TraBiMap: Reducing Privacy Concerns in Trajectory Analysis with Randomized Data Representations

Paul Walther
School of Engineering and Design
Technical University of Munich
Munich, Germany
paul.walther@tum.de

Xuanshu Luo
School of Engineering and Design
Technical University of Munich
Munich, Germany
xuanshu.luo@tum.de

Martin Werner
School of Engineering and Design
Technical University of Munich
Munich, Germany
martin.werner@tum.de

Abstract

Personal trajectory data, increasingly acquired by more and more GNSS-enabled devices, is currently underutilized as privacy concerns prohibit its comprehensive excavation to improve reaction and planning opportunities for communities and individuals. However, due to their distinctive data structure, existing data anonymization methods are difficult to apply directly to trajectories. Therefore, inspired by recent probabilistic representations of geographic information, we present TraBiMaps, a randomized data structure for trajectories based on Bloom Filters using cryptographic hash functions that can efficiently store and evaluate rasterized trajectories with a high level of individual privacy. We further provide a preliminary privacy analysis of TraBiMap and pose additional research questions in this field.

CCS Concepts

• **Security and privacy** → *Data anonymization and sanitization.*

Keywords

Trajectories, Bloom Filter, GloBiMaps, TraBiMap, Privacy, Anonymity

ACM Reference Format:
Paul Walther, Xuanshu Luo, and Martin Werner. 2024. TraBiMap: Reducing Privacy Concerns in Trajectory Analysis with Randomized Data Representations. In *2nd ACM SIGSPATIAL International Workshop on Geo-Privacy and Data Utility for Smart Societies (GeoPrivacy'24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3681768.3698496

1 Introduction

In recent years, individual trajectory tracking data has increased massively. Most handheld devices (like mobile phones) and vehicles can track movements using GNSS-enabled services more accurately than ever before, offering great potential for thorough mobility analysis. A common research goal here is to obtain underlying movement patterns applicable for future movement prediction [11, 15], benefiting various fields like emergency response, traffic routing, market research, city planning, and more [15].

GeoPrivacy'24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1147-3/ 24/10
https://doi.org/10.1145/3681768.3698496

However, tracking personal trajectories raises privacy concerns, as location data enables inference on, e.g., religious or sexual preferences, habits, and social customs [1, 15]. Various studies showed that linkage and re-identification attacks are often successful in trajectory databases [6, 11]. Meanwhile, several regions worldwide have already set up regulations to ensure data privacy, e.g., the European Union's *General Data Protection Regulation*[1]. Therefore, trajectory data anonymization is essential to prevent threats of privacy violations and avoid potential legal risks [9].

In general, trajectory data is often stored in databases where each record is denoted a *trajectory identity*, a *timestamp*, and a *location*. Due to their high dimension, sparseness, and sequential character, privacy metrics like k-anonymity [19], and ϵ-differential privacy [7] generally used for tabular-like data types are often not directly applicable to trajectories. Existing privacy mechanisms either allow simple adversarial attacks [9, 15] or reduce the utility of anonymized trajectories [15]. Consequently, only limited datasets are published regarding individual mobility trajectories, whereas datasets published are not representative, and utilization after anonymization is low. At the same time, data structures based on randomized representations of geo-data were presented, which use hash encodings and binary representation schemas featuring Bloom Filters for collective storage of complex geospatial data objects, e.g., GloBiMaps [22], which allow for high lossy data compression capability with low error rates.

With this paper, we identify the potential of hash-based data representation for privacy-preserving trajectory analysis and propose further research in this field: We encourage the investigation of cryptographic Bloom Filter representations for storing summarized trajectory information and propose a new data representation for trajectories: the *TraBiMap*. Further, we explain how TraBiMap ensures privacy and anonymity for publishing and processing trajectory data sets with high utility in analysis tasks. To our knowledge, this poses a new research direction with high potential. It combines the principles of k-anonymity, l-diversity, and differential privacy with storing spatial data in randomized representations.

2 Preliminaries

This section provides fundamental background and concepts to facilitate the following elaboration and analysis of TraBiMap.

2.1 General Privacy Concepts

Datasets usually contain different attributes from a privacy perspective. So-called *quasi-identifiers* denote attributes that alone cannot

[1] https://gdpr.eu/, last accessed 22.07.2024

uniquely identify an individual in a dataset. Still, a combination of quasi-identifiers can be used to deanonymize records [1]. In comparison, *sensitive attributes* are data attributes that must be kept confidential and thus need to be anonymized by privacy-preserving algorithm [1].

The most basic privacy measure is *k-anonymity* [19], which tells an individual's quasi-identifiers have to be equivalent to at least $k-1$ other individuals with which they form an equivalence class. Another concept is *l-diversity*, which extends the k-anonymity by measuring how diverse the values of one sensitive attribute are within one equivalence class to avoid the problem that no matter how high the k is, individuals may still be disclosed with additional information [14]. Furthermore, *t-closeness* is introduced as another privacy dimension and describes the concept of ensuring that the distance between the distributions of sensitive attributes in an equivalence class and the complete database is low [13].

In our case, we focus on privacy and anonymity of trajectories. The standard k-anonymity is not directly applicable here as their sequentiality property allow exploitation with additional information [15]. There is not a fixed set of quasi-identifiers and sensitive information, but rather all items, in our case trajectory points, are both, as it depends on the attacker's knowledge of which trajectory point is what. Therefore, Terrovitis and Mamoulis define an additional k^m-anonymity, where m is an attacker's maximum knowledge (the maximum number of deanonymized trajectory points) [20]. Another method is (k, δ)-*anonymity* where spatial cylinder-like structures describe trajectories and thereby anonymize them through this broader spatial extent [1]. The analysis above reveals two additional requirements for anonymizing trajectories: both the number of regions a user visits (*ubiquity*) and the number of users in each region (*congestion*) cannot be too few to allow anonymization[2].

Another way to anonymize the data in a database is the approach of *differential privacy* where an artificial noise is added probabilistically (normally in a Laplacian distribution or with Randomized Responses) [7].

2.2 Privacy Mechanisms for Trajectories

Various privacy mechanisms are described in the literature to apply the above privacy dimensions to trajectories. *Interactive* privacy mechanisms allow the user to query the dataset of the owner and results to those queries are crafted to protect privacy [7]. In comparison, *non-interactive* approaches sanitize, respectively anonymize the dataset before the release, and there is no interaction with the owner necessary after publishing [7]. Furthermore, we differentiate between the *offline mode* of protecting existing data, which was captured upfront and is provided for analysis and *online mode*, where real-time data of moving objects is analyzed in [15].

The most basic concept to avoid deanonymization in databases is *generalization*, e.g., group instances of one quasi-identifier together to make them less explicit [18]. For trajectories, we can conduct this idea by increasing the spatial granularity, e.g. by assigning an area instead of a point for each trajectory step [1]. An approach to solve the consecutive information loss due to generalization is the hierarchy-free multidimensional approach [12]. However, trying to generalize trajectories always results in the curse of dimensionality, as every coordinate of each trajectory point may be seen as a quasi-identifier for an individual [2, 11]. Furthermore, the cardinality of locations is also very high [11], which makes selecting the generalization degree hard to ensure total anonymity [1].

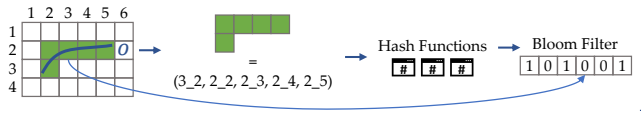

Figure 1: Visualization of the GloBiMaps: The input object o is discretized, hashed, and stored in a Bloom Filter.

Another approach is *supression*, which means deleting all information, which reduce anonymity [18]. For trajectory databases, this would degrade all trajectories outside anonymity sets, which means outliers. An alternative is point suppression, which is the same as a maximum space generalization: the respective suppressed point can be anywhere in space afterward, so it holds no information anymore [2]. Furthermore, *condensation* [3] perturbs the data in a way that preserves inter-attribute correlation to generate new samples but loses the original dataset in the process [1]. Similarly, *space translation* achieve k-anonymity by using a trajectory distance function to cluster trajectories and afterward sample an anonymized set of trajectories from these clusters [1].

Further, *differential privacy* was applied to trajectories [10, 17] and is said to be "the strongest unconditional privacy protection technology currently known" [8]. Issues are that often semantically non-meaningful trajectory points are sampled, so it is difficult to obtain useful mobility patterns for downstream tasks, and this allows for easier identification of artificially sampled points and, therefore, de-anonymization.

2.3 Bloom Filters and GloBiMaps

One of the most representative probabilistic data structures are *Bloom Filters* (BFs), which are used to store sets [4]. A BF in its simplest form consists of a binary array with m slots and a set of k pairwise independent hash functions mapping from the original domain to the range $(0, m-1)$. An empty filter is all zero. For inserting an element into this set description (filter), the element is hashed with all the hash functions, and all slots of the BF denoted by at least one hash value are set to 1. To check whether an element is stored in this BF, again, it is hashed, and if all slots denoted by the hash values are 1, it is returned `true` [21]. An important property of BFs is that there are no false negatives. If a query denotes an item to be stored, it is definitely in the stored set.

To take advantage of BF for trajectory representation, *GloBiMaps* was proposed as a probabilistic data structure for rasterized geometric objects [22]. The main idea is to give all raster cells unique identifiers and store the identifiers of cells covered by a spatially extended object as a set in a BF. The approach is visualized in Figure 1. Especially for sparse geometric objects, it holds an advantage as short BFs can represent complex data with a low false positive rate in limited space [22]. So far, GloBiMaps for trajectory data was limited to the spatial domain only and ignored the time domain [21, 22].

An alternative approach for using BFs with mobility data are *Spatial BFs* to encode spatial information for location-based services

in a specific variant of the BF holding the location information in a set-based format [5, 16]. They explain the privacy-preserving property of this approach in two proposed protocols. However, this is not directly applicable to trajectories.

3 Proposed Methodology

The following presents opportunities for trajectory anonymization by using probabilistic representations, evaluates them on their privacy contribution qualitatively.

3.1 Representing Trajectories with Bloom Filters: TraBiMap

The most straightforward representation of trajectories in the BF-based GloBiMaps representation is to just encode their rasterized spatial footprint as shown in Figure 1 and explained by Werner [22]. While this method might be beneficial to answer queries like *"Has any trajectory touched a specific rasterized cell?"* it does not allow for the differentiation of single trajectories. Such questions like, *"How high is the probability of an individual passing by cell A and cell B?"* cannot be answered. The second question is especially interesting in understanding crowd movements and can help with future planning and adaption to often traveled ways, which might be beneficial in city planning and, e.g., market analysis in a mall. Additionally, this is not an anonymized or privacy-preserving storage format, as the normally used hash functions for BF representations like Murmur hash allow for an inverse of the hashing operation. Therefore, they may allow the decrypting of single trajectories from the global GloBiMaps Footprint of all trajectories.

To improve a BF is assigned to every raster cell, which serves as a set representation of all trajectories T_i passing through the respective cell. This allows to compare the binary BF representations of different cells to detect whether similar trajectories passed by these cells. BF vectors being closer with regards to, e.g., Hamming distance, probably have had similar trajectories passing by. This enables analysis of the second posed question above but still falls short in describing actual movement but rather only location patterns (*"Have similar people been in cell A and cell B?"*).

To make movements analyzable, we propose a new method to represent trajectories with BFs, which we call *Trajectory Binary Map (TraBiMap)* and visualize in Figure 2. Compared to previous approaches, we do not use the rasterized cells as a basis for the probabilistic representation but instead, the crossed cell boundaries (red arrows in Figure 2). This represents a movement instead of a location only. To do so, we define a global raster covering all trajectories with a raster size r, a width of w, and height of h cells (in Figure 2 $w = 6$ and $h = 4$). Given a trajectory dataset $D = \{T_1, ..., T_n\}$ with n trajectories, we can then determine all crossings c of cell boundaries per trajectory $C_{T_i} = \{c_1, ..., c_j\}$. For our representation, we then denote a separate short Bloom Filter $BF_{c_{ab}}$ of length m to every cell crossing c_{ab} between raster cells a and b ($(2,3)$ and $(2,4)$ in Figure 2). We then insert all trajectory IDs T_i into each $BF_{c_{ab}}$ for which this cell crossing c_{ab} is in C_{T_i}. Inserting thereby works as in standard BFs: For each trajectory in D, the trajectory ID T_i is hashed by k hash functions, and results are mapped to the position set $[0, m-1)$. To ensure privacy preservation, we thereby only use cryptographic hashes. In the next step, all BFs for cell crossings in

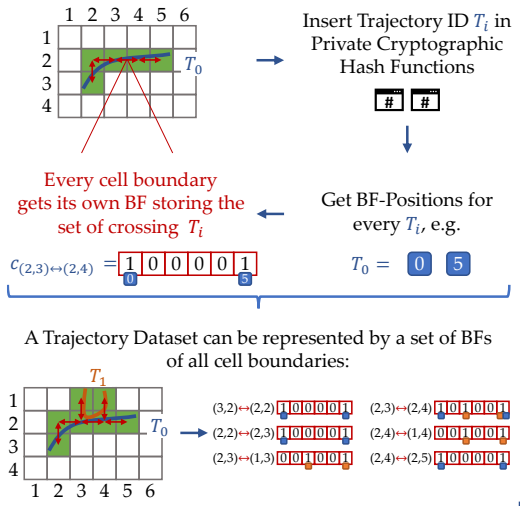

Figure 2: Visualization of the proposed new TraBiMap representation of rasterized trajectories by their cell boundaries, enabling the anonymized representation of whole trajectory datasets in a set of Bloom Filters

C_{T_i} are set to one at these respective positions. After all trajectories are inserted, we end up with $(w-1)*h+w*(h-1)$ BFs for all c_{ab}.

By construction, these binary BF vectors are more similar if the same individuals cross the respective cell boundaries. Therefore, the similarity of crowds that crossed certain cell boundaries may be analyzed by the similarity between these binary vectors. For the similarity, binary vector distance measures can be used. In the future, it needs to be analyzed which measures work best and whether the proposed approach preserves enough utility of the anonymized dataset.

3.2 Privacy Analysis of TraBiMap

In the following, the approach is evaluated against the privacy principles described in Section 2.1. Our approach does not need the dataset to be highly *ubiquitous*, as single users may not be identified in the proposed representation due to cryptographic encoding in BFs. Meanwhile, high *congestion* is necessary for sufficient diversity in the BFs, which ensures privacy. If just one person crosses a cell border, their unique encoding is leaked and can be tracked through the dataset, at least on a probabilistically.

The standard *k-anonymity* is thereby not fully applicable to our approach, as we do not have public individual quasi-identifiers but instead store only the cryptographically encrypted summarized information of all individuals passing a certain cell border. In this case, a single BF can only be seen as a quasi-identifier for the exact subset of all individuals who have passed this border. However, due to construction, even two exactly similar BFs can denote different sets of people with a defined probability due to overlaps in hash values. Therefore, our approach ensures individual privacy as only aggregated information is presented, which cannot be directly decoded into individuals as long as a minimum level of k trajectories goes through every cell boundary. This is somehow similar to the principle of differential privacy. The BFs add a random noise to

the distribution. Still, the proposed approach has no issues with incoherent points, as the BF does not allow for the explicit tracking of single trajectories. Furthermore, in theory, this k-anonymity may also be ensured for a smaller amount of trajectories per cell boundary based on the random representation of values with high likelihood. The sets of people being stored in a BF can then be seen as equivalence classes for *l-diversity*. Conclusively, it needs to be ensured that BF vectors do not allow for adversarial attacks. With enforced variety amongst the BFs encoding of one trajectory at different locations, this individual's other movements cannot be revealed. Approaches are the right filter size to allow for value overlaps in the BF or adding random noise to the BFs. Alternatively, reducing the overall variety of all BFs such that there are too many potential points an individual might have been to prevent revealing this individual's actual movements. This also increases *t-closeness* as more similar filters may occur with different actual initial trajectory IDs being mapped into them. This can be evaluated by checking the spatial distributions of locations with similar BFs.

The proposed representation further still allows for known trajectory IDs and used hash functions to check whether this trajectory passed a certain cell boundary. Therefore, it needs to be ensured to use private hash functions or trajectory IDs, which only the individual or a trusted data publisher knows.

4 Conclusion and Open Research Questions

Our proposed TraBiMap representation allows for an anonymized storage of trajectory datasets which theoretically still allow for a crowd movement analysis. However, for actual usage of this approach, additional research questions have to be answered:

(1) What is the best way to represent trajectories with randomized data structures?
(2) How to choose the right parameters for rasterization, BF sizes, and the number of hash functions for the proposed TraBiMap approach?
(3) How to determine the utility of the data anonymized with randomized data structures?
(4) How do these approaches anonymize real-world datasets?

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 507196470.

References

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *2008 IEEE 24th International Conference on Data Engineering*. IEEE. https://doi.org/10.1109/icde.2008.4497446

[2] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2010. Anonymization of moving objects databases by clustering and perturbation. *Information Systems* 35, 8 (2010), 884–910. https://doi.org/10.1016/j.is.2010.05.003

[3] Charu C. Aggarwal and Philip S. Yu. 2002. A Condensation Approach to Privacy Preserving Data Mining. In *Advances in database technology–EDBT 2004*, Elisa Bertino (Ed.). Lecture Notes in Computer Science, Vol. 2992. Springer-Verlag, New York, 183–199. https://doi.org/10.1007/978-3-540-24741-8{_}12

[4] Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM* 13, 7 (1970), 422–426. https://doi.org/10.1145/362686.362692

[5] Luca Calderoni, Paolo Palmieri, and Dario Maio. 2015. Location privacy without mutual trust: The spatial Bloom filter. *Computer Communications* 68 (2015), 4–16. https://doi.org/10.1016/j.comcom.2015.06.011

[6] Wei Chen, Hongzhi Yin, Weiqing Wang, Lei Zhao, Wen Hua, and Xiaofang Zhou. 2017. Exploiting Spatio-Temporal User Behaviors for User Linkage. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (ACM Digital Library)*, Ee-Peng Lim (Ed.). ACM, New York, NY, 517–526. https://doi.org/10.1145/3132847.3132898

[7] Cynthia Dwork and Aaron Roth. 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3-4 (2013), 211–407. https://doi.org/10.1561/0400000042

[8] Zhen Gu and Guoyin Zhang. 2022. Trajectory Data Publication Based on Differential Privacy. *International Journal of Information Security and Privacy (IJISP)* 17, 1 (2022), 1–15. https://doi.org/10.4018/IJISP.315593

[9] Patricia Guerra-Balboa, Àlex Miranda-Pascual, Thorsten Strufe, Javier Parra-Arnau, and Jordi Forné. 2022. Anonymizing Trajectory Data: Limitations and Opportunities. https://doi.org/10.5445/IR/1000148633

[10] Xi He, Ashwin Machanavajjhala, Cecilia Procopiuc, Divesh Srivastava, and Graham Cormode. 2015. DPT : differentially private trajectory synthesis using hierarchical reference systems. In *Proceedings of the VLDB Endowment*. VLDB Endowment Inc., 1154–1165. https://wrap.warwick.ac.uk/74440/

[11] Fengmei Jin, Wen Hua, Thomas Zhou, Jiajie Xu, Matteo Francia, Maria E. Orlowska, and Xiaofang Zhou. 2022. Trajectory-Based Spatiotemporal Entity Linking. *IEEE Transactions on Knowledge and Data Engineering* 34, 9 (2022), 4499–4513. https://doi.org/10.1109/TKDE.2020.3036633

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. 2006. Mondrian Multidimensional K-Anonymity. In *Proceedings of the 22nd International Conference on Data Engineering, 2006, ICDE '06*, LING LIU (Ed.). IEEE Computer Society, Los Alamitos, Calif., 25. https://doi.org/10.1109/ICDE.2006.101

[13] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE Service Center, Piscataway, NJ, 106–115. https://doi.org/10.1109/ICDE.2007.367856

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L -diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 3. https://doi.org/10.1145/1217299.1217302

[15] Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, Jussara Almeida, Taylor Anderson, Walid Aref, Gennady Andrienko, Natalia Andrienko, Yang Cao, Sanjay Chawla, Reynold Cheng, Panos Chrysanthis, Xiqi Fei, Gabriel Ghinita, Anita Graser, Dimitrios Gunopulos, Christian S. Jensen, Joon-Seok Kim, Kyoung-Sook Kim, Peer Kröger, John Krumm, Johannes Lauer, Amr Magdy, Mario Nascimento, Siva Ravada, Matthias Renz, Dimitris Sacharidis, Flora Salim, Mohamed Sarwat, Maxime Schoemans, Cyrus Shahabi, Bettina Speckmann, Egemen Tanin, Xu Teng, Yannis Theodoridis, Kristian Torp, Goce Trajcevski, Marc van Kreveld, Carola Wenk, Martin Werner, Raymond Wong, Song Wu, Jianqiu Xu, Moustafa Youssef, Demetris Zeinalipour, Mengxuan Zhang, and Esteban Zimányi. 2024. Mobility Data Science: Perspectives and Challenges. *ACM Transactions on Spatial Algorithms and Systems* 10, 2 (2024), 1–35. https://doi.org/10.1145/3652158

[16] Paolo Palmieri, Luca Calderoni, and Dario Maio. 2015. Spatial Bloom Filters: Enabling Privacy in Location-Aware Applications. In *Information Security and Cryptology*, Moti Yung, Jianying Zhou, and Dongdai Lin (Eds.). SpringerLink Bücher, Vol. 8957. Springer, Cham, 16–36. https://doi.org/10.1007/978-3-319-16745-9{_}2

[17] Sina Shaham, Gabriel Ghinita, Ritesh Ahuja, John Krumm, and Cyrus Shahabi. 2023. HTF: Homogeneous Tree Framework for Differentially-Private Release of Large Geospatial Datasets with Self-Tuning Structure Height. *ACM Transactions on Spatial Algorithms and Systems* 9, 4 (2023). https://doi.org/10.1145/3569087

[18] Latanya Sweeney. 2002. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 571–588. https://doi.org/10.1142/S021848850200165X

[19] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570. https://doi.org/10.1142/S0218488502001648

[20] Manolis Terrovitis and Nikos Mamoulis. 2008. Privacy Preservation in the Publication of Trajectories. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*. IEEE. https://doi.org/10.1109/mdm.2008.29

[21] Martin Werner. 2015. BACR: Set Similarities with Lower Bounds and Application to Spatial Trajectories. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Mohamed Ali (Ed.). ACM, 1–10. https://doi.org/10.1145/2820783.2820802

[22] Martin Werner. 2021. GloBiMapsAI: An AI-Enhanced Probabilistic Data Structure for Global Raster Datasets. *ACM Transactions on Spatial Algorithms and Systems* 7, 4 (2021), 1–24. https://doi.org/10.1145/3453184

Received 10 September 2024; accepted 29 September 2024