Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 369–378
https://doi.org/10.54941/ahfe1007191

Privacy-Preserving Human Mobility Clustering With Self-Organizing Trees

Kade Shoemaker, Theodore Tourneux, Priya Naphade, Brandon Kelly, Corey Ducharme, and Steve Hardy

Deloitte Consulting LLP, USA

ABSTRACT

The rapid growth of mobility data from phones, sensors, and connected systems has made it easier than ever to track and analyze how people move in the real world. This data can drive smarter decisions in urban planning, public health, and commercial services. At the same time, it raises tough privacy trade-offs. Individuals can be identified from even sparse data, especially with recent Trajectory User Linking (TUL) methods. In this paper, we implement a user segmentation algorithm for human mobility data designed to cluster individuals based on geospatial pattern-of-life data while censoring all personally identifiable information. Our algorithm addresses known privacy issues motivated by TUL approaches by extending a recent self-organizing tree model to represent a population of user trajectories rather than individual trees per user. This provides a hierarchical structure of user patterns of life across different geographical locations without exposing sensitive location details. Our findings indicate this method provides accurate clustering representations while balancing user privacy.

Keywords: Trajectory user linking, Mobility clustering, Privacy

INTRODUCTION

The increase of mobile devices, applications, social media, Internet of Things (IoT) sensors, and Global Positioning System (GPS) technology has led to an unprecedented volume of human spatiotemporal data (Cats, 2024; Ebrahimpour et al., 2020) offering new opportunities for modelling human behaviour. Previous foundational work has shown that human movement is highly regular and distinctive, with individuals returning to a small set of frequently visited locations (González et al., 2008). This regularity enables segmentation by mobility patterns but also introduces privacy risks, as even sparse trajectories can be re-identified with minimal outside information (de Montjoye et al., 2013).

Formal privacy frameworks, including geo-indistinguishability and k-anonymity, have been proposed to address these concerns (Andrés et al., 2013). Differential privacy, statistical noise addition, and synthetic data generation have been explored to protect user identity while enabling aggregate pattern extraction (Jiang et al., 2013; Rao et al., 2020; Rui et al., 2011). However, recent advances in TUL have further increased privacy risk as users anonymized in one dataset are now able be linked across other datasets, compromising privacy (Gao et al., 2017; Zhou et al., 2018).

Received September 5, 2025; Revised November 28, 2025; Accepted December 8, 2025; Available online April 1, 2026

RELATED WORK

Trajectory-User Linking

The majority of recent TUL techniques have focused on deep learning approaches. In general, these approaches implement advances in representational learning techniques to learn embeddings for trajectories, users, and locations. MainTUL (mutual distillation) (Chen et al., 2022) and AttnTUL (hierarchical attention) (Chen et al., 2024), combine recurrent and attention-based encoders to extract multi-level spatio-temporal features. Hierarchical graph contrastive methods ((HGTUL (Chang et al., 2025), ScaleTUL (Zhang et al., 2025)) attempt to capture complex inter-trajectory relations and improve robustness under sparse data. All these deep learning methods have demonstrated issues with complexity, cost, and scalability tradeoffs.

In contrast, simple and scalable methods explicitly designed for large scale application on inexpensive hardware have also been proposed. Najjar and Mede (2022) proposed a method based on simple heuristics and efficient compression of the data applied to 100k users. Christensen et al. (2024) proposed a method based on self-organizing trees with a compressed encoding scheme evaluated on more than 250k users.

Privacy

The intersection of geospatial analytics and privacy has been extensively explored in recent years, particularly with the growth of location-based services and mobile tracking. The concept of k-anonymity, originally formulated by Sweeney (2002), provides a systematic framework designed to ensure that individual records cannot be uniquely identified in released datasets, often by generalizing or suppressing identifying attributes. Building on this, (Monreale et al., 2010) specifically adapted k-anonymity for movement and trajectory data, introducing methods that generalize location paths so that each anonymized trajectory is indistinguishable from at least k–1 others, thus safeguarding individual privacy while maintaining analytical utility.

Reconstruction error is widely used for evaluating models across many domains, including human behavior and trajectory prediction. Zheng et al. (2014) highlight reconstruction error and related metrics as standards for assessing the accuracy of inferred mobility patterns, emphasizing their importance in urban computing and smart city research. Krumm and Horvitz (2006) further exemplify this by employing reconstruction error to measure the divergence between predicted and actual agent destinations from partial movement trajectories, providing a concrete benchmark for the efficacy of trajectory inference models.

METHOD

We adapt a per-user location encoding scheme from prior work (Christensen et al., 2024) on trajectory user linking and leverage it to enable privacy-preserving patterns across a population.

This encoding scheme provides several essential properties:

Location Invariance: Users are encoded relative to their own behavior, enabling cross-geographic pattern discovery.

Privacy-Enabling: Raw location identifiers are never used in the tree. Encoded data cannot reveal physical locations without the per-user mapping f_u.

Pattern of Life Preservation: Rankings emphasize habitual patterns (home, work) over rare visits.

The motivation for the usage of this encoding method is since mobility patterns are often behaviorally similar even when geographically distinct. For example, two users may both follow a "home → work → home" pattern despite living and working in different cities.

Per-user encoding solves this by ranking locations relative to each user's behavior rather than using global location identifiers. A user's most-visited location is encoded as 0, their second-most as 1, and so on, allowing users with similar routines but different geographic locations to produce identical encoded patterns that the tree can cluster together.

For each user u_i, we construct a personalized location ranking based on visit frequency. Given user u_i's check-in history at locations $\mathcal{L}$ and timeframe $\mathcal{T}\ \left\{(\ell,t)\mid \ell\in\mathcal{L},t\in\mathcal{T}\right\}$:

1. Count visits to each unique location ℓ
2. Sort locations by visit count in descending order
3. Assign rank: most visited → 0, second most → 1, etc. (Ties are assigned distinct ranks at random)
4. Set all ranks above a max value L to L itself.

This approach can be defined a mapping function $f_u:\mathcal{L}\rightarrow\{0,1,\ldots,L,L+1\}$ where $\mathcal{L}$ is the set of all locations ℓ and L is a fixed maximum:

$$f_u(\ell)=\begin{cases}\operatorname{rank}_u(\ell) & \text{if } \operatorname{rank}_u(\ell)<L\\ L & \text{otherwise}\end{cases}$$

For example, if $L=2$, all locations beyond the first and second most common (ranks 0 and 1 respectively) are set to rank 2 to represent all "other" locations. Figure 1 demonstrates this transformation.

Over time, we discretize into T fixed-length bins for each unique day a user is observed (we use $T=48$ for 30-minute bins over 24 hours). In cases when there are no observations for a given time bin, we assign the user to a special "missing" location at $L+1$. When multiple check-ins occur in the same time bin, we average the one-hot vectors and normalize so the total of each time bin along the location dimensions sums to 1 to avoid overweighting heavy check-in user trajectories. Each day d for user u is represented as matrix:

$$\mathbf{X}_d^{(u)}\in\mathbb{R}^{T\times(L+2)}$$

where each row encodes activity during one time bin and each column represents the location of that activity from 0 to L+1.

We also leverage the same self-organizing tree algorithm from (Christensen et al., 2024); but instead of building individual trees per user, we apply the same algorithm but construct a single tree for the user population, repurposing the technique from a TUL algorithm to a method of encoding a population's pattern of life in a privacy preserving way. An example of a leaf node representation for a user trajectory can be seen in Figure 2.

This approach offers a distinct advantage over traditional differential privacy (DP) methods commonly used in geospatial data (Jiang et al., 2013). While DP relies on injecting random noise to obfuscate individual trajectories—often trading off analytical accuracy to boost privacy—our approach eliminates the need for noise altogether. Instead, privacy is achieved by discarding any reference to absolute geographic locations. All location data is represented solely as usage-based ranks unique to each user. The critical mapping from these ranks back to physical places is never stored or shared, meaning that reconstructed trajectories expose *patterns of life* and behavioral rhythms (such as "home → work → home") but cannot be traced to real-world addresses or individuals by any third party. In addition, applying the self-organizing population tree on top of these per-user encodings further obscures individual-level patterns by grouping similar routines together, providing an effective clustering utility for population analysis.

This approach maintains high fidelity with respect to daily routines and temporal activity-supporting tasks like anomaly detection, behavioral clustering, or infrastructure planning while fundamentally blocking reidentification attacks. As a result, practitioners can safely analyze and share aggregate movement features that preserve actionable structure for population-level insights with no risk of reconstructing personal whereabouts or identities.

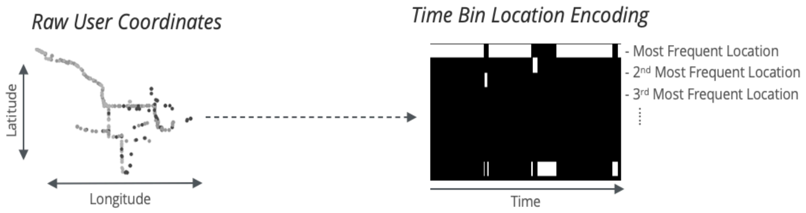
Raw User Coordinates
Latitude
Longitude
Time Bin Location Encoding
- Most Frequent Location
- 2nd Most Frequent Location
- 3rd Most Frequent Location
⋮
Time

Figure 1: Example of encoding from raw coordinates.

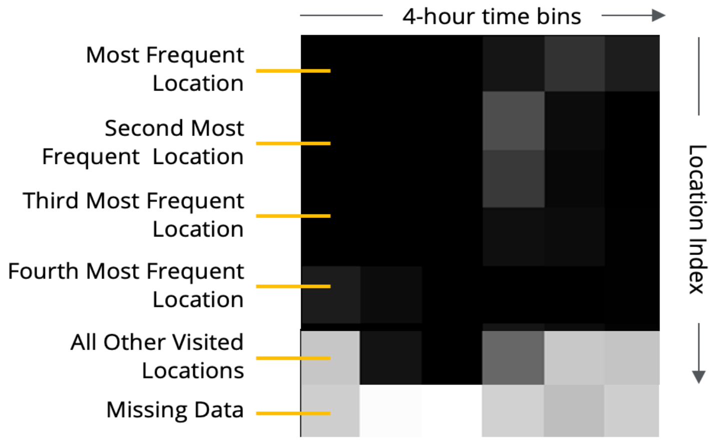

Figure 2: Notional example node representing a day of check ins for the population with L = 5 andT = 6.

EXPERIMENTAL DESIGN

To evaluate the effectiveness of our approach, we measure reconstruction error to quantify the fidelity of our population tree, and k-anonymity to assess privacy guarantees over several open datasets.

Datasets

Like the previous works of (Christensen et al., 2024; Najjar and Mede, 2022), we use 4 open-source datasets that are commonly used in TUL evaluation: Gowalla, Brightkite, Weeplaces, and Foursquare. Each dataset contains records of users' location observations at specific times. We group trajectories at daily timescales and filter out any users with less than 5 trajectories or 10 total check-ins. Periods of time when a trajectory has no check-ins is mapped to the aforementioned "missing data" dimension $L+1$.

We perform an 80:20 test train split on a temporal level, where both the test and train sets of data contain the same users, but train contains the first 80% of check-ins ordered by occurrence and test contains the final 20% of check-ins.

Evaluation

To assess both the privacy guarantees and utility of our population tree approach, we use 3 complementary metrics: **reconstruction error**, **k-anonymity,** and **privacy-utility tradeoff.** We report these numbers from a seeded run but validate each with at least two identical runs with different seeds to ensure rough consistency in our reported metrics.

Reconstruction error measures how well tree node prototypes represent the actual mobility data assigned to them. As the dimensions of a trajectory can change depending on the values of T and L, we compute both a Euclidean difference metric for interpretability and a non-Euclidian difference metric to be dimension size agnostic to avoid overweighting smaller values of L.

We compute the L2 Norm as a Euclidean metric. For a day d assigned to node n, we compute the average the normalized error across all T time bins.:

$$\text{error}(d,n) = \frac{1}{T}\sum_{t=1}^{T} \| d[t,:L+1] - n.\text{data}[t,:L+1] \|_2$$

where:

$\boldsymbol{d[t,:L+1]}$ represents all feature dimensions for time bin t excluding "missing" locations

***n*.data** is the node's data (mean pattern)

$\|\cdot\|_2$ is the Euclidean (L2) norm.

For our non-Euclidean metric, we use Kullback–Leibler (KL) Divergence $D_{KL}(p\,||\,q)$. Given a user's mobility pattern for a single day represented as a probability distribution p over Llocations across T time bins, and its assigned tree node prototype represented as probability distribution q, we compute the KL divergence as:

$$D_{KL}(p\,||\,q) = \frac{1}{T}\sum_{t=1}^{T}\sum_{l=0}^{L} p_{t,l} \log \frac{p_{t,l}}{q_{t,l}}$$

where

$p_{t,l}$ is the known probability of a user being at location l at time bin t

$q_{t,l}$ is the estimated probability of a user being at location l at time bin t given by the best fit leaf node.

Both $p_{t,l}$ and $q_{t,l}$ are normalized over L with a small addition of $\varepsilon = 10^{-10}$ to prevent numerical instability.

We measure **k-anonymity**(ℓ): for each leaf node ℓ, as the number distinct users that have had daily trajectories mapped to the leaf node, and indistinguishable from other users' trajectories that cluster. A higher k-value at a leaf means a user's encoded routine is shared with more peers, making it more difficult for an adversary to isolate or re-identify anyone from that node. We use the count of unique **users** over user days, as a node with only one user's trajectory days assigned to it can be easily mapped to the user itself.

Finally, we evaluate the fundamental tension that stronger privacy guarantees (higher k-anonymity) often reduce utility (higher reconstruction error). We quantify this trade-off by analyzing the relationship between k-anonymity and reconstruction error across leaf nodes with the **Pearson correlation coefficient** $\boldsymbol{\rho}$ between

k-anonymity(ℓ): Number of unique users who have trajectories mapped to ℓ

mean_error(ℓ): Average L2 Norm error for all days mapped to ℓ. We can use a Euclidean metric as all leaf nodes are the same dimension, and thus comparable.

We calculate the ρ as

$$\rho = \text{corr}\left(\left\{k\text{-anonymity}\left(\ell\right)\right\}, \left\{\text{mean\_error}\left(\ell\right)\right\}\right)$$

and interpret the values in the following manner

Positive correlation ($\rho > 0$)**:** Higher privacy → higher error (standard tradeoff)

Negative correlation ($\rho < 0$)**:** Higher privacy → lower error (natural clusters where diverse users share genuine patterns)

Weak correlation ($\rho \approx 0$)**:** Privacy and utility are somewhat independent

RESULTS

We evaluate our population-based TUL approach on four standard mobility datasets (Brightkite, Gowalla, Weeplaces, Foursquare). For each dataset, we select a subset of 400 users and assess model performance against our evaluation metrics. Hyperparameters are varied, specifically testing tree depths of 4, 5, and 6, as well as L-values ranging from 2 to 6. This process identifies the hyperparameter configurations per dataset. To contextualize our results, we benchmark our approach against DBScan, a standard clustering algorithm in geospatial analysis. Additionally, we conduct privacy-utility trade-off experiments, analyzing how model utility changes as tree depth increases for each dataset.

We observe in our results in Table 1 that our population-based TUL approach significantly outperforms DBScan across all four mobility datasets on our selected metrics. Specifically, our method achieves much higher k-anonymity as DBScan struggles to handle unique user-days. This indicates our approach allows for stronger privacy guarantees, while also yielding lower reconstruction errors (both L2 norm and KL-divergence), reflecting better data fidelity.

Table 1: Benchmark against DBScan for 400 users at optimal hyperparameters.

Dataset	Approach	Optimal Tree Hyper-Params		K-Anonymity (Test)		Reconstruction Error (Test)	
		Tree Depth	*Distinct Locations*	*Min*	Median	*L2 Norm*	*KL-Divergence*
BrightKite	*Tree (Ours)*	5	4	**28**	84	**1.28**	**0.75**
	DBScan	-	-	2	400	1.91	0.81
Gowalla	*Tree (Ours)*	5	3	**30**	73	**1.30**	**0.73**
	DBScan	-	-	1	400	1.87	1.02
Weeplaces	*Tree (Ours)*	5	2	**62**	110	**1.18**	**0.65**
	DBScan	-	-	1	400	1.89	0.96
Foursquare	*Tree (Ours)*	5	6	**9**	75	**1.22**	**0.63**
	DBScan	-	-	1	400	1.77	0.74

The results in Table 2 illustrate **low** privacy-utility trade-off as tree depth is varied across all evaluated datasets. Within all nodes for a given tree, we see that the Pearson correlation between the number of unique agents in each leaf node and reconstruction error is minimal, or even negative. This indicates that smaller clusters (leaf nodes with fewer users) do not experience higher error, and that our model achieves consistently strong performance regardless of leaf population size. Additionally, the scaling of minimum k-anonymity in line with the number of leaf nodes reflects that our approach produces highly balanced clusters—privacy gains are distributed evenly, rather than concentrated in a few large groups.

Table 2: Experiment results by dataset and tree depth at L = 4.

Dataset (N Users)	Tree Depth	K-Anonymity (Test)	Reconstruction Error (Test L2 Norm)	Privacy Utility Trade-Off (ρ)
BrightKite (8,733)	5	1,529	1.29	–0.45
	4	2,843	1.36	–0.37
	3	4,527	1.42	0.11
Gowalla (17,112)	5	2,757	1.29	0.32
	4	5,292	1.33	0.31
	3	8,833	1.39	0.28
Weeplaces (12,758)	5	3,257	1.32	–0.33
	4	5,568	1.37	–0.16
	3	8,359	1.43	–0.26
Foursquare (10,000)	5	1,472	1.17	–0.11
	4	2,798	1.19	0.03
	3	5,171	1.26	–0.13

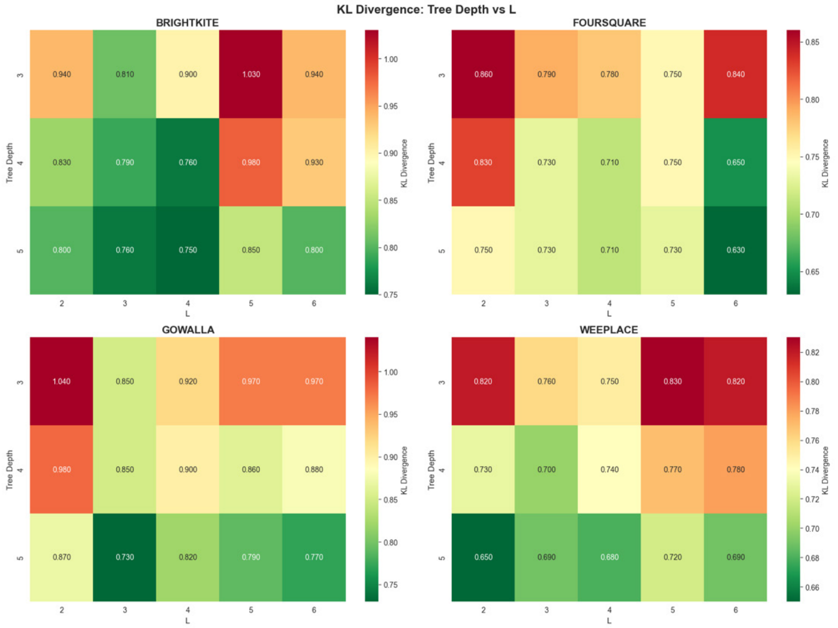

Figure 3: KL-divergence for different values of L and tree depth. We see that performance increases with larger tree depth, but the optimal L is different for each dataset.

CONCLUSION

Overall, our findings demonstrate that our method affords flexible and fair control over privacy and utility, enabling practitioners to tune model parameters to suit user mobility privacy needs and downstream analytic needs. Users benefit from uniformly from better privacy guarantees and consistent performance, even as clustering becomes more granular. The combination of increased privacy (as measured by k-anonymity) without lower reconstruction error suggests that representing user trajectories with population tree leaf nodes can support a range of downstream pattern-of-life analytical tasks, while offering meaningful privacy enhancements. Potential applications include identifying peak activity periods, distinguishing typical mobility archetypes (such as commuters, travelers, or local visitors) for infrastructure planning, and estimating population-level flow patterns.

It is important to note that, while our clustering approach substantially raises the barrier for some privacy risks, a detailed examination of the types of potential privacy attacks and the specific protections afforded is beyond the current scope. Further investigation will be needed to more precisely characterize adversarial risks and quantify how these privacy measures reduce real-world re-identification or inference threats. Despite the strong clustering performance observed, the current approach is limited by the fixed, uniform depth of our population tree and by the absence of adaptive pruning. To further enhance performance and flexibility, future work will focus on developing pruning strategies that assess the similarity between child node pairs. This may allow us to reduce the total number of leaf nodes while maintaining strong pattern-of-life representation across broad user populations.

ACKNOWLEDGMENT

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication. As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

REFERENCES

Andrés, M.E., Bordenabe, N.E., Chatzikokolakis, K., Palamidessi, C., 2013. Geo-indistinguishability: differential privacy for location-based systems, in: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13. Association for Computing Machinery, New York, NY, USA, pp. 901–914. https://doi.org/10.1145/2508859.2516735

Cats, O., 2024. Identifying human mobility patterns using smart card data. Transport Reviews 44, 213–243. https://doi.org/10.1080/01441647.2023.2251688

Chang, F., Zhu, X., Hu, Z., Qin, Y., 2025. HGTUL: A Hypergraph-based Model for Trajectory User Linking. https://doi.org/10.48550/arXiv.2502.07549

Chen, W., Huang, C., Yu, Y., Jiang, Y., Dong, J., 2024. Trajectory-User Linking via Hierarchical Spatio-Temporal Attention Networks. ACM Trans. Knowl. Discov. Data 18, 85:1-85:22. https://doi.org/10.1145/3635718

Chen, W., Li, S., Huang, C., Yu, Y., Jiang, Y., Dong, Y., 2022. Mutual Distillation Learning Network for Trajectory-User Linking. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/274

Christensen, E., Shoemaker, K., Hardy, S., 2024. Trajectory User Linking With Self Organizing Trees, in: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection, GeoAnomalies '24. Association for Computing Machinery, New York, NY, USA, pp. 28–31. https://doi.org/10.1145/3681765.3698450

de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd: The privacy bounds of human mobility. Sci Rep 3, 1376. https://doi.org/10.1038/srep01376

Ebrahimpour, Z., Wan, W., Velázquez García, J.L., Cervantes, O., Hou, L., 2020. Analyzing Social-Geographic Human Mobility Patterns Using Large-Scale Social Media Data. ISPRS International Journal of Geo-Information 9, 125. https://doi.org/10.3390/ijgi9020125

Gao, Q., Zhou, F., Zhang, K., Trajcevski, G., Luo, X., Zhang, F., 2017. Identifying Human Mobility via Trajectory Embeddings., in: IJCAI. pp. 1689–1695.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782. https://doi.org/10.1038/nature06958

Jiang, K., Shao, D., Bressan, S., Kister, T., Tan, K.-L., 2013. Publishing trajectories with differential privacy guarantees, in: Proceedings of the 25th International Conference on Scientific and Statistical Database Management, SSDBM '13. Association for Computing Machinery, New York, NY, USA, pp. 1–12. https://doi.org/10.1145/2484838.2484846

Krumm, J., Horvitz, E., 2006. Predestination: Inferring destinations from partial trajectories, in: International Conference on Ubiquitous Computing. Springer, pp. 243–260.

Monreale, A., Andrienko, G.L., Andrienko, N.V., Giannotti, F., Pedreschi, D., Rinzivillo, S., Wrobel, S., 2010. Movement data anonymity through generalization. Trans. Data Priv. 3, 91–121.

Najjar, A., Mede, K., 2022. Trajectory-User Linking Is Easier Than You Think, in: 2022 IEEE International Conference on Big Data (Big Data). Presented at the 2022 IEEE International Conference on Big Data (Big Data), pp. 4936–4943. https://doi.org/10.1109/BigData55660.2022.10020360

Rao, J., Gao, S., Kang, Y., Huang, Q., 2020. LSTM-TrajGAN: A deep learning approach to trajectory privacy protection. arXiv preprint arXiv:2006.10521.

Rui, C., Benjamin, C., Fung, M., Desai, B.C., 2011. Differentially private trajectory data publication. CoRR, abs/1112.2020.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems 10, 557–570.

Zhang, H., Chen, W., Zhao, X., Qi, J., Jiang, G., Yu, Y., 2025. Scalable Trajectory-User Linking with Dual-Stream Representation Networks. Proceedings of the AAAI Conference on Artificial Intelligence 39, 13224–13232. https://doi.org/10.1609/aaai.v39i12.33443

Zheng, Y., Capra, L., Wolfson, O., Yang, H., 2014. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 5, 1–55.

Zhou, F., Gao, Q., Trajcevski, G., Zhang, K., Zhong, T., Zhang, F., 2018. Trajectory-User Linking via Variational AutoEncoder., in: IJCAI. pp. 3212–3218.