Summer Internship Project

Report

Optimal privacy protection of mobility data

Submitted by

Arthur GOARANT

Under the guidance of

Sophie Cerf, Adrien Luxey, Rémy Raës

Spirals team

Inria Lille

Summer Internship 2023

Abstract

The widespread usage of location data in delivering geo-personalized content to mobile device users is raising increasing concerns regarding privacy. Geolocation attacks, such as Point of Interest (POI) attacks, exploit this data to extract personal information about users.

To address this issue, various protection mechanisms have been developed, including the introduction of noise into actual trajectories to prevent the retrieval of POIs through such attacks.

A recent approach [1] suggests using the information of the future positions of the user to increase privacy, though it is currently restricted to an offline implementation.

In this report, we present a novel methodology that repurposes the FLAIR [2] storage system into a mobility predictor and combines it with the p mpc-H [1] method to establish an online privacy protection mechanism. By integrating these techniques, we aim to optimize the privacy preservation of mobility data while maintaining real-time applicability.

Contents

Chapter 1

Introduction

In the context of ever-increasing applications using mobility data, concerns have been raised about the potential threat to privacy of their users. By leveraging user positions and information about the time spent at each location, it becomes possible to extract points of interest (POIs) and access sensitive information. The identification of POIs is typically based on defining specific regions where users spend a significant amount of time. For example, a POI attack algorithm [3] [4] has the potential to extract sensitive information such as an individual's home, workplace, or preferred places of leisure. To mitigate this privacy risk, several privacy protection mechanisms have been developed, some of which operate only offline (e.g., Promesse [5]and the current version of mpc), while others offer online protection (e.g., geo-I [6]).

These mechanisms commonly employ obfuscation techniques, where location data is intentionally perturbed with spatial noise before transmission to the service. For instance, geo-I applies time-dependent only noise to the user's position. Recent advancements, such as the p mpc-H approach, have exploited the unique characteristics of location data, such as temporal correlation and human mobility patterns and especially by utilizing future mobility prediction, to achieve enhanced privacy protection. However, the current implementation of p mpc-H is limited to offline computation, relying on the user's future positions to improve privacy.

Building upon this previous work, a logical progression is to develop a user position predictor and integrate it into the p mpc-H algorithm. This is the main objective of this internship.

To the best of our knowledge, while some papers provide machine learning based algorithms to predict future locations, they often provide predictions in the form of a location index rather than precise x, y coordinates, which are necessary in order to use the p mpc-H algorithm. Therefore, it is needed to develop an estimator that meets the requirements of our approach.

In this study, we define privacy in terms of data distortion [7], while utility quantifies the accuracy or fidelity of the obfuscation process.
The objective of our research is to maximize privacy while maintaining a convincing level of utility. Privacy preservation is of paramount importance in protecting sensitive information derived from mobility data. However, it is equally crucial to strike a balance by ensuring that the utility provided by the application or service is maintained at an acceptable level. The desired level of utility may vary depending on the specific application, as different scenarios might have different requirements and trade-offs between privacy and utility. Additionally, we aim to analyze the impact of lower accuracy mobility predictions on the effectiveness of the mpc-H approach and observe if its great results are still maintained. Moreover, we consider the future extension of this system to mobile phone usage, although its feasibility requires further examination.
Furthermore, this research remains aligned with the objective of eventually extending the system to encompass mobile phone usage, although the feasibility of such an extension requires further investigation.

Chapter 2

Related works

In this chapter, we review several existing works that aim to protect privacy in the context of mobility data. While each approach shares the common goal of preserving privacy, they employ different techniques and exhibit variations in their implementation capabilities.
Additionally, we will explore related works in the field of mobility prediction to explore the different mobility predictors we could use for an online implementation of the p mpc-H algorithm.

2.1 Common definitions

We first provide definitions for several key concepts used throughout the paper. These definitions establish a common understanding of terms and metrics employed in the context of privacy protection of mobility data.

Points of Interest (POI): A Point of Interest refers to a specific region in which a user spends a significant amount of time. The exact size and duration required to qualify as a POI can vary depending on the specific application or context. In this paper, we define a POI as a region where a user stays for a certain duration, which will be further specified based on the experimental setup.

Privacy: Privacy, in the context of this paper, refers to the degree to which an algorithm or attacker can detect points of interest (POIs) with or without the application of an obfuscation algorithm. To quantitatively measure privacy, we define a spatial distortion metric. While various definitions of privacy are possible, such as comparing the number of obtained POIs before and after applying obfuscation, we adhere to this particular definition for several reasons. Firstly, it provides a mathematically well-defined metric that facilitates the formulation and optimization of privacy-preserving algorithms. Additionally, this definition aligns with the privacy metric used by the p mpc-H algorithm, which serves as the basis for our online implementation.

Utility: Utility measures the extent to which the obfuscated position generated by a privacy protection mechanism aligns with the actual position of the user. It quantifies the accuracy or fidelity of the obfuscation process. A commonly used metric for utility is the mean distance between the obfuscated position and the real position of the user.

2.2 Tools to ensure privacy

One notable work is the PROMESSE algorithm, which focuses on smoothing GPS traces both temporally and geographically to remove points of interest (POIs) from the input trace. This method operates in an offline manner and involves the deletion of certain temporal information. While PROMESSE successfully mitigates the risk of POI extraction, its offline implementation restricts its applicability in real-time scenarios.

Another approach, geo-I, adopts a blind obfuscation technique that applies time-dependent noise to all locations and at all times, following a specific formula. This method enables online implementation and provides a certain level of privacy preservation. However, its effectiveness and utility may vary depending on the specific noise formula employed.

The p mpc-H algorithm represents another notable contribution, known for its impressive results in privacy protection. It operates on the assumption of having access to the entire trajectory, as it relies on predicting the future positions of the user. Currently, p mpc-H is implemented offline due to the requirement of knowing the future positions. This approach demonstrates strong privacy preservation capabilities but lacks the feasibility of real-time application.

While these existing works offer valuable insights into privacy protection mechanisms, they predominantly focus on either offline implementations or specific obfuscation techniques. In contrast, our study aims to leverage the FLAIR algorithm as a predictor within the p mpc-H framework, enabling online implementation and exploring the potential of combining predictive capabilities with privacy preservation. By employing FLAIR's trajectory modeling and prediction abilities, we seek to achieve effective privacy protection while maintaining satisfactory utility in real-time scenarios.

2.3 Mobility prediction models

Mobility predictions can be made at various levels of granularity, including user-wise, point of interest (POI)-wise, and trajectory-wise. User-wise predictions aim to estimate the user's next location based on their historical behavior and other contextual information, such as time of day and weather. POI wise predictions, on the other hand, aim to predict the likelihood that a user will visit a specific POI, such as a restaurant or a museum. Finally, trajectory-wise predictions aim to estimate the user's full trajectory or route over a given period of time. In our case, our goal was to predict the user's next location accurately within a short time frame of 5 minutes or so, which falls under the user-wise prediction category.

To identify existing methods for human mobility prediction, we conducted a literature survey of recent articles, emphasizing more recent methodologies proposed and discussed in the field. A comprehensive survey [8]that references numerous papers on mobility prediction exists, encompassing diverse prediction methods and levels.

We limited the time range for the publication date of the articles to ensure relevance to current research and advancements in the field. It is worth noting that while recent models often employ machine learning techniques for mobility prediction, their objectives do not entirely align with ours. Most existing models provide predictions in the form of a region index, indicating the next region the user is likely to be in. However, for our purpose of combining a mobility predictor with the p mpc-H framework, accurate x, y coordinates are essential.

One notable paper in this domain is DeepMove [9], which utilizes a neural network algorithm to predict human mobility patterns. GCDAN [10], on the other hand, employs a recurrent neural network architecture for mobility prediction. Additionally, WhereNext [11] utilizes decision trees to choose the most likely continuation of the current trajectory.However, these methods often require predictions to be formulated as region indexes, which does not align with the specific aim of our research.

Chapter 3

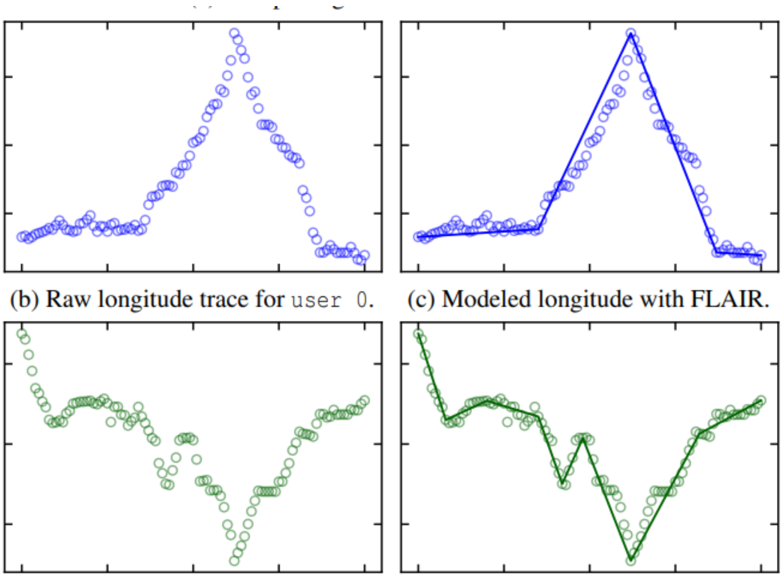
Background

In this chapter, we provide the necessary background information to understand the motivation behind combining the p mpc-H algorithm with the FLAIR predictor in our proposed privacy protection mechanism. Our ultimate objective is to develop an efficient and effective approach that maximizes privacy (denoted as p) while maintaining a satisfactory level of utility (denoted as u).

3.1 FLAIR

FLAIR is a storage system based on a piece-wise linear approximation technique. It enables efficient compression of mobility data and facilitates modeling of a user's trajectory with limited memory usage.
Each trajectory transformed by FLAIR comprises a collection of ”models”, which consist of a starting point and a linear coefficient. By appropriately selecting the FLAIR parameter epsilon, these models can accurately represent the user's trajectoryFig. 3.1 .

We typically denote a model as (A,x,t)
- A being the linear coefficient of the current model
- x, the position at which the model starts
- t, the time at which the model starts
Because FLAIR's approximation is continuous by definition, the last model ends when a new one starts. Thus, there is no need to specify any ending time or point for the model.



(b) Raw longitude trace for `user 0`.

(c) Modeled longitude with FLAIR.

(d) Raw latitude trace for `user 0`.

(e) Modeled latitude with FLAIR.

Figure 3.1: FLAIR compacts any location stream as a sequence of segments, obtained from a piece-wise model.

3.2 The p mpc-H algorithm

We denote the horizon H as the number of next positions we want to predict. The p mpc-H algorithm is designed to optimize privacy protection while maintaining a lower bound on utility (i.e while staying as close as possible to the real position of the user). It operates on a received sample provided by the user, assuming that the next H samples are known (which implies an offline implementation). The algorithm tackles the problem of finding the optimal obfuscated position that maximizes privacy, as measured by a specific privacy metric, while ensuring a certain level of utility.
To achieve this, p mpc-H formulates an optimization problem to find the obfuscated position that maximizes privacy while respecting our constraints whenever it receives a sample (for every k). It basically looks for the obfuscated position $\tilde{z}$ which is going to maximize privacy, knowing the H future positions, while maintaining our utility below a certain level:

$$\max_{(\delta x_i, \delta y_i)_{i=1}^{H} \in \mathbb{R}^H \times \mathbb{R}^H} \qquad \sum_{j=1}^{H} p(\tilde{z}(k+j))$$

$$\delta x_i^2 + \delta y_i^2 \leq \Delta^2$$

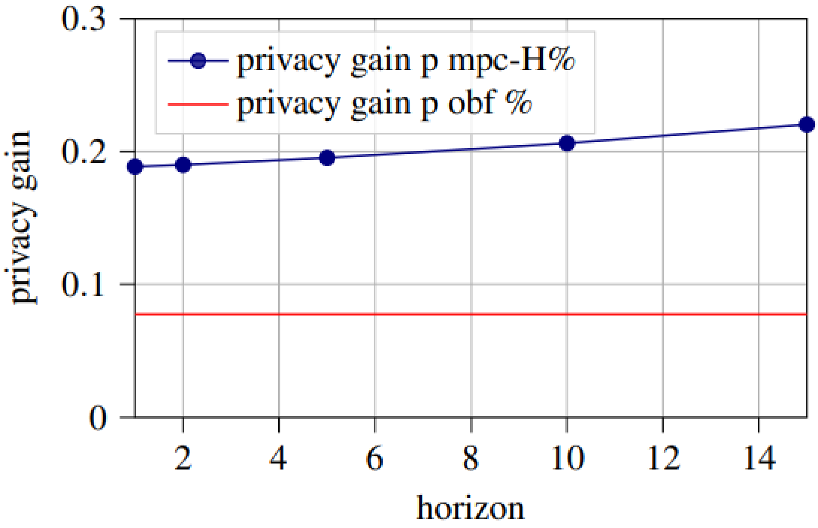


Figure 3.2: p mpc-H's performances as a function of the horizon H, as compared to geo-I

The p mpc-H algorithm manages to reach convincing results even with smaller values of horizons. Fig. 3.2 Since we could expect FLAIR's predictor to exhibit relatively decent performance for smaller horizon values, we could hope for satisfactory enough results.

Nonetheless, it remains crucial to acknowledge that higher horizon values may lead to significantly diminished prediction quality with FLAIR. Therefore, expecting similar performance for higher horizon values would be overly optimistic given FLAIR's limitations.

Chapter 4

System overview

4.1 Predicting with FLAIR

Due to its piece-wise linear approximation nature, FLAIR can also be leveraged as a predictor by simply extending the last supplied model to forecast the user's next positions.
In order to deal with constant sampling frequencies, we first harmonize the trajectory data: the received samples within a 30-second interval (referred as the harmonized sampling frequency) are averaged and fused into a single representative sample. In cases where no data is received during that interval, the corresponding sample is marked as not received.
The next k predictions are then computed using $x + (k * sampling_time) * A$.
Although this prediction approach is relatively simplistic, it can be remarkably effective for smooth trajectories or when the prediction horizon is small.
By using FLAIR as a predictor, we fulfill the requirement of expressing coordinates and also benefit from substantial memory savings, as only a limited number of models needs to be stored. Within our aim of one day using an obfuscation algorithm directly from the phone, this can prove very useful and may not always be reached using more complex prediction algorithms, or without a third-party.

4.2 Combining p mpc-H and FLAIR's prediction for an online obfuscation algorithm

With our newly developed mobility predictor in place, we can seamlessly integrate it into the code of the pmpc-H algorithm by replacing the future positions of the user with our homemade predictions. By directly plugging in our trajectory predictions into the algorithm, the optimization problem is solved using the anticipated positions of the user.

Our objective is then to examine the performance of this combined approach and compare it with state-of-the-art algorithms, such as geo-I. This analysis allows us to assess the potential benefits of leveraging FLAIR in comparison to existing approaches and gain insights into the trade-off between utility and privacy preservation.

Chapter 5

Results

5.1 FLAIR's predictive performances

We proceed to evaluate the prediction performance of FLAIR using the cabspotting dataset. To ensure comparability, we transformed it using the pymap library to express coordinates in meters. We present the results in a comparative table, as a function of the prediction horizon (Table 5.1)

While FLAIR as a predictor generally demonstrates respectable performance for smaller horizon values, its efficacy notably diminishes when faced with abrupt sharp turns in trajectories at the moment of the prediction or for larger horizon values(Fig. 5.1).

5.2 p mpc-H with FLAIR's predictions

Our evaluation reveals that the performance of our combined FLAIR-p mpc-H approach is highly dependent on the chosen value of the horizon. Notably, for smaller horizon values, our model almost consistently outperforms geo-I in terms privacy (Fig. 5.2) .This outcome can be attributed to FLAIR's ability to capture and predict the user's trajectory within shorter prediction horizons more accurately. However, as the horizon value increases, the quality of FLAIR's predictions

Horizon	1	3	5	10	15
mean X error	20.3	46.8	68.5	304.2	602.3
mean Y error	17.2	51.2	100.4	374.2	704.9

Table 5.1: FLAIR's mean errors as a predictor (on one cabspotting trajectory)

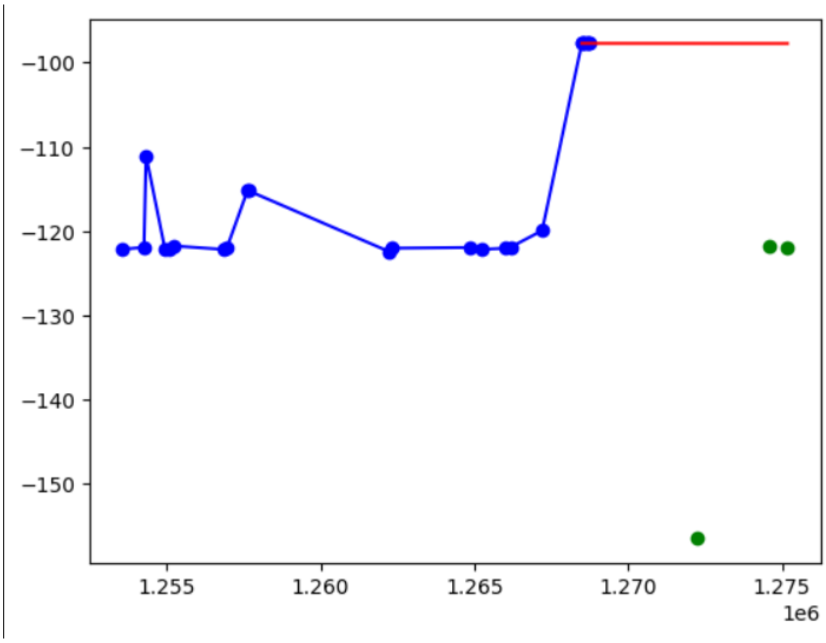


Figure 5.1: FLAIR's poor predictive performances for a sharp turn.

decreases and leads to reduced performance compared to geo-I (Fig. 5.3) .Thus, careful consideration of the horizon value is crucial to ensure optimal results when using our approach. Additionally, the offline implementation of p mpc-H consistently outperforms both geo-I and the online implementation, which was to be expected.

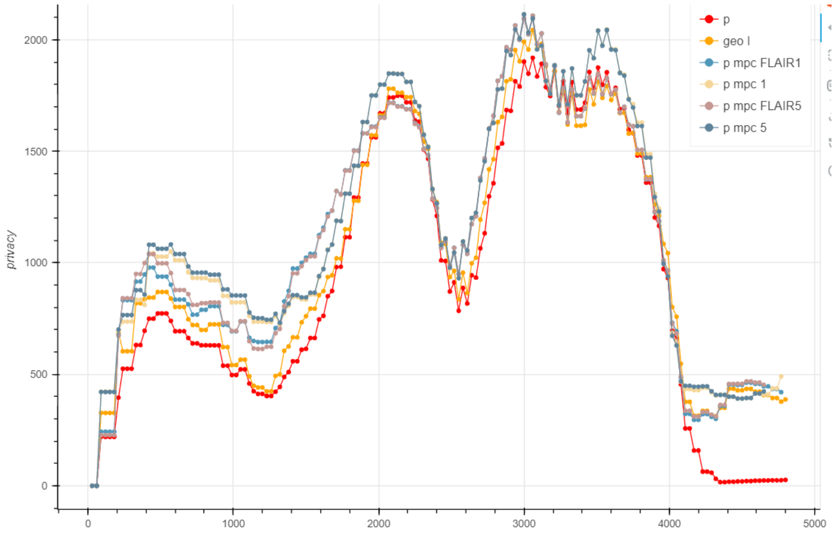


Figure 5.2: privacy values for small values of horizon

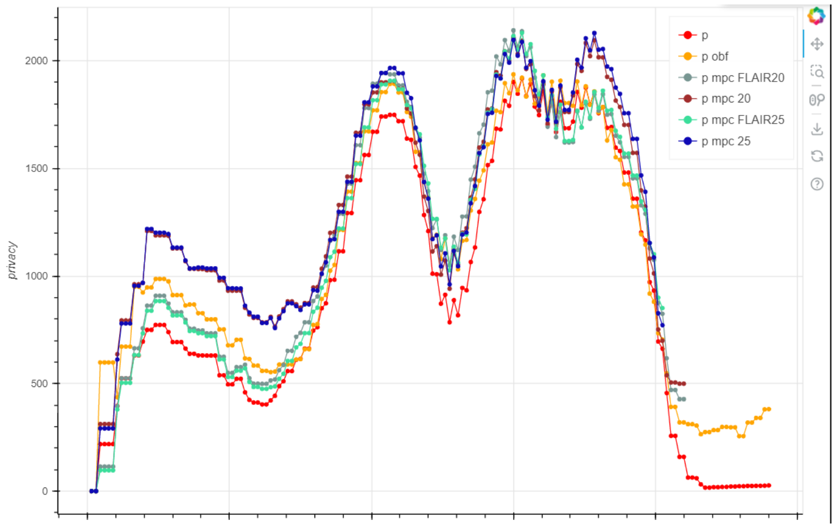


Figure 5.3: privacy values for larger values of horizon

Chapter 6

Conclusion

These results demonstrate the feasibility of utilizing FLAIR as a predictor within the p mpc-H framework to achieve effective privacy protection in an online scenario, reaching more convincing results than the state of the art Geo-I algorithm.
While our results demonstrate the effectiveness of our online implementation, there remain promising avenues for future research. As this work was mostly conducted on a single trajectory, further analysis on diverse trajectories and datasets would provide valuable insights into the specific scenarios where our model outperforms geo-I.
Additionally, it would be insightful to examine further cases where the privacy metric falls below a certain threshold, as this indicates a possibility to identify a POI and therefore a significant privacy threat. While our code provides the possibility to highlight such situations, the chosen privacy limit is arbitrary. It could be interesting to link this value to a concrete possibility of identifying a POI using a POI attack algorithm. This way, it would be possible to accurately identify the situations where a POI of a certain duration and certain radius can be obtained.
Our code and analysis remain available on github

Acknowledgment

I would like to thank the entire team of Spirals where I had the privilege of completing my internship, for their warm and welcoming environment.

I would specifically express my deepest gratitude towards my three tutors Sophie, Adrien and Rémy. Their time, patience, and pedagogy have played a pivotal role in shaping my understanding of the subject and improving my work. I am truly grateful for their kindness and willingness to share their knowledge and expertise.

I have spent a wonderful internship and couldn't be happier of all the things I have learned and the people I have met.

Bibliography

[1] E. Molina, M. Fiacchini, S. Cerf, and B. Robu, "Optimal privacy protection of mobility data: a predictive approach," in IFAC WC 2023 - 22nd IFAC World Congress, (Yokohama, Japan), July 2023.

[2] Anonymous, "FLAIR: Storing unbounded data streams on mobile devices to unlock user privacy at the edge," in Submitted to Journal of Systems Research - March 2023, 2023. under review.

[3] R. Hariharan and K. Toyama, "Project lachesis: Parsing and modeling location histories," vol. 3234, pp. 106–124, 10 2004.

[4] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personally meaningful places: An interactive clustering approach," ACM Transactions on Information Systems, vol. 25, July 2007.

[5] V. Primault, S. B. Mokhtar, C. Lauradoux, and L. Brunie, "Time distortion anonymization for the publication of mobility data with high utility," 2015.

[6] N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," 11 2014.

[7] S. Cerf, B. Robu, N. Marchand, S. B. Mokhtar, and S. Bouchenak, "A control-theoretic approach for location privacy in mobile applications," in 2018 IEEE Conference on Control Technology and Applications (CCTA), pp. 1488–1493, 2018.

[8] H. Zhang and D. Lingcheng, "Mobility prediction: A survey on state-of-the-art schemes and future applications," IEEE Access, vol. PP, pp. 1–1, 12 2018.

[9] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," Proceedings of the 2018 World Wide Web Conference, 2018.

[10] W. Dang, H. Wang, S. Pan, P. Zhang, C. Zhou, X. Chen, and J. Wang, "Predicting human mobility via graph convolutional dual-attentive networks," in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (P. Nagarkar, ed.), (United States of America), pp. 192–200, Association for Computing Machinery (ACM), 2022. Publisher Copyright: © 2022 ACM.; ACM International Conference on Web Search and Data Mining 2022, WSDM 2022 ; Conference date: 21-02-2022 Through 25-02-2022.

[11] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: A location predictor on trajectory pattern mining," pp. 637–646, 06 2009.