Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3103

SEMANTIC PRIVACY RISKS IN SOCIAL TRAJECTORY PUBLICATION

ZhenZhen Wu, YanFei Yuan*
College of Cyber Security, Tarim University, Alar 843301, Xinjiang, China.
Corresponding Author: YanFei Yuan, Email: yanfeiyuan@taru.edu.cn

Abstract: The publication of trajectory data in mobile applications raises significant privacy concerns for users. When combined with behavioral pattern analysis, the semantic similarity of published trajectories can be exploited by attackers to infer users' travel motivations, posing substantial risks to personal privacy. In this paper, we simulate such attacks by proposing an observation-based algorithm to infer user travel behavior and develop a corresponding privacy risk quantification mechanism. Extensive experiments on real-world datasets validate the effectiveness of the proposed risk quantification approach, providing a foundation for the further development of semantic privacy protection schemes.
Keywords: Mobile application; Behavioral semantics; Privacy inference; Risk quantification

1 INTRODUCTION

The rapid advancement of cloud computing, IoT, and wireless communication technologies has accelerated the growth of mobile communication services [1]. These services are now widely adopted across diverse domains including social entertainment, intelligent transportation, military and defense, as well as government and business operations [2]. To deliver functional and personalized services—such as point-of-interest searches, navigation, and weather forecasts—mobile applications often require users to share their location trajectories [3]. However, the publication of such trajectory data raises significant privacy concerns among users [4].

In response, a variety of location privacy protection schemes have been developed. These include traditional techniques such as anonymization, generalization, obfuscation, and perturbation [5]. For instance, anonymization blends real locations with fake ones [6]; generalization and obfuscation replace precise points with broader regions [7]; and perturbation introduces noise into actual location coordinates [8]. Additionally, geographic indistinguishability applies differential privacy to constrain the distinguishability between real and synthetic locations [9]. Despite their utility, these methods primarily safeguard user privacy within the geographic domain.

Nevertheless, privacy in mobile travel encompasses not only geospatial location protection but also behavioral privacy in the semantic dimension—such as inferring a user's travel motivation. To address this, Dai et al. introduced a semantic generalization method that reduces the semantic similarity between reconstructed trajectories and the user's actual behavior [10]. Our prior work [11] enhanced this approach by adaptively adjusting generalization strength based on the user's access roles at different locations. More recently, we further refined semantic generalization by incorporating behavioral patterns to govern privacy sensitivity-aware strength [12].

These efforts highlight the need for a clear criterion to quantify the appropriate degree of semantic generalization in location privacy protection. Without such a guideline, published trajectories may either reveal user travel motivations due to excessive semantic similarity, or degrade service quality due to overly distorted semantic information.

To bridge this gap, this paper proposes a quantification mechanism that assesses the risk of privacy leakage arising from the semantic similarity of published trajectories. This mechanism offers a quantitative criterion to guide semantic generalization in location privacy protection. The main contributions of this work are as follows.

2 SEMANTIC SIMILARITY

We first represent and quantify the semantic similarity between location functional attributes to support subsequent behavioral inference based on published trajectories.

2.1 Trajectory Data Preprocessing

Location data generated from user mobility, often captured by global navigation satellite systems (GPS, GLONASS, Galileo, BDS), form geographic trajectories (Geo-Trajectories). These trajectories can be formally represented as a sequence of visited points, Geo_Traj={vi}Geo_Traj={vi}, where each point vi=(ti,li), vi=(ti,li) consists of a timestamp ti and a geographic coordinate li=(lati, longi, li=(lati, longi).

The check-in locations within a user's mobile trajectory can be broadly classified into two types: moving points and staying points. Moving points describe the path, speed, and distance of travel. In contrast, the purpose of a user's social travel is often revealed by the functional attributes of the places where they stay—for instance, sleeping at home at night, working at an office during the day, or shopping at a mall. This paper focuses exclusively on stay locations and

the transitions between them, disregarding moving points and the specific geographic paths taken by the user [3]. We annotate the behavioral semantics of a user by analyzing the functional attributes of their identified stay locations.
User social travel typically involves two forms of staying: remaining at a specific location for a prolonged period, or wandering within a defined area. Based on these patterns, we formally define the concept of a stay point and describe a method for extracting these points from user travel trajectories.

2.2 Behavioral Semantic Annotation

We infer the semantic functional attributes of stay points by leveraging open-source web data related to the corresponding locations or regions. This process facilitates the discovery of a user's travel purpose at each stay point and reveals the underlying behavioral semantics of their social trips. The assignment of such semantic properties to user -visited locations is commonly referred to as semantic annotation.
One approach to semantic annotation involves analyzing the functional properties of Points of Interest (POIs) near a user's stay location using map applications—such as Google Earth, Baidu Map, or OpenStreetMap [11]. In this method, typical semantic categories are selected and applied to label the stay location. Common semantic selection algorithms include proximity-based principles, quantity-first selection, and TF-IDF.
Alternatively, open-source user-generated content from social network platforms—such as Foursquare, Instagram, Sina Weibo, Meituan, and RenRen—can be utilized [7,13]. This includes travel check-ins, location-related tweets, comments, reviews, and ratings. Through semantic analysis of such social data, the functional attributes of a user's stay location can be effectively derived.
These procedures enable the transformation of raw geographic trajectories into behavior-oriented sequences composed of semantic attributes from stay points. This structured representation supports the characterization of users' social mobility patterns and enables in-depth behavioral semantic analysis.

2.3 Hierarchical Semantic Similarity Architecture

The semantic-functional properties of Points of Interest (POIs) in the real world typically exhibit a hierarchical similarity structure. For instance, junior high schools, high schools, and universities all fall under the broader category of educational institutions, while also encompassing more specific subtypes. Similarly, Chinese and Western cuisine belong to the food category; shopping malls and supermarkets can be grouped under shopping; and activities such as dining, shopping, KTV, bars, and cinemas are often classified as entertainment.
Building upon these observations, we construct a hierarchical tree structure to characterize semantic similarities among POIs. First, the specific semantic functional attributes of POIs are treated as leaf nodes in the tree. We then perform semantic generalization and clustering at the finest possible granularity, grouping entities such as KTV, bars, cinemas, and supermarkets into progressively broader categories. Through iterative application of this process, a hierarchically organized semantic tree is ultimately established.
This semantic tree captures similarity relationships among semantic attributes through its layered organization, analogous to kinship structures in human society. It provides a powerful technical framework for conducting semantic analysis in mobile computing scenarios.

2.4 Association Representation of Semantic Similarity

To characterize the mapping relationships of behavioral semantics between synthetic and actual trajectories, we quantify location semantic similarities using the constructed semantic tree. This quantification establishes the basis for evaluating privacy leakage risks in trajectory publication.
The semantic similarity quantification process proceeds as follows: First, we select an arbitrary leaf attribute as the starting point and traverse upward through the semantic tree layer by layer. During this backtracking process, we cluster the newly encountered attributes ($A\tau$) from the leaves of each intermediate node. The mapping probability from the departure attribute to each newly traversed attribute, denoted as $Ri(oj)$ is calculated to be inversely proportional to the size of $A\tau$ and decays with the number of iterations.
This upward traversal continues until the root node is reached, establishing semantic similarity association probabilities from the departure node to all other attributes. Finally, we systematically replace the departure attribute and repeat the entire traversal process until every attribute has served as a starting point. Through this comprehensive procedure, we obtain complete semantic similarity associations between all attribute pairs.

3 BEHAVIORAL INFERENCE ASSOCIATED WITH SEMANTIC SIMILARITY OF PUBLISHED TRAJECTORY

We represent the stochastic process of user social mobility following the principle of HMM, and reveal the risks of privacy leakage induced by the semantic similarity of published trajectories to users' real behaviors.

3.1 User Behavioral-Pattern Construction

Following data preprocessing and behavioral semantic annotation, we obtain the user's behavioral trace in mobile scenarios. To model these traces, we construct behavioral patterns of user social travel using a supervised learning approach based on the Markov Chain principle.

Two key challenges must be addressed: first, user travel behavior is closely tied to their current life-state; second, it exhibits strong temporal dependencies. For instance, users typically return home at night, work during daytime hours, and engage in leisure activities during holidays. Traditional Markov models fail to capture these complex spatio-temporal-life-state associations, focusing solely on spatial transitions [14].

To overcome these limitations, we enhance the Markov model by constructing a time partition-based extension matrix M~ {aij} and multiple life-state-specific matrices AM~ Mls. This extension incorporates temporal dimensions into the spatial transitions, representing each transition as aij~ ((ti,si)→(tj,sj)) rather than simply (li,lj). This formulation captures both the user behavior and its temporal context. Additionally, by maintaining separate matrices for different life-states, we prevent interference between distinct behavioral patterns during characterization.

3.2 Behavioral-Semantic Mapping Associations between Published and Actual Trajectories

Semantic similarity constitutes the fundamental basis for mapping associations Ri(oj) between user behaviors in published and actual trajectories. As established in Section 3.1, we characterize these mapping associations by incorporating the computed semantic similarities.

The association probabilities Ri(oj) between any two attributes are derived through backtracking operations on the semantic tree. We construct a two-dimensional observation matrix BM={Ri(oj)} for our Hidden Markov Model (HMM) using these probabilities. This matrix effectively represents the semantic associations between published observation attributes oj and potential actual behavioral semantics si.

3.3 Social-Mobility Stochastic Processes

Building upon the constructed HMM, we model the stochastic process of invisible user social mobility in mobile scenarios. We define an HMM-based forward variable αα and specify its computational method, describing its evolution process. During evolution, target forward variables are computed and intermediate results are stored in a matrix structure. These stored results subsequently support efficient behavioral inference while reducing computational overhead.

3.4 Observation-based Behavioral Inference

In mobile social applications, attackers may observe users' shared location trajectories. Through behavioral pattern analysis of published trajectories, they can infer users' upcoming behaviors. This section simulates such inferential attacks to quantify the mobile privacy leakage risk associated with trajectory publication.

As published trajectories consist of sequentially shared access locations, we dynamically quantify privacy leakage risk by characterizing the attacker's inference probability before and after each location observation. This iterative approach ultimately achieves comprehensive privacy risk assessment for the complete published trajectory.

Let $P^{-}(s)$ represent the attacker's prior inference probability about the user's actual behavior before observing a released location, and $P^{+}(s \mid z)$ denote the posterior probability after observation. We further transform these formulae using the characterized HMM-based social mobility and the defined forward variable:

$\Delta = P^{+}(s|z) - P^{-}(s)$

$P^{+}(s|z) = p(st|o1,\cdots,ot-1,ot= z,\lambda)$

$=p(o1,\cdots,ot-1,ot= z,st|\lambda)/p(o1,\cdots,ot-1,ot= z|\lambda)$

$P^{-}(s) = p(st|o1,\cdots,ot-1,\lambda)$

$=p(o1,\cdots,ot-1,st|\lambda)/p(o1,\cdots,ot-1|\lambda)$

4 EXPERIMENTAL EVALUATION

4.1 Datasets and Setup

Mobile Dataset: We utilized the real-world Geolife dataset from the Microsoft Research Asia project led by Yu Zheng's group to evaluate the performance of BSPri. This dataset captures daily life trajectories of 182 users over a five-year period, primarily within Beijing. Our experiments focus on mobility data inside Beijing's Sixth Ring Road.

Open-source Libraries: Through Baidu Map API, we collected public POIs within the target area to construct a POI Library containing 90,494 entries. Their semantic attributes were extracted to form an open-source Semantic Library with 44 standardized categories. Using time/distance thresholds of 1min/10m and 30min/100m, we identified 9,495 stay points from the Geolife dataset, establishing a Staying-points Library as personal mobility data.

Privacy-preservation Setup: Our semantic generalization mechanism for trajectory reconstruction operates through backtracking on the semantic tree during similarity characterization. This process uses leaf attributes of intermediate nodes reached after backtracking as candidate anonymous semantic types, with backtracking levels indicating the degree of semantic generalization. We conducted experiments to quantify behavioral semantic leakage risks across different generalization levels.

4.2 Experimental Results of Privacy Risks

We deployed inference algorithms in scenarios both with and without life-state differentiation to assess travel behavior leakage risks during weekdays and holidays. The following figures presents privacy leakage metrics through both dynamic progression and average values.

Comparative analysis of Figures 1 and 2 reveals two key findings: First, with one-level backtracking, consistently high metric values indicate that weak semantic generalization significantly increases attackers' inference probability, substantially exposing travel privacy. Second, while three-level backtracking effectively constrains overall metric values, sporadic peaks suggest the necessity of excluding high-similarity attribute types in privacy protection design - a consideration we incorporate in subsequent work.

Figure 3 demonstrates that average inference probabilities decrease substantially with increasing backtracking levels, confirming semantic generalization as an effective countermeasure against behavioral-semantic inference attacks. Although non-life-state-differentiated curves closely resemble weekday patterns, holiday curves exhibit lower inference probabilities and more pronounced variation trends under semantic generalization, reflecting users' diverse behavioral patterns during holidays.

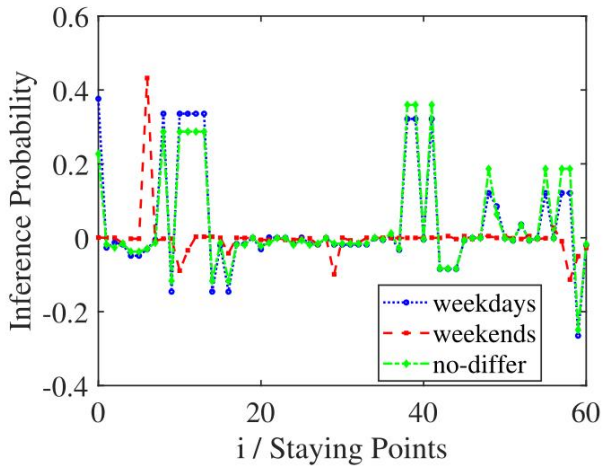

Figure 1 Dynamics of Inference Probability in Different Scenarios with bk = 1

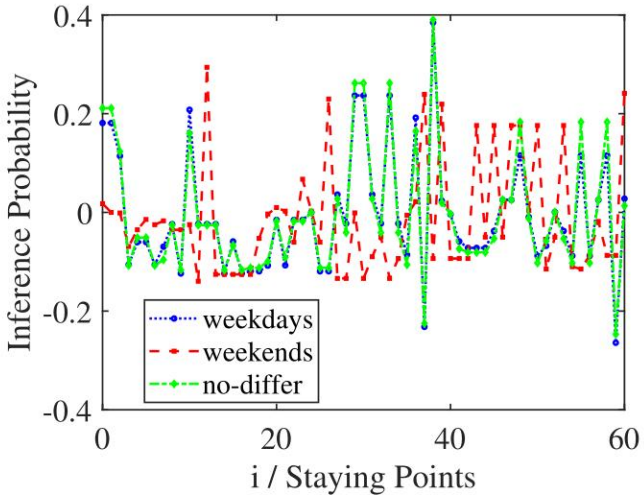

Figure 2 Dynamics of Inference Probability in Different Scenarios with bk = 3

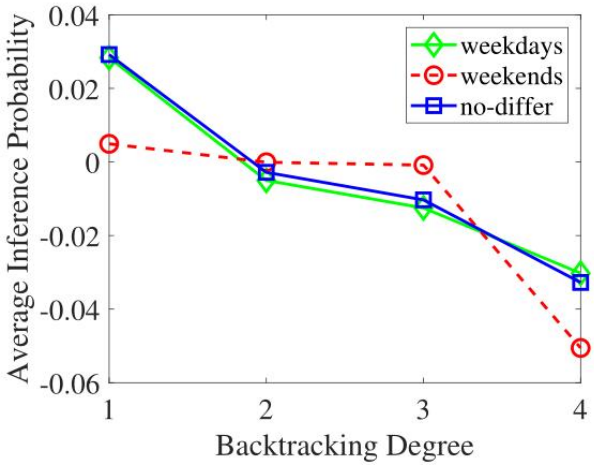

Figure 3 Average Inference Probability in Different Scenarios

5 CONCLUTION

Addressing the privacy concerns in mobile trajectory publication, this paper presents a privacy risk quantification mechanism. We simulate inference attacks by developing an observation-based behavioral inference algorithm, which characterizes the privacy leakage risk stemming from the semantic similarity between published trajectories and users' actual behaviors. Extensive experiments on real-world datasets validate our approach by demonstrating the leakage risks under varying degrees of semantic generalization, thereby providing a foundation for designing subsequent semantic privacy protection schemes.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Qiu G, Tang G, Li C, et al. Differentiated location privacy protection in mobile communication services: A survey from the semantic perception perspective. ACM Computing Surveys(CSUR), 2023, 56(3): 1-36.
[2] Jiang H, Li J, Zhao P, et al. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. ACM Computing Surveys(CSUR), 2021, 54(1): 1-36.
[3] Zheng Y. Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology(TIST), 2015, 6(3): 1-41.
[4] Jin X, Zhang R, Chen Y, et al. Dpsense: Differentially private crowdsourced spectrum sensing. In Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security(CCS), Vienna, Austria, 2016: 296-307.
[5] Primault V, Boutet A, Mokhtar SB, et al. The long road to computational location privacy: A survey. IEEE Communications Surveys and Tutorials, 2019, 21(3): 2772-2793.
[6] Zhang J, Li C, Wang B. A performance tunable cpir-based privacy protection method for location based service. Information Sciences, 2022, 589: 440-458.
[7] Zhao W, Zhou N, Zhang W, et al. A probabilistic lifestyle-based trajectory model for social strength inference from human trajectory data. IEEE Transactions on Information System, 2016, 35(1): 1-28.
[8] Song C, Raghunathan A. Information leakage in embedding models. In Proceedings of the 27th ACM SIGSAC Conference on Computer and Communications Security(CCS), Gather.town, virtual platform, 2020: 377-390.
[9] Andrés ME, Bordenabe NE, Chatzikokolakis K, et al. Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, 2013: 901-914.
[10] Dai Y, Shao J, Zhang D. Personalized semantic trajectory privacy preservation through trajectory reconstruction. World Wide Web, 2018, 21(4): 875-914.
[11] Qiu G, Guo D, Shen Y, et al. Mobile semantic-aware trajectory for personalized location privacy preservation. IEEE Internet of Things Journal, 2020, 8(21): 16165-16180.
[12] Qiu G, Tang G, Li C, et al. Behavioral-semantic privacy protection for continual social mobility in mobile internet services. IEEE Internet of Things Journal, 2024, 11(1): 462-477.
[13] Yang C, Sun M, Zhao WX, et al. A neural network approach to jointly modeling social networks and mobile trajectories. IEEE Transactions on Information System, 2017, 35(4): 1-28.
[14] Phan N, Wang Y, Wu X, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Arizona USA, 2016: 1309-1316.