') = d(x, x')$ and $\delta(x, x') = 0$ we recover the standard definition of metric-DP [3]. With metric-DP, the closer are the inputs the more indistinguishable they are, instead for local-DP the level of indistinguishability is constant for any input. We can relate these two definitions by relaxing the condition on neighboring points. Using a metric-DP mechanism a user can set a definition of neighboring inputs (e.g. $d(x, x') = r$) to get local-DP guarantees within a radius with privacy budget $\varepsilon = \varepsilon^* r$.

2.2. Locality Sensitive Hashing

Definition 4 (Locality Sensitive Hashing). *Consider a metric input space* $(\mathcal{X}, d)$*. Given real values* $r_1 > 0$*,* $r_2 > r_1$*,* $0 \leq p_1 \leq 1$*, and* $0 \leq p_2 \leq 1$*, a family* $\mathcal{H}$ *of hash functions* h *drawn according to a distribution* $D_{\mathcal{H}}$ *is called* (r_1, r_2, p_1, p_2)-sensitive *if for any* $x, x' \in \mathcal{X}$

$$\begin{aligned} &(i) \quad d(x, x') \leq r_1 \Rightarrow Pr_{h_1 \in D_{\mathcal{H}}}[h(x) = h(x')] \geq p_1, \\ &(ii) \quad d(x, x') \geq r_2 \Rightarrow Pr_{h \in D_{\mathcal{H}}}[h(x) = h(x')] \leq p_2. \end{aligned}$$

The LSH family used in [9] has the peculiar property that $\Pr[h(x) \neq h(x')] = d_{\mathcal{X}}(x, x')$ where $d_{\mathcal{X}}$ is a dissimilarity function with image in $[0, 1]$. We will provide a metric DP analysis on the more general definition, which is used to construct LSH for trajectories [5, 6]. To get a more accurate estimate of similarity is usually used LSH amplification which involves using multiple independent concatenations of hash functions to hash each data point.

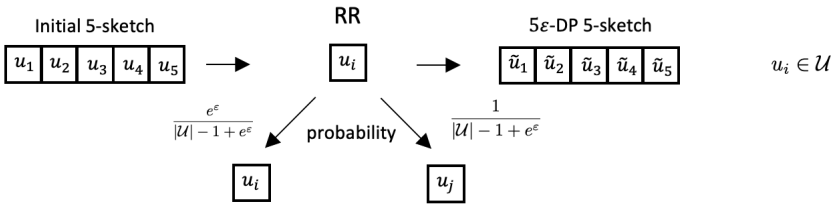


Figure 4: Randomized Response RR applied to 5-concatenation LSH sketch. The RR is applied 5 times independently on each sketch realization

3. Locality sensitive hashing and randomize response

Following the idea of [9], we study how the concatenation of κ hash functions $H = (h_1, \ldots, h_\kappa)$, where each h_i is an LSH function that maps secret inputs in a finite space $h_i : \mathcal{X} \to \mathcal{U}$, changes with κ independent randomized response mechanisms. We will call this algorithm LSH-RR. In [9] the authors demonstrated that the random hyperplane projection LSH family [10] provides $(\xi_\alpha, \delta_\alpha)$ - extended DP

$$\xi_\alpha(x, x') = \varepsilon\kappa(d_\mathcal{X}(x, x') + \alpha),$$
$$\delta_\alpha(x, x') = \exp(-\kappa D_{KL}(d_\mathcal{X}(x, x') + \alpha || d_\mathcal{X}(x, x'))),$$

with $D_{KL}(.||.)$ as the Kullback-Leibler divergence and for any $\alpha > 0$. Practically, a user can decide his/her own level of indistinguishably by setting $d_\mathcal{X}(x, x') = r$ to get a $(\xi_\alpha, \delta_\alpha)$ - local DP mechanism with neighboring points such that $d_\mathcal{X}(x, x') \leq r$. The α parameter can be computed numerically by fixing the probability δ_α.

We found that this is also true for the more general definition of (r_1, r_2, p_1, p_2) - LSH as long as the user agrees in a level of indistinguishably set at $d(x, x') = r_1$.

Theorem 1 (General LSH and RR local privacy guarantees)**.** *Given a (r_1, r_2, p_1, p_2) - LSH family, the LSH-RR algorithm satisfies $(\xi_\alpha, \delta_\alpha)$ - local DP for any neighboring points defined as $d(x, x') \leq r_1$, with*

$$\xi_\alpha(x, x') = \varepsilon\kappa(1 - p_1 + \alpha),$$
$$\delta_\alpha(x, x') = \exp(-\kappa D_{KL}(1 - p_1 + \alpha || 1 - p_1)).$$

With this theorem, it is now theoretically possible to get local DP sketches of trajectories, using the LSH construction proposed in [5, 6] by fixing the cardinality of the hash space, with neighboring distance according to the Frechét distance. In figure 4 is represented the application of randomized response on a general 5-sketch, the result is a 5ε-DP sketch. With theorem 1 we can guarantee higher indistinguishability ξ_α using the same privacy budget ε, ensuring a more accurate RR mechanism.

4. Utility analysis

We would like that the private κ-sketch can still be used to efficiently search nearest neighbors in a dataset indexed by non-private κ-sketches. To achieve this, the private mechanism, which can

be seen as a query transformation in a standard LSH, should not change too much the collision probability of the underlying LSH mechanism. The theory of LSH with query transformation has been developed in [12] where the authors defined the concept of *asymmetric LSH*. The only difference with definition 4 is that $\mathcal{H}$ now represents a family of pairs of functions. For a hash function $h : \mathcal{X} \to \mathcal{U}$ with $|\mathcal{U}| = O(1)$ we proved this theorem

Theorem 2 (Asymmetric LSH-RR with query privacy)**.** *Given $p_0(\varepsilon)$ the probability to release a truthful answer with a local-DP mechanism $\mathcal{M}$, a (r_1, r_2, p_1, p_2) - LSH is mapped to an asymmetric LSH with the function $(\mathcal{M} \circ h, h)$ and $(r_1, r_2, p_1(\varepsilon), p_2(\varepsilon))$ - sensitivity, with*

$$p_i(\varepsilon) = p_i\left(\frac{|\mathcal{U}|p_0(\varepsilon) - 1}{|\mathcal{U}| - 1}\right) + \frac{1 - p_0(\varepsilon)}{|\mathcal{U}| - 1} \qquad \textit{for } i = 1, 2.$$

To test the utility of asymmetric LSH we need to limit the difference between its collision probabilities with those of LSH.

Proposition 1 (Asymmetric LSH-RR utility)**.** *If $p_i > 1/|\mathcal{U}|$ then for any $\delta \in (0, 1]$ we have that*

$$|p_i(\varepsilon) - p_i| \leq \delta \iff p_0(\varepsilon) \geq 1 - \frac{\delta(|\mathcal{U}| - 1)}{p_i|\mathcal{U}| - 1}.$$

For the randomized response mechanism we have $p_0(\varepsilon) = e^\varepsilon/(|\mathcal{U}| - 1 + e^\varepsilon)$, so we can control the utility of the asymmetric LSH only if

$$\varepsilon \geq \ln\left[\left(\frac{p_i}{\delta} - 1\right)|\mathcal{U}| - \left(\frac{1}{\delta} - 1\right)\right].$$

That means we need to set the privacy budget $\varepsilon = \Omega(\log(\mathcal{U}))$, when $p_i/\delta = O(1)$. In the case of the LSH-RR studied in [9] the hash space is $\mathcal{U} = \{0, 1\}$, so it is possible to control the utility without jeopardizing the privacy. In the trajectory case the hash space $\mathcal{U}$ can increase significantly, e.g. for the random grid case [5] if we have a $c \times c$ grid we would get $|\mathcal{U}| = 2^{c^2}$ if we consider trajectories with at most one visit per cell, therefore the utility-privacy trade-off is quite unbalanced.

5. Conclusion and Research Directions

We studied how hash values in the finite domain can satisfy differential privacy using the randomized response mechanism. We further developed the theory introduced in [9] in the more general definition of LSH showing that still it is possible to have metric differential privacy guarantees. In particular, theorem 1 with the trajectory LSH mechanism provided in [5, 6] gives metric differential privacy guarantees for indistinguishably of trajectories according to the Fréchet distance. However, the utility of this mechanism is compromised when the cardinality of the hash space is high.

Further research directions are to replace RR with other LDP mechanism (such as exponential mechanism on the hash values) to enhance utility. We are confident that theorem 1 is true for any LDP mechanism.

Acknowledgments

I wish to thank my supervisor Prof. Francesco Silvestri and my co-supervisor Martin Aumüller. This work is supported by the Ph.D. fellowship DM 352 and partially by the Unimpresa project Big Mobility.

References

[1] J. Krumm, Inference attacks on location tracks, in: Pervasive Computing: 5th International Conference, PERVASIVE 2007, Toronto, Canada, May 13-16, 2007. Proceedings 5, Springer, 2007, pp. 127–143.

[2] C. Dwork, Differential privacy, in: Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33, Springer, 2006, pp. 1–12.

[3] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, Geo-indistinguishability: Differential privacy for location-based systems, in: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, 2013, pp. 901–914.

[4] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, pp. 604–613.

[5] A. Driemel, F. Silvestri, Locality-sensitive hashing of curves, arXiv preprint arXiv:1703.04040 (2017).

[6] M. Astefanoaei, P. Cesaretti, P. Katsikouli, M. Goswami, R. Sarkar, Multi-resolution sketches and locality sensitive hashing for fast trajectory processing, in: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2018, pp. 279–288.

[7] F. Turati, C. Cotrini, K. Kubicek, D. Basin, Locality-sensitive hashing does not guarantee privacy! attacks on google's floc and the minhash hierarchy system, arXiv preprint arXiv:2302.13635 (2023).

[8] M. Aumüller, A. Bourgeat, J. Schmurr, Differentially private sketches for jaccard similarity estimation, in: Similarity Search and Applications: 13th International Conference, SISAP 2020, Copenhagen, Denmark, September 30–October 2, 2020, Proceedings 13, Springer, 2020, pp. 18–32.

[9] N. Fernandes, Y. Kawamoto, T. Murakami, Locality sensitive hashing with extended differential privacy, in: Computer Security–ESORICS 2021: 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4–8, 2021, Proceedings, Part II, Springer, 2021, pp. 563–583.

[10] M. S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, 2002, pp. 380–388.

[11] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Foundations and Trends® in Theoretical Computer Science 9 (2014) 211–407.

[12] A. Shrivastava, P. Li, Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips), Advances in neural information processing systems 27 (2014).