Assessing Shadows in Mobility: Beyond Spatiotemporal Patterns

Lucas Félix, Anne Josiane Kouam, Aline Carneiro Viana, Nadjib Achir, Jussara Almeida

Assessing Shadows in Mobility: Beyond Spatiotemporal Patterns

Lucas Félix[1], Anne Josiane Kouam[2], Aline Carneiro Viana[3], Nadjib Achir[3], Jussara Almeida[1]

Univesidade Federal de Minas Gerais[1], TU Berlin[2], Inria[3]

1 INTRODUCTION

Mobility is a fundamental aspect of human life that influences numerous dimensions of individual and societal well-being. It is well known that mobility serves as a signature of human behavior. By identifying an individual's or group's mobility patterns, insights can be gained into various aspects of their daily lives, such as the places they visit, work, and live. This understanding also highlights their vulnerability and loss of privacy.

Supporting this view, a study by de Montjoye et al. (2013) [2] demonstrated that only four spatiotemporal points in a cellular mobility dataset are sufficient to uniquely identify 95% of individuals in a population of 1.5 million. This finding highlights the significant vulnerability of users to re-identification, even in datasets with spatial and temporal coarseness. To quantify this vulnerability, the authors proposed a metric called *uniqueness* that measures the number of time-ordered unique displacements a user makes based on their trajectories. The more unique users' movements are, the easier it is to re-identify them [2]. However, *we argue that merely measuring a user's spatiotemporal visiting patterns may not be enough to prevent re-identification, as individual mobility behavior can still be highly distinctive.* It is noteworthy that re-identification techniques that explore more than the users' spatial-temporal patterns could capture vulnerabilities in a high-dimensional space (e.g. embeddings) that are not captured by the uniqueness alone [3].

To assess the shadows in human mobility not captured by uniqueness, we propose a method to quantify how distinguishable a user is from a behavior perspective, based on the distance to the nearest neighbors in a multi-dimensional behavioral space. Specifically, a user that is d distant from k-nearest neighbors (k-NN) is considered indistinguishable by a factor of d from at least $k-1$ other users concerning a set of behavioral metrics. Defining the bounds for d and k that ensure indistinguishability for a user is crucial for quantifying behavioral vulnerability individually but is outside this paper's scope. However, we use d and k as parameters in the uniqueness study.

The users are modeled through standard metrics that capture their circadian mobility behaviors from different perspectives. We categorize these metrics by their capability to capture daily routine, mobility preferences, and visiting uncertainty in space and time. Using these straightforward metrics provides high interpretability of our results, helping identify features that most effectively characterize users' vulnerability. This interpretability allows researchers to understand user behavior and the factors contributing to their vulnerability, facilitating the design of privacy-enhancing methods.

2 METHODOLOGY

We follow the intuition that users that are *far* (i.e., high d) from their neighbors in a multi-dimensional behavioral space can be easily re-identified. On the other hand, users *close* (i.e., low d) to a large enough set of $k-1$ neighbors can blend into a k crowd with users having similar mobility behaviors, hence are k-protected.

We call S the set of users in our dataset, where each user $s \in S$ is associated with a set of $M = \{m_1, ..., m_i\}$ well-known behavioral mobility metrics, each corresponding to a dimension in the $|M|$-dimensional space. Each user is assigned a distance $d \geq 0, d \in \mathbb{R}$, representing the average distance between s and its k-nearest neighbors in the $|M|$-dimensional space. We call S_m^v, the top-v users that are more distant from their k neighbors, where $v > 0, v \in \mathbb{Z}$.

Extract Representation Each metric m provides a different perspective of the users' behavior. We categorize them into 4 groups: **Spatial:** Radius of gyration (RG), 2-RG (e.g., the radius of gyration of the two most visited places), max. displacement, avg and std of jump lengths (i.e. distance between the points visited), number of visits, and number of unique locations; **Temporal:** Avg and std waiting time; **Spatial-Temporal:** Diversity, regularity, stationarity; **Predictability:** Real entropy. Note that such literature metrics capture spatiotemporal routine, mobility preferences, and uncertainties in human behavior, aspects that uniqueness lacks.

For the metrics **pre-processing**, we normalize the data using a min-max scaler to ensure consistency across different feature scales. We use the Manhattan distance to measure distances, as it performs better in high-dimensional spaces than Cosine and Euclidean similarities.

Uniqueness: We set the value of k as the number of unique users given by the uniqueness metric (u) [2]. User uniqueness is determined across time windows of size $|t_w|$; if unique in any window, the entire trajectory is deemed unique. u quantifies how often an ordered combination of all visited places appears across time windows. Lower u indicates greater vulnerability. Users in set S are associated with $u > 0, u \in \mathbb{Z}$, and $S_{u=1}$ comprises the most vulnerable users with the minimum uniqueness $u = 1$.

Uniqueness (u) varies with temporal aggregation (i.e., time window t_w); for $t_w = 1$ (one hour), due to few displacements per hour on average of our data (cf. Table 1), vulnerability is low. Increasing t_w increases unique combinations, increasing the number of unique users, similar to findings by Montjoye et al. [2]. To set t_w, we compare values varying from 1 to 15 hours. With $t_w = 1$, 18% of users are unique; for $t_w > 1$, over 80% are unique, peaking at $t_w = 4$ with 88% in $S_{u=1}$. Thus, we adopt $t_w = 4$ for our evaluations.

We seek to compare vulnerable users according to their uniqueness, i.e., $S_{u=1}$, with users that are the most vulnerable, i.e., isolated, according to their mobility behavior, i.e., S_m^v. We select $k = 10$ neighbors and consider the top v most vulnerable users such that $|S_m^v| = |S_{u=1}|$. Figure 1 provides an overview of our methodology.

3 EXPERIMENTAL EVALUATION

# user	Avg. places per user	Avg. places per user and day	Avg. places per hour and day	Avg. diff places per user	Avg. diff places per user per day
585020	156.16 ± 0.12	15.61 ± 0.017	1.12 ± 0.00	9.79 ± 0.10	2.77 ± 0.004

Table 1: Statistics of the Shanghai after pre-processing.

Dataset overview. We use the non-public and fully anonymized Shanghai CDR dataset (cf. Table 1). We refine this dataset through a trace refinement process. We apply a data completion strategy to the dataset to mitigate the temporal gaps common to CDR datasets, as in [1]. Next, we employ a 200mx200m square tessellation based

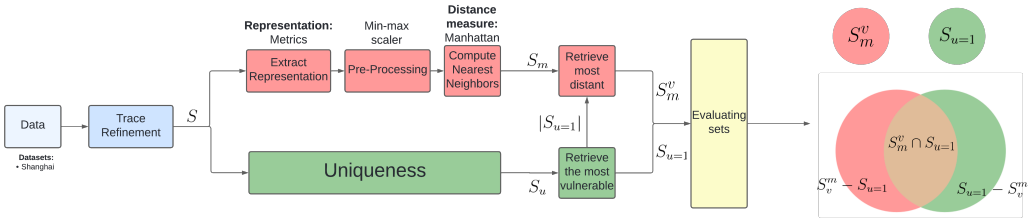

Figure 1: Methodology overview.

on *OpenStreetMap* (cf. *Scikit-mobility*) Lastly, we remove inactive users with a low number of records (i.e., less than 10 days and 120 records, as in [1]), given that these users can influence the identification of the mobility behavior captured in the dataset.

Table 1 presents some statistics (i.e., avg and 95% confidence interval) of the dataset after trace handling. The data indicates that users in Shanghai exhibit limited diversity in their place visits, averaging 2.77 unique places per day and 9.79 overall. Typically, users visit one place per hour, often alternating between common locations like home and work.

Verifying the intersection between vulnerable and isolated users. Figure 2B displays the distance distribution among Shanghai users, including average and minimum distances for sets S, $S_{u=1}$, and S_m^v. Interestingly, the averages across all sets are pretty similar. This is mainly due to the high proportion (about 88%) classified as vulnerable based on uniqueness, with $S_{u=1} \cup S_m^v$ encompassing approximately 95% of vulnerable users (Figure 2A). *This substantial overlap indicates our methodology effectively identifies isolated users considered vulnerable by uniqueness metrics.* Given the vulnerability prevalence (88% of users), this task is notably manageable.

Verifying the Manhattan distance between vulnerable and isolated users. When examining the minimum distance in Figure 2B, we observe distinct behaviors between $S_{u=1}$ and S_m^v. *The minimum distance value for users in$S_{u=1}$ is* 0*, indicating that at least ten users share identical behaviors, meaning that this user is protected from the behavioral perspective. Although these users have distinguishable visiting patterns for potential attackers, they exhibit a "common behavior", which makes it difficult to identify attacks that explore the behavior.* In such scenarios, protective measures can prioritize preserving the unique aspects of user behavior while maintaining overall similarity to the original dataset. This approach maximizes data utility (given that it would be more similar to the original dataset) while mitigating risks effectively.

Vulnerability patterns not captured by uniqueness. When examining the set $S_m^v - S_{u=1}$, we identify users who were not categorized as vulnerable by uniqueness metrics but are significantly distant from their neighbors in the metric space. Statistical testing did not reveal any discernible differences between the distributions of users in $S_m^v - S_{u=1}$ (users distant in metrics but not vulnerable by uniqueness) and $S_{u=1} - S_m^v$ (users vulnerable by uniqueness but not distant in metric space). This lack of distinction likely stems from the dataset's homogeneous user behavior, with most users falling within the intersection of these sets, capturing potentially outlier behaviors. *Within $S_m^v - S_{u=1}$, we identified users exhibiting high stationarity (near 1), indicating they frequent a limited number of places and could potentially be easily re-identified.* These users, however, have an average uniqueness of 217.5 ± 53.74, *suggesting they possess distinctive behavior despite being in areas with many other people. Thus, they exhibit behaviors that are not captured by the uniqueness metrics.*

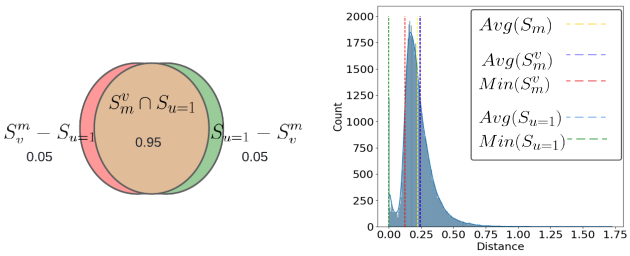

Figure 2: (a) Sets of users. (b) Manhattan distance distribution.

4 CONCLUSIONS

Our findings demonstrate promising potential for developing a technique to quantify user vulnerability and understand which behaviors contribute to vulnerability. We observed a significant overlap between the most isolated users in the multi-dimensional behavioral space and those deemed most vulnerable. This suggests that advancing in this direction could empower researchers to design improved privacy metrics, focusing on protecting the most vulnerable users and enhancing the utility of datasets.

Next Steps: Moving forward, our goals include validating our methodology across various datasets, especially those with fewer stationary users, expanding our behavior vulnerability approach (e.g., investigating the trade-off between d and k), and conducting deeper investigations into the factors contributing to user vulnerability. One limitation of our current approach is the reliance on the size of $S_{u=1}$ to select users in S_m^v. This selection method may inadvertently include users whose neighbors protect their behaviors. To mitigate this issue, we aim to develop an automated process for identifying the most distant users.

ACKNOWLEDGMENTS

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the CAPES-STIC-AMSUD 22-STIC-07 LINT project. All authors approved the final version of the manuscript.

REFERENCES

[1] L. Amichi, A. C. Viana, M. Crovella, and A. Loureiro. 2020. Understanding individuals' proclivity for novelty seeking. In *Inter. conf. on adv. in geo. info. sys.*
[2] YA De Montjoye, CA Hidalgo, M Verleysen, and VD Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* (2013).
[3] L. May Petry and et Al. 2020. MARC: a robust method for multiple-aspect trajectory class. via space, time, and semantic embed. *Inter. Journal of Geo. Info. Science* (2020).