Taylor & Francis
Taylor & Francis Group

Journal of Location Based Services

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tlbs20

Influence of tracking duration on the privacy of individual mobility graphs

Nina Wiedemann, Henry Martin, Esra Suel, Ye Hong & Yanan Xin

To cite this article: Nina Wiedemann, Henry Martin, Esra Suel, Ye Hong & Yanan Xin (2023): Influence of tracking duration on the privacy of individual mobility graphs, Journal of Location Based Services, DOI: 10.1080/17489725.2023.2239190

To link to this article: https://doi.org/10.1080/17489725.2023.2239190

Published online: 27 Jul 2023.

JOURNAL OF LOCATION BASED SERVICES
https://doi.org/10.1080/17489725.2023.2239190

Taylor & Francis
Taylor & Francis Group

Influence of tracking duration on the privacy of individual mobility graphs

Nina Wiedemann[a*], Henry Martin[a,b*], Esra Suel[a,c], Ye Hong[a] and Yanan Xin[a]

[a]Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland; [b]Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria; [c]Center for Advanced Spatial Analysis, University College London, London, UK

ABSTRACT

Location graphs, compact representations of human mobility without geocoordinates, can be used to personalise location-based services. While they are more privacy-preserving than raw tracking data, it was shown that they still hold a considerable risk for users to be re-identified solely by the graph topology. However, it is unclear how this risk depends on the tracking duration. Here, we consider a scenario where the attacker wants to match the new tracking data of a user to a pool of previously recorded mobility profiles, and we analyse the dependence of the re-identification performance on the tracking duration. We find that the re-identification accuracy varies between 0.41% and 20.97% and is affected by both the pool duration and the test-user tracking duration, it is greater if both have the same duration, and it is not significantly affected by socio-demographics such as age or gender, but can to some extent be explained by different mobility and graph features. Overall, the influence of tracking duration on user privacy has clear implications for data collection and storage strategies. We advise data collectors to limit the tracking duration or to reset user IDs regularly when storing long-term tracking data.

ARTICLE HISTORY
Received 15 February 2023
Accepted 15 July 2023

KEYWORDS
Privacy; tracking data; mobility graphs; time dependence; location privacy

1. Introduction and background

Companies are increasingly gathering and using spatio-temporal location data from personal mobile devices. User location data have substantially improved location-based services (LBS) and personalised offers (Keßler and McKenzie 2018). However, detailed mobility traces collected from individuals may contain sensitive personal data associated with high privacy risks (Banerjee 2019; Primault et al. 2018). A particular concern is the increasing integration of user data from different sources (Thompson and Warzel 2019), enabling companies to build more detailed and complete user profiles

CONTACT Nina Wiedemann nwiedemann@ethz.ch

*Authors contributed equally.

(Melendez and Pasternack 2019). Therefore, identifiability (and matching) of individuals from different datasets is a critical dimension of data privacy risk (Keßler and McKenzie 2018).

Previous studies showed that removing basic identity information from mobility traces is insufficient in this context, as users can be re-identified using the information on frequently visited locations (De Montjoye et al. 2013; De Mulder et al. 2008; Gambs, Killijian, and Del Prado Cortez 2014; Golle and Partridge 2009; Rossi, Walker, and Musolesi 2015; Zang and Bolot 2011). One solution proposed in the literature is to obscure the geographic coordinates to guarantee ε-differential privacy (Andrés et al. 2013; Duckham and Kulik 2005; Haydari et al. 2021; Wang et al. 2017) or k-anonymity (Charleux and Schofield 2020; Gruteser and Grunwald 2003; Shokri et al. 2010; Sweeney 2002). For reviews of geoprivacy attacks and protection methods, we refer readers to Kounadi et al. (2018) and Fiore et al. (2020). Nevertheless, location obfuscation and related methods only provide limited privacy protection. For example, Tong et al. (2022) extend the notion of 'location uniqueness' to 'trajectory uniqueness' and show that full trajectories may be exploited for improving re-identification, and Zhen et al. (2019) argue that k-anonymity does not protect from a semantic inference about visited locations.

Another promising possibility for privacy-preserving storage and processing of individual tracking data is given with so-called *location graphs* or *mobility networks* (Raubal, Bucher, and Martin 2021; Rinzivillo et al. 2014). In these graphs, nodes represent visited locations and edge weights correspond to the number of observed movements between these locations. Graph representations offer several benefits: 1) they can be enriched with node and edge features based on the application needs, 2) they are compact and grow sub-linearly in size with increasing tracking duration, 3) they still provide rich insight into mobility behaviour despite their compactness (Martin et al. 2023; Rinzivillo et al. 2014; Wiedemann, Martin, and Raubal 2022) and can be analysed efficiently with graph neural networks for various applications such as activity purpose imputation (Martin et al. 2018).

However, the privacy and unique identifiability properties of individual mobility graphs are not well understood. Recently, Manousakas et al. (2018) showed that the graph topology of personalised mobility graphs, even when all coordinate and time stamp information is removed from its nodes, is often uniquely identifiable. In this paper, we build upon their work and aim to understand the dependency of privacy preservation on tracking duration. Intuitively, location graphs over short periods contain less information about users and may reduce the risk of deanonymization. To investigate this possibility, we divide a tracking dataset of 137 users into distinct periods of different durations and analyse attack scenarios where a new location graph is matched to a pool of location graphs of known users. Our experiments indeed show that matching performance depends on the tracking duration of both pool data and new data; however, there is a considerable re-identification risk even with just a few weeks of tracking duration.

2. Materials and methods

2.1. Data and preprocessing

We analyse the time dependency of topology privacy on a high-quality tracking dataset, collected through the SBB Green Class 1 tracking study (Martin et al. 2019). The study was conducted by the Swiss Federal Railways (SBB) to evaluate the impact of a mobility-as-a-service offer on individuals' travel behaviour. Study participants are predominantly male with above-average income. All study participants were tracked over a full year using an application installed on their phone that segments tracking data into stationary periods called *staypoints*, labelled with activity purpose, and movement behaviour called *triplegs*, labelled with transport modes. All preprocessing is done in Python and PostgreSQL using the Trackintel movement data processing library (Martin et al. 2023). The staypoints are clustered into locations with the DBSCAN algorithm with the parameter $\varepsilon = 30m$, and a minimum number of one point per cluster, i.e. each staypoint is assigned to a location. The Trackintel library merges consecutive staypoints and triplegs into trips as long as they are not interrupted by an activity (staypoints with duration >25 min or labelled with a purpose other than wait and unknown) or by a temporal gap (here 25 minutes). Finally, when constructing the graph, we filter out users with low tracking coverage during the selected time period. The users are required to have a tracking coverage of at least 70% in at least one-third of the days. In our experiments, this leads to a varying number of 132–137 users depending on the time periods used.

Based on the sequence of locations and trips of a user, we construct the individual location graph (or mobility network) as described by Manousakas et al. (2018): In the graph $G(V, E)$, each location is one node, and each trip between two locations increases the weight of the directed edge by one. The edge weight $w(e)$ thus corresponds to the number of transitions during the observation period. To analyse the impact of different tracking periods, we build the graphs on subsets of the dataset that are created by binning the dataset into non-overlapping time periods of 1, 2, 4, 8, 16, 20, 24, and 28 weeks (see Figure 1).

Furthermore, we use the SBB Green Class 2 study (Martin et al. 2019), which was a smaller follow-up study where 50 different participants were tracked under similar conditions for a full year directly after the Green Class 1 study. The data is processed in the same way as the Green Class 1 data, but due to the lower number of users, we will only use this dataset to validate our results in section 3.2, section 3.3 and section 3.4.

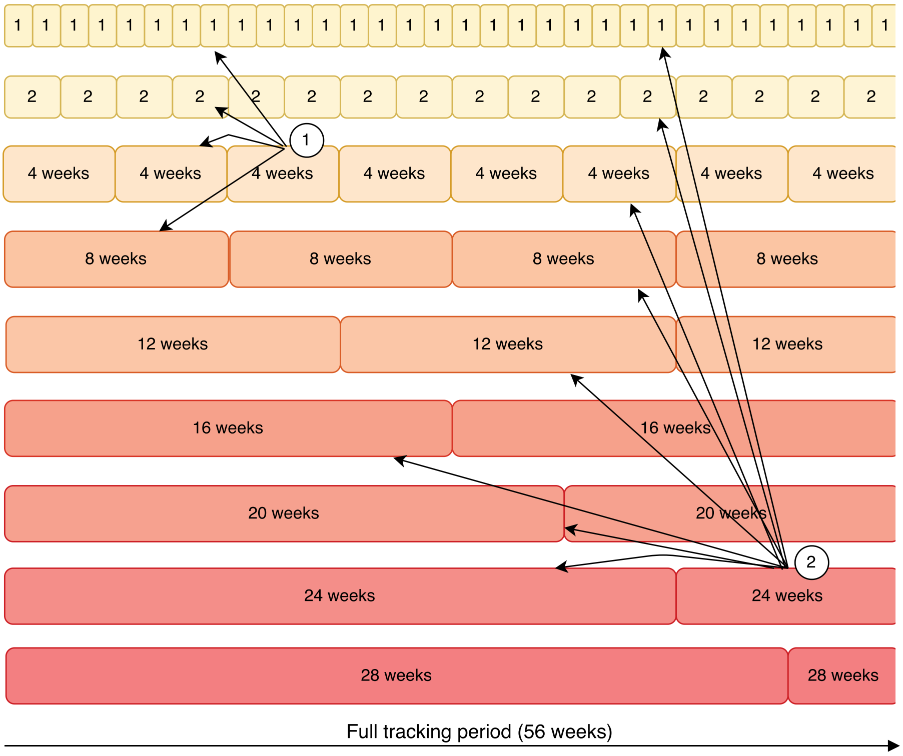


Figure 1. Experimental setup: The tracking data, comprising 56 weeks, are split into non-overlapping bins of varying duration. In the attack scenario, new tracking data from one period is matched to a pool of users at a previous time period. In example 1) the test data of four weeks length can be compared to the pool in the preceding 1, 2, 4 and 8 weeks. In the second example (marked as 2), a test user with tracking data from the second 24-week period is matched to users from all directly preceding tracking data, which includes one from each tracking duration except for 28 weeks.

2.2. Feature based graph matching

Graph matching describes the problem of either identifying if two graphs are isomorphic (exact graph matching) or identifying the best match from a set of candidate graphs (inexact graph matching) (Riesen, Jiang, and Bunke 2010). The exact solutions for both problems are computationally intractable, therefore we rely on heuristics to accomplish inexact graph matching. Related works have proposed so-called R-convolution graph kernels (Haussler 1999) that measure the difference between two graphs in terms of counts of certain substructures, such as paths. Similarly, we compare the distributions of selected graph features to approximate the graph similarity. We represent each graph in a fixed-size vector $v(G)$ that expresses graph characteristics, e.g., the distribution of node in-degrees. Two graphs G_1 and G_2 are compared in terms of the distance between their vector representations, $d(v(G_1), v(G_2))$. As distance metrics d, we test a simple Mean Squared Error (MSE), Kullback-Leibler divergence, and Wasserstein distance.

We experiment with five vector-based graph representations $v(G)$:

- $v_{indegree}$: Distribution of (unweighted) node in-degrees, i.e., the number of connections of one location to other locations. The distribution of in-degrees over the 20 most popular locations is used.
- $v_{outdegree}$: Similar to the in-degree, the distribution of out-degrees over the 20 locations with the highest out-degree is computed.
- $v_{transition}$: The distribution of transition weights over the 20 most popular trips. Intuitively, some users commute between very few locations more frequently than other locations, whereas some users transit more evenly among locations (Pappalardo et al. 2015).
- $v_{shortest\_path}$: The distribution of shortest-path lengths in the graph. All-pairs shortest paths were computed with the Floyd-Warshall algorithm (Floyd 1962; Warshall 1962). The ratio of shortest paths with length x for $x \leq 10$ is reported in $v_{shortest\_path}$.
- $v_{centrality}$: The betweenness centrality (Freeman 1977) of a node denotes its centrality in terms of network hops with respect to other nodes, which is bounded between 0 and 1. Since many nodes have low centrality in mobility graphs, we construct 10 bins from 0 to 1 in log space and report the number of nodes per centrality bin.

Finally, we concatenate all five graph descriptors into one combined vector v_{comb}.

2.3. Experiment design

We analyse the following privacy attack scenario: The adversary is a data broker with access to a pool of users and their tracking data. The attacker then gets access to additional tracking data of a test user, which she wants to match to the correct user in the pool to create a combined user profile. All tracking data are represented as weighted and directed individual location graphs without node or edge features such as coordinates. In the following, we define $u_i^{pool} (i \in [1..n])$ as the i-th user in a pool of n users, and $u_j^{test} (j \in [1..m])$ as a user of the test dataset, $D_{test} = \{u_j^{test}\}$. Let G_i^{pool} and G_j^{test} further denote the corresponding location graphs.

The adversary now aims to find the best match out of the pool users for each test user u_j^{test}. This is accomplished by computing the distance of the graph descriptors presented in section 2.2. The pairwise distances from a test user to all users of the pool are computed as $d(v(G_j^{test}), v(G_i^{pool}))$ and the pool users are ranked according to their distances. As a result, we obtain the rank assigned to the true match of a user in the pool. In other words, we are only interested in the rank that was assigned to the user in the pool that corresponds to the test user ($u_i^{pool} = u_j^{test}$) and the assigned rank $r_j = r(u_j^{test})$ means that this user had the r_j-highest similarity to herself compared to all other users in the pool.

To obtain statistically robust results, we evaluate the scenario on all possible tracking period combinations for the pool and the test user. Figure 1 gives an overview of the experimental setup and demonstrates that the tracking period combinations are not unique. For example, for our total tracking time of 56 weeks, there are 14 distinct 4-week periods and 7 distinct 8-week periods. We do not evaluate all possible combinations (here 98) but regard only combinations where the test user is matched to the closest, directly preceding tracking period in the pool. This choice of valid pool and test user pairs is exemplified by the black arrows in Figure 1. In section 3.6, we additionally consider periods that are not *directly* successive in order to understand the effect of temporal gaps between the pool and test user.

For every valid time bin combination for a given combination of tracking periods, we match every available test user to the users from the pool and evaluate the matching success using the metrics introduced below. All code for the experiments is publicly available[1], however, we can not publish the tracking dataset to protect the privacy of the study participants.

2.4. Metrics for re-identification performance

To evaluate the success of the matching attack, we employ two metrics: the top-k matching performance and the mean reciprocal rank (MRR). Both rely on the rank assigned to the true match of a test user in the pool as introduced above, $r(u_j^{test})$.

We then report the top-k matching performance in one set of test users D_{test} as

$$Acc(D_{test}, k) = \frac{1}{|D_{test}|} \sum_{u_j \in D_{test}} 1\{r(u_j^{test}) \leq k\}.$$

This considers a match as successful if the true match of the test user is among the top-k closest users in the pool.

Furthermore, we use the MRR as a second evaluation metric, defined as the average of the inverse of the ranks in a test dataset. It is a common metric in information retrieval and re-identification tasks (Craswell 2009). The MRR of a test set is

$$MRR(D_{test}) = \frac{1}{m} \sum_{u_j \in D_{test}} \frac{1}{r(u_j^{test})}.$$

The MRR can be interpreted as the harmonic mean of the ranks, with the property that good matches (high rank) have a much higher influence than bad matches (low rank).

3. Results and discussion

We run the experiment described in section 2.3 for all combinations of tracking periods and consecutive start times, resulting in 827 combinations. For each of these combinations, we attempt a matching for every user available in the dataset, which results in over 13 million user-to-user comparisons (Green Class 1). We find that the best matching performance is achieved with the combined graph descriptor v_{comb} and the mean squared error (MSE) as the similarity metric d. See Table 2 and section 3.2 for more details on this choice.

In the following, we report the MRR and top-k matching accuracy for each combination of the pool- and test-user tracking duration. We report the average result and the standard deviation if several accuracy results for a tracking period combination are obtained (due to multiple time bin combinations).

3.1. *Effect of tracking period on re-identification performance*

Figure 2 shows the average matching performance and the standard deviation for all duration combinations of the pool and the test users. All metrics show a significant dependency on both the duration of the pool and the duration of the test user data. This result implies that privacy-friendly applications should be designed such that their tracking duration is as short as possible. This is especially true when new tracking data is to be collected because a privacy-concerned person does not have control over the duration of the pool in our scenario, as the pool represents data already collected by a third party.

Furthermore, even for the shortest tracking duration that was tested (i.e., one week combined with one week), the re-identification capability of our simple matching strategy is substantially better than random (see Figure 2). A random rank assignment would result in a top-10 accuracy of 7.6%, compared to the accuracy of 19.4% from the shortest tracking duration. Thus, the graph representation, even without any additional context or coordinate information, is not anonymous, which is in line with the conclusion reported from Manousakas et al. (2018).

We further analysed the importance of the pool duration, the test user duration, and the difference between their durations, using linear regression with the duration as the independent variable and the average performance as the dependent variable. The resulting coefficients are shown in Table 1. While both duration variables positively impact the performance, the influence of test duration is slightly stronger. For every additional week of test tracking duration, the top-10 identification accuracy increases by 1.06% on average. As the pool is not under the user's control, a potential solution to minimise the privacy risk is to require data brokers to reset user IDs after a specific tracking period. Notably, Table 1 also reveals a major effect from the similarity of pool and test tracking duration, corresponding to the strong performance on the diagonals in Figure 2. This can be explained by the higher similarity of graphs constructed from the same tracking duration, making it easier to match the correct user.

Table 1. Regression analysis of the effect of the pool- and test-user tracking duration on the matching performance. Both positively affect the re-identification performance (=negative impact on privacy); however, the effect of the test duration is slightly higher. The matching performance is higher if the absolute difference between the pool and test user duration is low. All results are significant (p-values << 0.01).

	test duration	pool duration	absolute difference between pool and test duration	intercept	R2 score
MRR	0.01	0.01	−0.01	0.09	0.90
1-Accuracy	0.41	0.43	−0.42	2.42	0.90
5-Accuracy	0.83	0.82	−0.87	10.89	0.89
10-Accuracy	1.06	0.97	−1.10	18.76	0.87

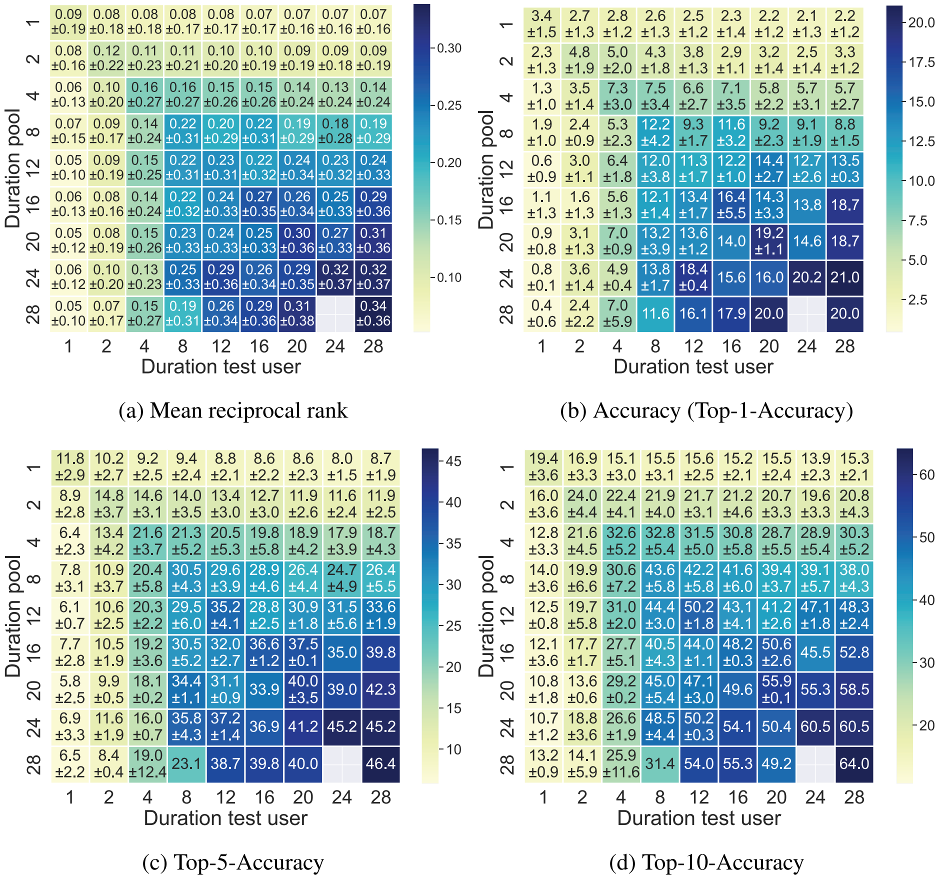
For the interpretation of the results, it is important to note that the results with small bins are statistically more robust than those with large bin combinations because more bins are available. For several combinations of large bins, only one trial was available; therefore, no standard deviation was reported, and no distinct time bins were available for the combination of 28 weeks pool duration and 24 weeks test tracking duration.

3.2. Ablation of approximate graph matching workflow

In section 2.2, we proposed several graph descriptors to calculate the distance between graphs. Table 2 lists the matching performance of different graph features and distance functions. We note that the distance function does not strongly affect the matching performance. In contrast, the features result in very different re-identification abilities. The transition weight and in-degree distribution are the most useful features, whereas node centrality obtains low matching capability. Based on the results in Table 2, we chose the MSE of all features combined, as this performs best on average according to three out of four error metrics. While our focus is on the time-dependency of privacy preservation, future work could analyse the limits of re-identification of location graphs by using more complex

Table 2. Matching performance of different combinations of features, distance functions, and evaluation metrics. The highest matching accuracy is achieved with an R-convolution kernel that computes the MSE between all graph-features distributions combined.

		Recip. rank		1-Accuracy		5-Accuracy		10-Accuracy	
Distance metric d	$v(G)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
KL-divergence	transition	0.10	0.19	3.92	9.60	13.19	28.46	21.73	40.80
	in_degree	0.10	0.20	3.29	10.40	12.05	24.80	20.83	41.60
	out_degree	0.09	0.17	3.07	7.32	11.74	24.80	21.06	36.29
	shortest_path	0.07	0.11	1.77	4.13	7.45	14.52	13.59	26.61
	centrality	0.04	0.06	0.79	2.02	4.14	8.06	8.52	15.32
	combined	0.16	0.35	7.96	23.33	21.41	51.20	31.25	62.40
MSE	transition	0.10	0.19	3.89	9.24	12.98	28.00	21.80	41.60
	in_degree	0.10	0.18	3.49	9.76	12.22	22.40	20.73	35.48
	out_degree	0.09	0.16	2.76	6.61	11.48	25.60	20.52	39.20
	shortest_path	0.07	0.11	1.89	4.04	7.80	15.32	14.16	29.03
	centrality	0.05	0.07	1.17	3.25	5.26	11.16	9.67	16.74
	combined	**0.17**	0.34	**8.40**	20.97	**22.36**	46.40	**32.73**	**64.00**
Wasserstein distance	transition	0.10	0.19	3.82	9.21	13.32	30.08	21.50	41.60
	in_degree	0.10	0.19	3.38	10.40	12.18	25.60	20.96	36.00
	out_degree	0.09	0.17	2.96	8.13	11.64	24.00	20.70	40.32
	shortest_path	0.06	0.11	1.64	4.20	6.48	16.00	11.94	24.00
	centrality	0.05	0.09	1.14	4.13	5.03	11.38	9.69	16.53
	combined	0.15	**0.36**	7.14	24.00	19.71	52.80	28.94	61.60
Sum all metrics	combined	0.16	0.36	8.07	22.50	21.81	**52.80**	31.67	62.40



(a) Mean reciprocal rank

(b) Accuracy (Top-1-Accuracy)

(c) Top-5-Accuracy

(d) Top-10-Accuracy

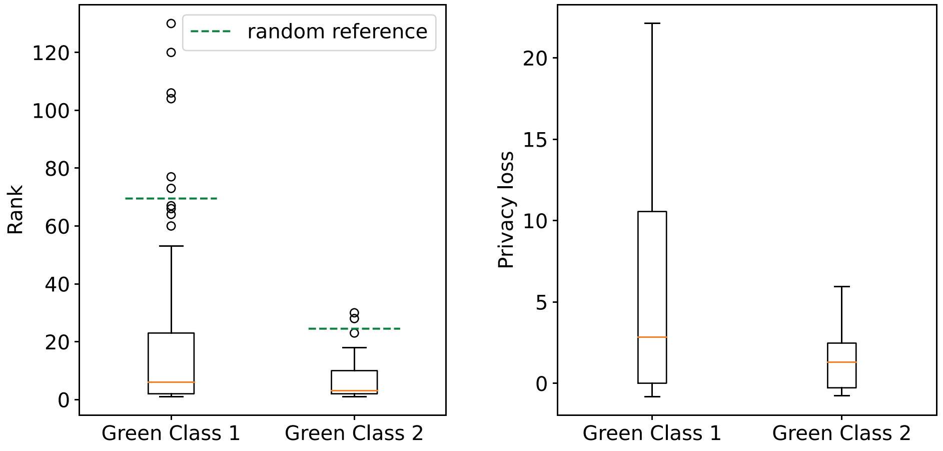
Figure 2. Dependency of matching performance on tracking duration. Top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user.

matching methods such as deep graph kernels (Yanardag and Vishwanathan 2015).

3.3. Validation of matching methodology based on related work and the Green Class 2 dataset

We first validate our results by conducting the same experiment on the Green Class 2 data described above. The full pool-user-duration matrices can be found in Figure A1 in Appendix A. The results show a similar dependency on pool and tracking duration, but, due to the lower number of users, the re-identification accuracy is generally higher (up to 82% top-10 accuracy) and the results are less stable.

We further compare our results on both datasets to the results reported by Manousakas et al. (2018). In their longitudinal study, Manousakas et al. (2018) split the tracking data user-wise into two parts at a random point in time, sampled uniformly between 30% and 70% of the whole period (around one year). The most comparable experiment from our study is the one where both the pool and the test duration are 28 weeks. Following the evaluation by Manousakas et al. (2018), we show the distribution of ranks and the 'privacy loss' in Figure 3. Although the absolute ranks are not informative due to the different pool sizes (132 users/27 users[2] for our dataset versus 1500), the re-identification ability can be compared in terms of the shift of true rank. Specifically, the mean of the true rank is shifted from 62 (random) to 17.1 (informed adversary) for Green Class 1 and from 13 (random) to 7.6 (informed adversary) for Green Class 2, whereas the experiment in (Manousakas et al. 2018, p. 13) yields a shift from 750



(a) Rank distribution

(b) Privacy loss [19]

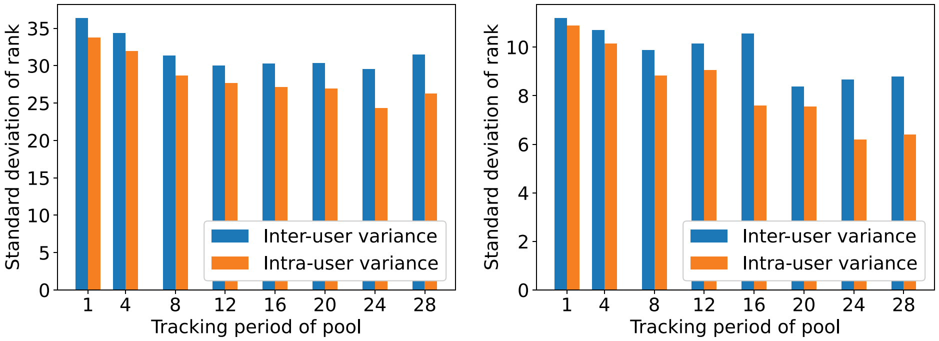
Figure 3. Evaluation of rank distribution and privacy loss as proposed by Manousakas et al. (2018).

to 140. The study by Manousakas et al. (2018, p. 14) also reported a median privacy loss of 2.52 which means ‘the informed adversary can achieve a median deanonymization probability 3.52 times higher than an uninformed adversary’. In our experiments, the median privacy loss is 2.85 for the Green Class 1 data and 1.31 for the Green Class 2 data. Overall, we reproduced the results successfully and extended their results with additional analysis of the impact of tracking durations.

3.4. Intra-user vs inter-user variability of re-identification performance

The main results of this study (Figure 2) are reported as average matching performance. We now further analyse the sources of variance of the matching performance by analysing the variance of the rank assigned to users during the matching. In particular, we aim to answer the following question: Is the variance due to strong differences between users (e.g., easy-to-match vs hard-to-match users), or due to a change in a user’s re-identification ability over time? To answer this question, we calculate the standard deviation between different users in the same timesteps (inter-user) and for the same user over several timesteps (intra-user).

Figure 4 shows that the inter-user standard deviation is consistently higher than the intra-user standard deviation for both datasets. This indicates the existence of user groups that are consistently hard or easy to match. Moreover, the intra-user standard deviation in general decreases as the tracking duration increases for both datasets, which can be explained by the higher stability of long-term location graphs.



(a) Green Class 1

(b) Green Class 2

Figure 4. Inter vs intra person variability of matching performance. The variance over users is higher than the variance over time bins. Intra-user variance decreases with growing tracking duration.

3.5. Factors that explain re-identification performance

Given the high variance in the re-identification ability over users, we further analyse features that could drive the degree of recognition of a user. For this purpose, we compute features commonly used to describe individual mobility behaviour, such as the radius of gyration, the typical trip distances, and the entropy of location sequences (random and real entropy) (Song et al. 2010). Additionally, we compute graph features proposed by Martin et al. (2023), which describe the complexity and centeredness of the location graphs. Last, we regard socio-demographics extracted from surveys in the context of the Green Class 1 and Green Class 2 studies, namely age, gender and whether the user subscribed to a public transport subscription in Switzerland (PT). Note that all features are computed as a single value for all users since there is only one value per user for sociodemographics and classical mobility features. We use the average value over both 28-week bins for the graph features to describe the user's stable graph topology.

In Table 3, the coefficients of a regression analysis with the above-presented features as independent variables and the normalised rank as the dependent variable are given. The normalised rank is the true user's rank in the matching process, normalised by the total number of users, which allows to combine the users of Green Class 1 with the ones from Green Class 2 in this study. We further checked the correlation r between attributes to exclude potential collinearity issues, but $r < 0.6$ for all pairs. A significant positive coefficient indicates that a feature hampers the re-identification ability since it leads to a higher rank. The model is fitted separately for each tracking duration (1, 2, 4, ..., 28), whereby we only consider scenarios with the same pool- and user tracking duration, corresponding to the diagonal of the matrices in Figure 2, and we average all available rank predictions for each user (i.e., average over time bins).

According to the regression coefficients (Table 3), socio-demographics do not affect the rank significantly. A higher radius of gyration makes a user harder to identify which might be related to an increased variability of the location graph over time due to a higher level of travel activity. For long durations, a high random entropy increases the identification performance. The random entropy increases if time is spent at many different locations which increases the complexity and uniqueness of a graph and therefore makes it easier to match. The graph features, in particular the journey length, also significantly affect the rank, but in an unexpected direction: More star-shaped graphs, indicated by low journey length, low hub size, and high transition β, yield higher ranks.

Table 3. Effect of mobility behaviour and socio-demographics on re-identification accuracy, i.e., the rank of a user. A linear model is fitted and the coefficients are reported. Significant coefficients (p-value below 0.05) are marked with (*).

		Classical mobility features				Graph features				Sociodemographics			
Duration	Const.	Radius of gyration	Random entropy	Real entropy	Median distance	Journey length	Hub size	Degree β	Transition β	Age	Sex	PT	R^2
1	32.62 (*)	1.54 (*)	−0.73	0.48	−0.04	−0.83	−0.65	0.04	0.77	−0.25	−0.39	−0.48	0.13
2	29.8 (*)	1.83 (*)	−0.62	0.19	0.06	−1.46 (*)	−0.1	−0.6	−0.46	−0.09	−0.61	−0.64	0.11
4	26.01 (*)	2.7 (*)	−0.93	−0.19	0.51	−2.5 (*)	−0.44	−1.06	−1.17	0.14	0.2	−0.98	0.15
8	20.56 (*)	2.65 (*)	−3.2 (*)	2.35	−1.44	−2.75 (*)	−1.23	−3.41 (*)	−0.01	−0.39	0.84	−1.16	0.17
12	19.24 (*)	2.52 (*)	−3.64 (*)	−0.18	−2.73 (*)	0.52	−1.92	0.35	2.28	1.63	0.6	1.09	0.14
16	20.33 (*)	2.98	−1.67	1.68	−1.05	−2.71	1.16	−3.66	2.34	0.56	0.95	0.88	0.07
20	17.75 (*)	1.92	−3.68	1.43	−0.51	−0.45	2.12	−1.8	0.42	1.69	−0.21	0.82	0.04
24	16.3 (*)	2.82	−4.18	1.63	0.34	1.72	0.93	−0.26	5.62 (*)	0.71	−0.12	3.02	0.08
28	16.53 (*)	2.58	−7.27 (*)	2.99	0.53	0.43	4.7 (*)	−1.45	6.43 (*)	0.09	0.43	3.46	0.12

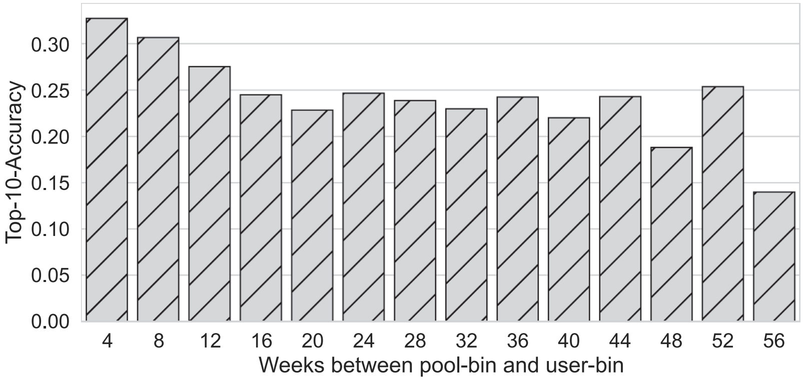


Figure 5. The re-identification accuracy decreases when there is a larger temporal gap between pool-bin and user-bin. However, the accuracy converges slowly and retains more than half of its former value even after one year.

3.6. Influence of temporal difference between pool and user tracking period

The experiments reported so far were restricted to consecutive time periods (see Figure 1). Here, we further analyse the effect of a temporal gap between the tracking periods. Since the number of possible combinations becomes very high in this setting, we restrict the analysis to one pool- and user duration combination and analyse the 1.56 million combinations where pool- and user duration are four weeks and the pool was recorded before the user duration. Figure 5 shows the results, where the top-10 re-identification accuracy is shown by the temporal gap. As expected, the matching performance decreases as we increase the duration of the gap. However, it stabilises already at around 16 weeks between pool and test user, and remains surprisingly high even for the longest gap of 56 weeks. This finding implies that saved location data can be exploited by an attacker for a long time.

4. Conclusion

In this work, we present a set of experiments to analyse how tracking duration influences the re-identification ability of individual location graphs. Our experiments on time-binned subsets of one-year tracking data show that the tracking duration indeed has a strong effect on the success of a privacy attack, with the re-identification accuracy at the longest tracking duration (28 weeks) being more than 3 times higher than when matching 1-week tracking data. We further show that the re-identification ability increases in roughly equal parts with increased tracking duration of the pool of candidate users on the one hand, and increased tracking duration of the test user on the other hand. Therefore, privacy-friendly applications should only require tracking data over periods that are as short as possible, and data brokers should be required to reset the user IDs of their data regularly to limit the pool duration. On top of that, long-term storage of tracking data should be impeded, since the re-identification accuracy only slowly decreases with increasing time between pool and test tracking period.

More generally, we confirm results from Manousakas et al. (2018) that location graphs without coordinates or additional context information are sufficient to re-identify users with a success rate significantly higher than random. At the longest tracking duration, the de-anonymisation probability of an informed adversary is 3.85 times higher than the one of an uninformed adversary for our dataset. Our work reveals many opportunities for further work on location-graph privacy. For example, we found that certain users are consistently hard or easy to be identified. Characterization of these user groups should be explored in future work. We take a step in this direction with our analysis of the relation to different mobility-behaviour features and socio-demographics, but our results hint at more complex characteristics that make a user hard to re-identify. Evidence from more diverse datasets may help to find such influence factors. The reproduction of our experiments on new datasets is straightforward as the individual location graphs have very few requirements (e.g., no specific features or labels needed). At the same time, future work could also regard the re-identification risk of more complex location graphs, e.g., amended with temporal information.

Finally, it is important to mention that we only employed a simplistic matching strategy, and a more sophisticated matching approach, such as learning graph similarities with deep neural networks (Guixiang et al. 2021), could lead to even higher success rates for matching. The results should therefore be considered as a lower bound of possible matching success. The presented analysis however augments the understanding of the privacy risk of tracking data – even if it is reduced to topology – and can improve the regulation of anonymisation practices.

Notes

1. https://github.com/mie-lab/topology_privacy.
2. For long time bin durations, not all users matched the criteria set for the tracking coverage.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the ETH Zürich Foundation [MI-01-19]; Hasler Stiftung [1-008062].

CRediT author statement

Wiedemann, Nina: Conceptualization, Methodology, Software, Visualization, Writing – Original Draft, Review & Editing; **Martin, Henry**: Conceptualization, Methodology, Software, Writing – Original Draft, Review & Editing; **Suel, Esra**: Conceptualization, Writing – Original Draft, Review & Editing; **Hong, Ye**: Conceptualization, Writing – Review & Editing; **Xin, Yanan**: Conceptualization, Writing – Review & Editing;

Informed consent statement

The Green Class 1 and 2 studies were conducted by SBB. The participants provided informed consent to be tracked over the study period, and for their data to be shared for research purposes.

References

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis, and Catuscia Palamidessi. 2013. "Geo-Indistinguishability: Differential Privacy for Location-Based Systems." In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, Berlin, Germany, 901–914.

Banerjee, Syagnik. 2019. "Geosurveillance, Location Privacy, and Personalization." *Journal of Public Policy & Marketing* 380 (4): 0 484–499. https://doi.org/10.1177/0743915619860137 .

Charleux, Laure, and Katherine Schofield. 2020. "True Spatial K-Anonymity: Adaptive Areal Elimination Vs. Adaptive Areal Masking." *Cartography and Geographic Information Science* 470 (6): 0 537–549. https://doi.org/10.1080/15230406.2020.1794975.

Craswell, Nick. 2009. "Mean Reciprocal Rank." *Encyclopedia of Database Systems*, 1703. Springer US

De Montjoye, Yves-Alexandre, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 30 (1): 0 1–5. https://doi.org/10.1038/srep01376.

De Mulder, Yoni, George Danezis, Lejla Batina, and Bart Preneel. 2008. "Identification via Location-Profiling in GSM Networks." In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, Alexandria Virginia USA, 23–32.

Duckham, Matt, and Lars Kulik. 2005. "A Formal Model of Obfuscation and Negotiation for Location Privacy." In *Pervasive Computing* 3468 152–170. Berlin Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11428572_10.

Fiore, Marco, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. 2020. "Privacy in Trajectory Micro-Data Publishing: A Survey." *Transactions on Data Privacy* 13: 91–149.

Floyd, Robert W. 1962. "Algorithm 97: Shortest Path." *Communications of the ACM* 50 (6): 0 345. https://doi.org/10.1145/367766.368168.

Freeman, Linton C. 1977. "A Set of Measures of Centrality Based on Betweenness." *Sociometry* 40 (1): 35–41. https://doi.org/10.2307/3033543.

Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez Del Prado Cortez. 2014. "De-Anonymization Attack on Geolocated Data." *Journal of Computer and System Sciences* 800 (8): 0 1597–1614. https://doi.org/10.1016/j.jcss.2014.04.024.

Golle, Philippe and Kurt Partridge. 2009. "On the Anonymity of Home/Work Location Pairs." In *International Conference on Pervasive Computing*, Nara, Japan, 390–397. Springer.

Gruteser, Marco and Dirk Grunwald. 2003. “Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services* Boppard, Germany, 31–42.

Guixiang, Ma, Nesreen K Ahmed, Theodore L Willke, and Philip S Yu. 2021. “Deep Graph Similarity Learning: A Survey.” *Data Mining and Knowledge Discovery* 35:0 688–725. https://doi.org/10.1007/s10618-020-00733-5.

Haussler, David. 1999. Convolution Kernels on Discrete Structures. Technical report, Department of Computer Science, University of California.

Haydari, Ammar, Michael Zhang, Chen-Nee Chuah, Jane Macfarlane, and Sean Peisert. 2021 https://arxiv.org/abs/2112.08487. “Adaptive Differential Privacy Mechanism for Aggregated Mobility Dataset.” *arXiv e-prints. arXiv:2112.08487 [cs].*

Keßler, Carsten, and Grant McKenzie. 2018. “A Geoprivacy Manifesto.” *Transactions in GIS* 220 (1): 0 3–19. https://doi.org/10.1111/tgis.12305.

Kounadi, Ourania, Bernd Resch, and Andreas Petutschnig. 2018. “Privacy Threats and Protection Recommendations for the Use of Geosocial Network Data in Research.” *Social Sciences* 7 (10): 191. https://doi.org/10.3390/socsci7100191.

Manousakas, Dionysis, Cecilia Mascolo, Alastair R Beresford, Dennis Chan, and Nikhil Sharma. 2018. “Quantifying Privacy Loss of Human Mobility Graph Topology.” *Proceedings on Privacy Enhancing Technologies* 20180 (3): 0 5–21. https://doi.org/10.1515/popets-2018-0018.

Martin, Henry, Henrik Becker, Dominik Bucher, David Jonietz, Martin Raubal, and Kay W. Axhausen. 2019. “Begleitstudie SBB Green Class-Abschlussbericht.” *Arbeitsberichte Verkehrs-und Raumplanung* 1439. https://doi.org/10.3929/ethz-b-000353337.

Martin, Henry, Dominik Bucher, Esra Suel, Pengxiang Zhao, Fernando Perez-Cruz, and Martin Raubal. 2018. “Graph Convolutional Neural Networks for Human Activity Purpose Imputation. In *NIPS spatiotemporal workshop at the 32nd Annual conference on neural information processing systems (NIPS 2018)*, Montreal, Canada.

Martin, Henry, Ye Hong, Nina Wiedemann, Dominik Bucher, and Martin Raubal. 2023. “Trackintel: An Open-Source Python Library for Human Mobility Analysis.” *Computers, Environment and Urban Systems* 101:101938. https://doi.org/10.1016/j.compenvurbsys.2023.101938.

Martin, Henry, Nina Wiedemann, Daniel J Reck, and Martin Raubal. 2023. “Graph-Based Mobility Profiling.” *Computers, Environment and Urban Systems* 100:101910. https://doi.org/10.1016/j.compenvurbsys.2022.101910.

Melendez, Steven, and Alex Pasternack. 2019. “Here are the Data Brokers Quietly Buying and Selling Your Personal Information.” *The Fast Company*. https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information.

Pappalardo, Luca, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. “Returners and Explorers Dichotomy in Human Mobility.” *Nature Communications* 60 (1): 0 8166. https://doi.org/10.1038/ncomms9166.

Primault, Vincent, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 2018. “The Long Road to Computational Location Privacy: A Survey.” *IEEE Communications Surveys & Tutorials* 210 (3): 0 2772–2793. https://doi.org/10.1109/COMST.2018.2873950.

Raubal, Martin, Dominik Bucher, and Henry Martin. 2021. “Geosmartness for Personalized and Sustainable Future Urban Mobility.” In *Urban Informatics*, 59–83. Springer Singapore. https://doi.org/10.1007/978-981-15-8983-6_6.

Riesen, Kaspar, Xiaoyi Jiang, and Horst Bunke. 2010. “Exact and Inexact Graph Matching: Methodology and Applications.” In *Managing and Mining Graph Data*, 217–247. Springer US. https://doi.org/10.1007/978-1-4419-6045-0_7.

Rinzivillo, Salvatore, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2014. The Purpose of Motion: Learning Activities from Individual Mobility

Networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, Shanghai, China, 312–318. IEEE.

Rossi, Luca, James Walker, and Mirco Musolesi. 2015. “Spatio-Temporal Techniques for User Identification by Means of GPS Mobility Data.” *EPJ Data Science* 40 (1): 0 11. https://doi.org/10.1140/epjds/s13688-015-0049-x.

Shokri, Reza, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. 2010. “Unraveling an Old Cloak: K-Anonymity for Location Privacy.” In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, Chicago, USA, 115–118.

Song, Chaoming, Qu Zehui, Nicholas Blumm, and Albert-László Barabási. 2010. “Limits of Predictability in Human Mobility.” *Science: Advanced Materials and Devices* 3270 (5968): 0 1018–1021. https://doi.org/10.1126/science.1177170 .

Sweeney, Latanya. 2002. “K-Anonymity: A Model for Protecting Privacy.” *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems* 100 (5): 0 557–570. https://doi.org/10.1142/S0218488502001648.

Thompson, Stuart A and Charlie Warzel. 2019. “The Privacy Project: Twelve Million Phones, One Dataset, Zero Privacy.” *The New York Times*. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html.

Tong, Wei, Yinggang Tong, Chang Xia, Jingyu Hua, Li Qun, and Sheng Zhong. 2022. “Understanding Location Privacy of the Point-Of-Interest Aggregate Data via Practical Attacks and Defenses.“ *IEEE Transactions on Dependable and Secure Computing* 20 (3): 2433–2449.

Wang, Leye, Gehua Qin, Dingqi Yang, Xiao Han, and Ma. Xiaojuan. 2017. “Geographic Differential Privacy for Mobile Crowd Coverage Maximization.” *arXiv: 1710,10477 [Cs]* 32. https://doi.org/10.1609/aaai.v32i1.11285.

Warshall, Stephen. 1962. “A Theorem on Boolean Matrices.” *Journal of the ACM (JACM)* 90 (1): 0 11–12. https://doi.org/10.1145/321105.321107 .

Wiedemann, Nina, Henry Martin, and Martin Raubal. 2022. “Unlocking Social Network Analysis Methods for Studying Human Mobility.” *AGILE: GIScience Series* 3:0 19. https://doi.org/10.5194/agile-giss-3-19-2022.

Yanardag, Pinar and S. V. N. Vishwanathan. 2015. “Deep Graph Kernels.” In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, Sydney, NSW, Australia, 1365–1374. New York, NY, USA: Association for Computing Machinery.

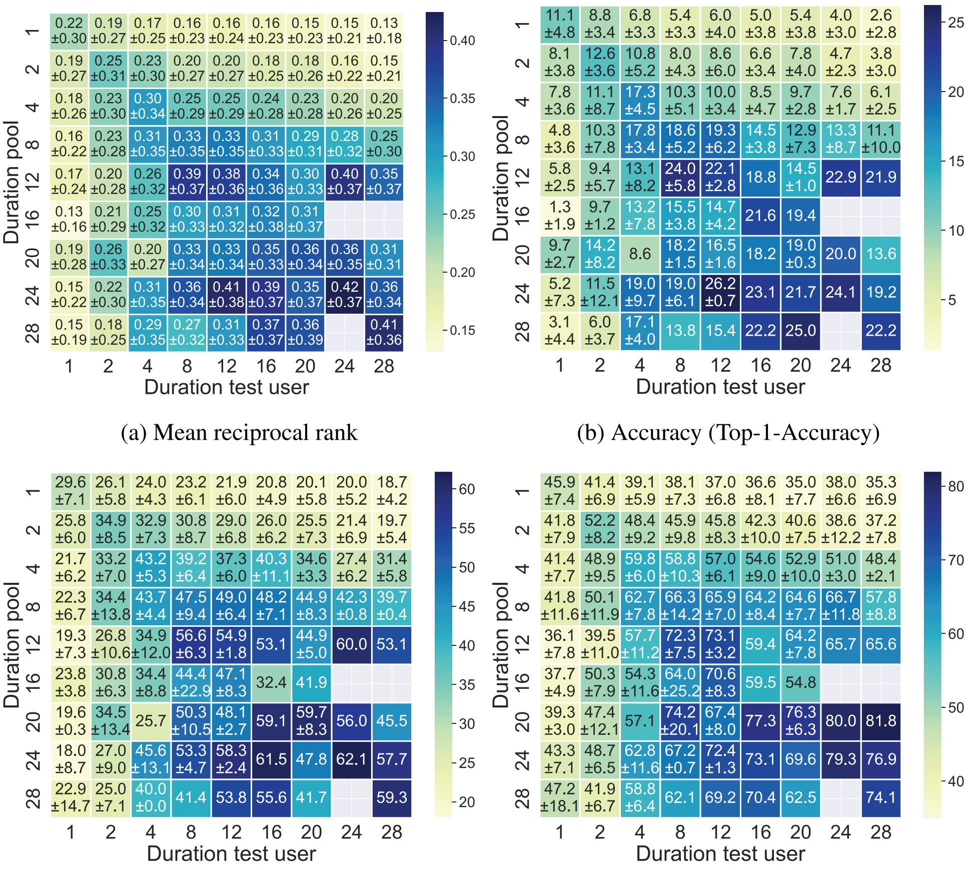
Zang, Hui and Jean Bolot. 2011. “Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study.” In *Proceedings of the 17th annual international conference on Mobile computing and networking*, Las Vegas, USA, 145–156.

Zhen, Tu, Kai Zhao, Xu Fengli, Li Yong, Li Su, and Depeng Jin. 2019. “Protecting Trajectory from Semantic Attack Considering *K* -Anonymity, *L*-Diversity, and *T*-Closeness.” *IEEE Transactions on Network and Service Management* 160 (1): 0 264–278. https://doi.org/10.1109/TNSM.2018.2877790.

Appendix

A Validation on Green Class 2

To validate our results for the Green Class 1 data, we compute the matching performance results on the Green Class 2 data accordingly. Figure A1 visualises the results corresponding to Figure 2. Due to the lower number of users in Green Class 2, the re-identification accuracy is generally higher, but the same patterns as for Green Class 2 can be observed: Both the pool and the test duration impact the matching performance, and the best results are obtained when pool and test duration are the same.



(a) Mean reciprocal rank

(b) Accuracy (Top-1-Accuracy)

(c) Top-5-Accuracy

(d) Top-10-Accuracy

Figure A1. Dependency of matching performance on tracking duration for the Green Class 2 data. Similarly to the results for Green Class 1, the top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user. Due to the lower number of users, the re-identification performance is higher, reaching up to 82% top-10 accuracy.