TransRisk: Mobility Privacy Risk Prediction based on Transferred Knowledge

XIAOYANG XIE, Rutgers University, USA
ZHIQING HONG, Rutgers University, USA
ZHOU QIN, Rutgers University, USA
ZHIHAN FANG, Rutgers University, USA
YUAN TIAN, University of Virginia, USA
DESHENG ZHANG, Rutgers University, USA

Human mobility data may lead to privacy concerns because a resident can be re-identified from these data by malicious attacks even with anonymized user IDs. For an urban service collecting mobility data, an efficient privacy risk assessment is essential for the privacy protection of its users. The existing methods enable efficient privacy risk assessments for service operators to fast adjust the quality of sensing data to lower privacy risk by using prediction models. However, for these prediction models, most of them require massive training data, which has to be collected and stored first. Such a large-scale long-term training data collection contradicts the purpose of privacy risk prediction for new urban services, which is to ensure that the quality of high-risk human mobility data is adjusted to low privacy risk within a short time. To solve this problem, we present a privacy risk prediction model based on transfer learning, i.e., TransRisk, to predict the privacy risk for a new target urban service through (1) small-scale short-term data of its own, and (2) the knowledge learned from data from other existing urban services. We envision the application of TransRisk on the traffic camera surveillance system and evaluate it with real-world mobility datasets already collected in a Chinese city, Shenzhen, including four source datasets, i.e., (i) one call detail record dataset (CDR) with 1.2 million users; (ii) one cellphone connection data dataset (CONN) with 1.2 million users; (iii) a vehicular GPS dataset (Vehicles) with 10 thousand vehicles; (iv) an electronic toll collection transaction dataset (ETC) with 156 thousand users, and a target dataset, i.e., a camera dataset (Camera) with 248 cameras. The results show that our model outperforms the state-of-the-art methods in terms of RMSE and MAE. Our work also provides valuable insights and implications on mobility data privacy risk assessment for both current and future large-scale services.

CCS Concepts: • **Networks** → **Sensor networks**; • **Information systems** → *Location based services.*

Additional Key Words and Phrases: Heterogeneous Datasets, Mobility Patterns, Privacy

ACM Reference Format:
Xiaoyang Xie, Zhiqing Hong, Zhou Qin, Zhihan Fang, Yuan Tian, and Desheng Zhang. 2022. TransRisk: Mobility Privacy Risk Prediction based on Transferred Knowledge. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 82 (June 2022), 19 pages. https://doi.org/10.1145/3534581

Authors' addresses: Xiaoyang Xie, xiaoyang.xie@rutgers.edu, Rutgers University, New Brunswick, New Jersey, USA; Zhiqing Hong, zhiqing.hong@rutgers.edu, Rutgers University, New Brunswick, New Jersey, USA; Zhou Qin, zhou.qin@rutgers.edu, Rutgers University, New Brunswick, New Jersey, USA; Zhihan Fang, Zhihan.fang@rutgers.edu, Rutgers University, New Brunswick, New Jersey, USA; Yuan Tian, yuant@virginia.edu, University of Virginia, Charlottesville, Virginia, USA; Desheng Zhang, desheng.zhang@cs.rutgers.edu, Rutgers University, New Brunswick, New Jersey, USA.

2474-9567/2022/6-ART82 $15.00
https://doi.org/10.1145/3534581

1 INTRODUCTION

Many large-scale urban services have explicitly or implicitly collected anonymized mobility data of residents to understand personal mobility patterns, e.g., cellphone data (CDR and Connection) [27][28], vehicular GPS data (Vehicles) [34], and electronic toll collection data (ETC) [36]. Unfortunately, while providing the improvement for urban services, the collection of human mobility data is sensitive because it is shown by the previous works [9][10] that it is possible to re-identify users from the dataset. So before the large-scale data collection, for privacy protection, the operator may need to understand the privacy risk by utilizing the privacy risk measurement approaches, i.e., the attack models, which identify the probability of a user's trace being exposed by an attack (see formal definition in Section 2.2). However, it is challenging to efficiently assess the privacy risk of users by privacy risk measurement approaches because most of them are based on the high time complexity matching functions.

To solve the efficiency issue, some works have been proposed based on the predictive Machine Learning model to approximately measure the privacy risk. For example, Pellungrini *et al.* [25] presents a privacy risk prediction model based on Random Forest to predict the privacy risk with the extracted mobility features of the users from vehicular data; Rocher *et al.* [29] utilizes a generative model to estimate the success of re-identifying users from the crowd given incomplete demographic data. Nevertheless, most existing predictive approaches measuring the privacy risk are based on long-term data (e.g., one-month data) to achieve a higher prediction accuracy. However, for a newly deployed urban service, if the service operators want to set up the spatial-temporal granularity of the collected data to balance the data utility and potential privacy risks, the ideal case is to collect short-term data to assess privacy risks first without direct large-scale fine-grained data collection (details in Section 2.3). In this case, the existing works were not applicable given their requirement for long-term training data.

One solution is to use data already collected in other services, complementing short-term data of a new service to assess the privacy risks of this new service without collecting large-scale data of this new service. In particular, lots of data from existing urban services have been collected and shared with the research community, and the privacy risk of data from these services has been properly assessed (e.g., cellphone[9], vehicles[25]). Meanwhile, transfer learning has shown its effectiveness in the privacy and security community. For example, [26] proposes an asymmetric multitask learning approach for image data to transfer a view-invariant representation from the source dataset to an unlabelled target dataset. [18] also presents a deep domain adaptation model to perform the cross-domain person re-identification by utilizing the domain-invariant and specific features of images. Inspired by these works, to explore the feasibility of using the collected data, we present a mobility privacy risk prediction model based on transfer learning called TransRisk. TransRisk predicts the privacy risk of users of a new service based on (1) a *target dataset* with short-term mobility data newly collected from this new service, and (2) a *source dataset* with long-term mobility data legally collected from another service.

However, employing transfer learning for privacy risk assessment also brings some new challenges. (1) Most previous prediction models for privacy risk assessment utilize the aggregate mobility features of users from the dataset to predict the privacy risk. Simply utilizing the same aggregate mobility features may not be effective to transfer learning privacy risk prediction. (We show a detailed analysis in Section 2.4.) (2) Two records from two mobility datasets with similar spatial-temporal information may have different influences on the privacy risks due to user behaviors in different datasets. The model should be able to capture global influences for all spatial-temporal records. For example, a CDR record in a company during office hours may have a lower influence on the user's privacy risk compared to that of a record in late-night, because there might be many phone calls during office hours in a company. While a vehicle record in the same place and at the same time may have a higher influence because there might not be many vehicles driving during office hours, especially in an industrial region.

To address the above challenges, (1) we unify the spatial granularity of all mobility datasets by a spatial-temporal tensor to represent the spatial and temporal information of users, which remains the detailed spatial-temporal information of the user, instead of the aggregate mobility features; (2) we employ an attention-based convolution layer to capture the global influence of each unit in the spatial-temporal tensor.

To evaluate the applicability of TransRisk to real-world applications, we envision that TransRisk could be applied to assist the traffic camera surveillance system in a city. In some cases, the government of a city may utilize the traffic camera surveillance system to capture the vehicle that violates the traffic rules, and publish the information of the violation, including the plate number, location, time, etc [11]. By employing TransRisk, the government could set up a safety data configuration to protect the people or regions with high privacy risks. Hence, in this paper, we regard the camera dataset as the target dataset and evaluate TransRisk on it with various source datasets. The key contributions of this paper are as follows.

- To the best of our knowledge, we conduct the first case study on the mobility privacy risk prediction based on transfer learning across heterogeneous mobility datasets. In particular, our study on five large-scale mobility datasets covers a broad spectrum of spatial and temporal granularity. The large-scale study enables us to compare these datasets and analyze them for valuable insights.

- We present an efficient mobility privacy risk prediction model, TransRisk, to predict the privacy risk of users based on transfer learning. Compared to previous work, TransRisk unifies multiple mobility datasets and employs an additional input, spatial-temporal tensor, to represent the spatial-temporal information of users from mobility data. Also, we design multiple embedding sequential layers for multiple inputs from a source dataset and a target dataset and design multiple loss functions to train our model simultaneously.

- We implement TransRisk in Shenzhen based on at least one month of real-world data from multiple real-world datasets (details in Section 2.1). We evaluate TransRisk with different combinations of existing mobility datasets as source data for comprehensive analyses on target data, i.e., camera data. Our results reveal some valuable insights regarding the impact of source datasets on privacy risk prediction. A comprehensive comparison study with multiple state-of-the-art methods is provided in detail and the results show TransRisk improves the prediction performances in terms of RMSE and MAE, given the short-term data.

The rest of the paper is organized as follows. Section 2 introduces the background and motivation of this paper. Section 3 shows and describes the design of our method. Section 4 shows the evaluation result of our method and the baselines with different metrics and factors. Section 5 reviews the related work. Section 6 presents some discussions about limitations, potential implications, and privacy and ethics. Section 7 concludes the paper.

2 BACKGROUND AND MOTIVATION

In this section, we first introduce the details of the source datasets collected from the real-world urban services for analyses in motivation. Then we show the definitions of re-identification attack and privacy risk. Finally, we investigate the advantage of transfer learning and inputs beyond aggregated mobility features.

2.1 Datasets

We have access to one-month real-world data sets from several service providers and the Shenzhen Committee of Transportation (SCT). As shown in Table 1, we consider four categories of data sets from four urban mobile systems, i.e., (i) vehicles GPS data from a vehicular insurance tracking system, (ii) cellphone CDR data from a cellphone system, (iii) cellphone connection data from a cellphone system, and (iv) toll transaction data from an electronic toll collection system, which detect individual mobility patterns from four combinations of different extents of spatial granularity and temporal continuity. In particular, because the mobility features of users for

each dataset on weekdays and weekends are significantly different, in the rest of this paper, for one-week data, we only consider the data on weekdays.

- Vehicular GPS dataset **(Vehicles)** covers the locations data of 10 thousand vehicles in Shenzhen, where their records were collected every 10 seconds. The data was collected through onboard devices installed inside vehicles, which are mainly used for insurance purposes. This dataset includes the 1.2 TB data from January 2016 to February 2016, containing the GPS locations of all involved vehicles when they are turned on.
- Call detail record dataset **(CDR)** and cellphone connection dataset **(CONN)** include the coarse-grained spatial information (tower locations) and temporal information of cellphone users, which the total size of the dataset is around 650 GB. CDR data was collected when cellphone users use their cellphones for phone calls and sending messages. While CONN data was collected when cellphone users use their cellphones for network connection. Both of them contained more than one million active users in October 2013 from 5 thousand towers in Shenzhen.
- Electronic toll station dataset **(ETC)** includes the coarse-grained spatial information (toll station locations) and the discrete temporal information of toll station transactions for June 2016 which were collected when vehicles enter and left the highway in Shenzhen. In particular, the data size is around 900 MB and 85 toll stations were capturing 6.5 thousand vehicles per hour.

Table 1. Details of Datasets

Mobility Datasets	# of ID	# of Days	Daily Records	Spatial Granularity	Temporal Continuity
Vehicles	10K	2 months	13M	GPS (153.86 m^2)	Continuous (Per 30 sec)
CDR	1.2M	1 month	14M	Tower (0.57 km^2)	Discrete
CONN	0.78M	1 month	152M	Tower (0.57 km^2)	Continuous (Per 2 min)
ETC	156K	1 month	818K	Station (25.3 km^2)	Discrete

2.2 Re-identification Attack and Privacy Risk

2.2.1 Definition of Re-identification Attack. We use one **re-identification attack** model defined in [9] to quantify the **privacy risk** of users from a mobility dataset in this paper. This attack model envisions an adversary having access to a public dataset where many users' traces were anonymized. And he/she knows for sure a user U's partial information via a separate data source or real-world observation, e.g., K leaked spatial-temporal records. For example, if the adversary knows when U leaves her home and when U arrives at her work, the adversary uses these $K = 2$ spatial-temporal records to find which trace belongs to U in the public dataset.

Formally, we define one attack as a matching function $M = (K, S, B)$. B is an anonymized set of traces (i.e., a spatial-temporal record sequence) and each trace belongs to a user. In the matching function, we uniformly sample K spatial-temporal records to obtain an arbitrary sequence S_K (which is the leaked K spatial-temporal records) from a trace S belonging to a particular user ($S \in B$). Based on S_K, we search a trace subset $b(S_K)$ from B such that $S_K \in b(S_K)$ and find out how many traces are in $b(S_K)$. A trace S is characterized as **re-identified** if $|b(S_K)| = 1$. The matching function returns 1 if $|b(S_K)| = 1$ and returns 0 otherwise.

2.2.2 Definition of Privacy Risk. The privacy risk of a user is regarded as the probability of the user being re-identified from a dataset by an adversary with the re-identification attack model. In this paper, given the above

re-identification attack model, we define the privacy risk of a user U as

$$Pr(U) = \sum_{i=1}^{N} \frac{M(K,S,B)}{N}, \tag{1}$$

where $M(K,S,B)$ is the matching function defined in Sec.2.2.1, i is a counter, and N stands for the total times that we simulate the re-identification attack (we set it to 100 in this paper).

Because the matching function has a very high time complexity, existing privacy risk prediction models learn a function f to predict the privacy risk, instead of using the matching function M, i.e., $Pr(U) = f(K,S,B)$. Compared to them, our goal is to learn the f function through transfer learning, which is defined in Section 3.

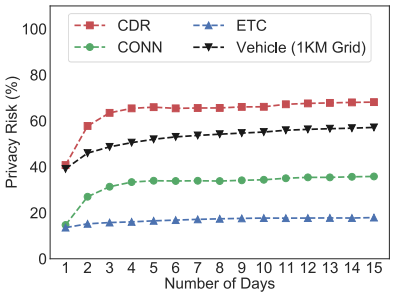


Fig. 1. Privacy Risk Evolution with $K = 4$ (The x-axis denotes the number of days of data is used for calculating the privacy risk, and the y-axis denotes the average user privacy risk of the users involved in the corresponding data.)

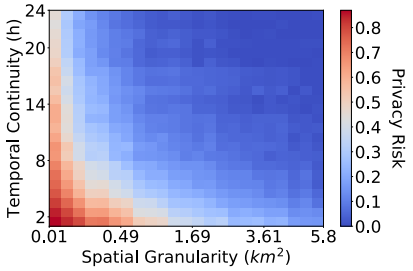


Fig. 2. Privacy Risk Contour for Vehicles with $K = 4$ (The spatial granularity denotes the area of a cell in a grid with a specific side, e.g., the area of a cell in $1km \times 1km$ grid is $1(km)^2$.))

2.3 Why Privacy Risk Prediction with Short-term Data

To study the impact of date length of mobility data on the privacy risk of users, we measure the average privacy risk of users from three weeks of data of four mobility datasets, including CDR, Connection, ETC, and Vehicles. In particular, according to our measurement and the previous study[5] on the uniqueness at GPS level, for the vehicle data with original spatial granularity at GPS level, it is able to identify it from the crowd with 90% probability even we only know one of its accurate records. However, this kind of attack is not practical because it is almost impossible to obtain the exact GPS data of a vehicle externally. We try different sizes of spatial granularity and temporal continuity to discretized the vehicle data and analyze the privacy risk as shwon in Figure 2. The choice of discretization has a significant impact on the privacy risk. To avoid the very high privacy risk and very low privacy risk that may cause imbalance issues to the prediction, we prefer to use an intermediate spatial granularity. So, we measure the vehicular system privacy in the setting of $1km$ square grids (i.e., dividing a city into a grid where each cell is $1km \times 1km$).

From Figure 1 we find the privacy risk of users increases with the increase of the length of days for each mobility dataset. For example, the user privacy risk of three weeks (15 weekdays) cellphone Connection data is around 38% while that of one-day cellphone Connection data is only 18%, which increase more than 110%. On the other hand, the increases for all datasets start to be flat after one week. From the result, we argue that it is necessary to predict the privacy risk with short-term data to reduce the privacy risk of users when a new urban service starts to collect data. Besides, the result shows the privacy risk of 15 weekdays data could be used as the label for users of mobility dataset because it is more stable than the privacy risk of data within one week.

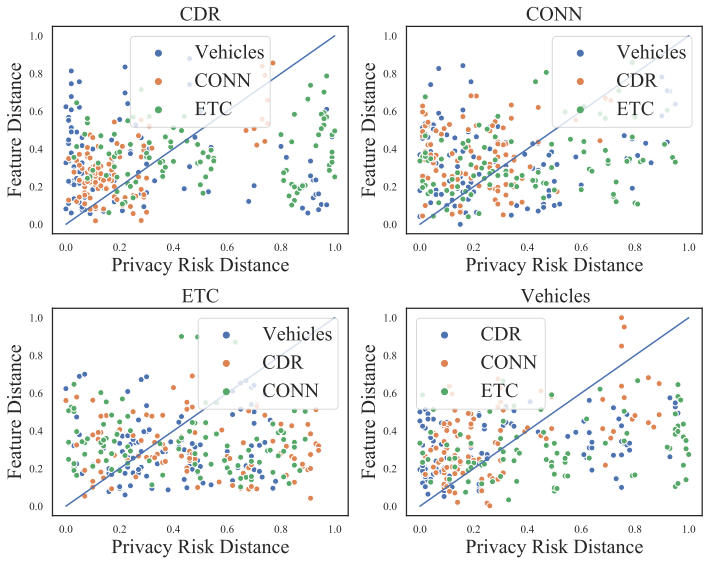


Fig. 3. Distances Comparison

2.4 Why Aggregate Mobility Features are Not Enough for Transfer Learning

We obtain the scatter plot where each dot in the figure is representing one user from the source dataset (e.g., CDR) and one user from the target dataset (e.g., Vehicle). The X-axis is the absolute distance between these two users' privacy risks, and the Y-axis is the l2 pairwise distance between their aggregated mobility features. The aggregated mobility features of users from each dataset are defined in Section 3. We show the result for each dataset in Figure 3. We find most dots do not lie close to the diagonal line in the figure, which means the short distance between two users' mobility features does not equal the short distance between their privacy risks. Hence, only using aggregated mobility features is not enough for transfer learning privacy risk prediction. One possible reason is the aggregated mobility features may lose some important information because of their different spatial granularity and temporal continuity. For example, a trip in the Toll Collection (ETC) dataset might only contain two records for one vehicle in general, while the trip in Vehicles GPS dataset could collect a dense trace of one vehicle. These two captured trajectories of the same trip have the same travel distance but their privacy risk levels are different because every day there should be multiple vehicles traveling between two stations on a highway (meaning low risk) but it is hard to find two vehicles that have exact the same dense spatial-temporal trajectory (meaning high risk). To solve this issue, in Section 3, besides the aggregated mobility features, we employ the spatial-temporal tensors as the additional input for source and target datasets.

3 DESIGN

In this section, we first show the overview of TransRisk. Then we introduce the detail of two inputs for embedding, i.e., the aggregated mobility feature vector and spatial-temporal tensor. Finally, we present layers for embedding and transfer learning for TransRisk and the involved loss functions.

3.1 Overview of TransRisk

Figure 4 shows the architecture of TransRisk, which includes two phases, i.e., the feature embedding phase and the knowledge transferring phase. The feature embedding phase contains two sequences of layers to embed source input (one trajectory of a user from source dataset) and target input (one trajectory of a user from target dataset) to obtain the embedding features, i.e., (1) TransRisk calculates the mobility features for a user (from

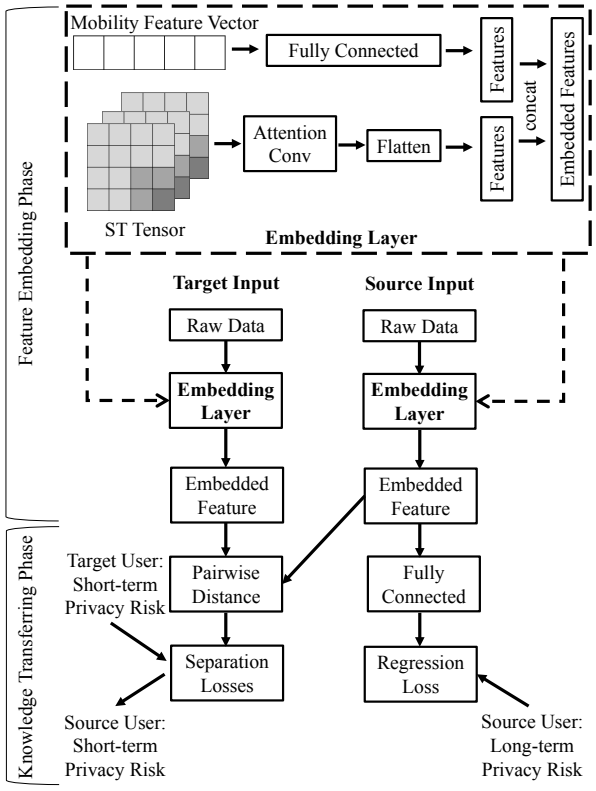


Fig. 4. Framework of TransRisk

source or target input) given her/his trajectory data and then feed the features into a neural network to obtain one feature vector; (2) TransRisk calculates the Spatial-Temporal Tensor for a user (from source or target input) given her/his trajectory data and then feed the tensor to an attention network to obtain another feature vector. (3) concatenate two output feature vectors into one feature vector to obtain an embedding feature vector. With the embedding features from source and target, in the knowledge transferring phase: (1) TransRisk utilizes a pairwise distance calculation function to compute the distance between the embedded features of source and target inputs and then calculate the loss between this distance with the label distance, i.e., the separation loss; (2) TransRisk employs a fully connected layer to obtain a predicted label for the source input and then calculate the corresponding loss as well, i.e., the regression loss; (3) TransRisk learns the weights for the privacy risk prediction of source data through the training of regression loss and then transfers the learned weights to target data through the training of separation loss.

3.2 Spatial-Temporal Tensor

To unify spatial-temporal trajectories of each mobility dataset into one standard, we employ the spatial-temporal tensor (ST tensor) T to represent the regular spatial-temporal mobility patterns of a user from a mobility dataset. We map the spatial information of a city to a geospatial grid G with $R \times C$ grid cells $c_{i,j}$, where $G = c_{1,1}, c_{1,2}, \ldots, c_{R,C}$ (R is 120 and C is 60). Then each visitable location in a mobility dataset could be associated with a grid cell based

on its coordinate {longitude and latitude}. We divide one day into N time slots and let T be a tensor with N matrices, where N is 12 and each time slot is 2 hours. Each matrix in a one-time slot has R rows and C columns, which are equal to the numbers of rows and columns for the grid of the city after the mapping. This matrix is defined as the probability matrix indicating the probability of a user visits a place at a time slot.

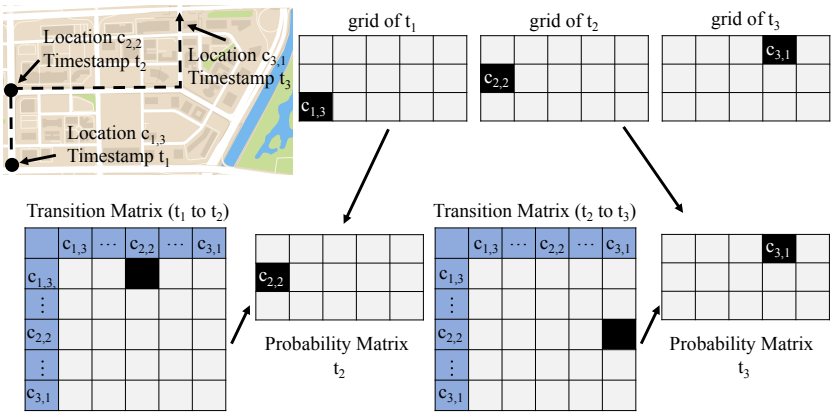


Fig. 5. Example of ST tensor

Figure 5 shows an example of how we obtain the probability matrices for a user. The left top map in figure 5 shows the whole trace of one user from a mobility dataset that contains only three records. After mapping, we obtain three grids for this trace where the color in a grid cell indicates the frequency of the user visits to that cell, and two transition matrices, i.e., transition matrices from t_1 to t_2 and t_2 to t_3. The color is deeper if the frequency is higher, as well as that for the transition matrix. Given the frequency grid of t_1 and the transition matrix for t_1 to t_2, for cell $c_{2,2}$ in grid t_2, we sum the products of each frequency in grid t_1 multiply its transition probability to $c_{2,2}$. We do this calculation for each cell of the grid to obtain the probability matrix for t_2. Similarly, we obtain the probability matrix for t_3. Note that there are N transition matrices for one day and the last transition matrix indicates the transition probabilities between cells from timestamp t_N to t_1.

Formally, an entry T_{ijk} in T is defined as the probability of the user visits $c_{i,j}$ at time slot k. By assuming a user's transition between locations follows the Markov property [33], we have

$$T_{ijk} \approx \sum_{r=1,c=1}^{R,C} F_{r,c,k-1} Tr_{c_{r,c},c_{i,j},k-1} \tag{2}$$

The first term in the above equation indicates the normalized visited frequency of a grid cell $c_{r,c}$ for the user at time slot $k-1$. The second term is a transition function that return the transition probability of the user transit from grid cell $c_{r,c}$ to grid cell $c_{i,j}$. We utilize the probability matrix instead of the frequency matrix to reduce the impact from the noise of data collection, such as the ping-pong effect (happens when the strength of signals of cell tower change significantly and will lead to abnormal records) for CDR data because the normal records will be dominated in the computation of probability. In addition, only the frequency matrix may not include the temporal transition of the trajectory data. Here, for data with a long time interval between two records, e.g., CDR, we will assume the user stays in the same place until a new record emerges.

3.3 Aggregated Mobility Feature Vector

In TransRisk, we also utilize aggregated mobility features to represent a user in a mobility dataset, which is extracted from his/her spatial-temporal trajectory. The mobility feature of a user is denoted as a vector X. In this paper, we utilize the classic personal aggregated mobility features defined from the previous works[25][2].

- **Random Entropy (RE):** We utilize the method from [12] to calculate random entropy of a user. It calculates the predictability of user visited locations based on the assumption that the user will visit each location with equal probability.
- **Uncorrelated Entropy (UE)**: We utilize the method from [23] to calculate the uncorrelated entropy of a user. This method calculates the predictability of visited locations of the user based on the historical visiting probability of each location without considering the temporal correlation.
- **Real Entropy (Enp):** We utilize the method from [31] to calculate the real entropy of a user. This method calculates the predictability of visited locations of the user based on the historical visiting probability of a time-ordered sub-sequence of the trajectory of the user. This method considers not only the historical visiting probability of locations but also the temporal order of locations.
- **Number of Visited Locations (NV)**: It computes the daily number of locations a user visited.
- **Maximal Distance from Home (MDH)**: It computes the maximal geographic distance between a user's home and a location the user visited. Here the home of a user is defined as the most frequently visited place of the user during the early morning and late night.
- **Number of Distinct Visited Locations (NDV)**: It computes the daily number of distinct locations a user visited.
- **Distance of Straight Line (DSL)**: It computes the total geographic travel distance of a user during a trip. We choose the maximal one from trips during the measurement period.
- **Maximal Distance (MD)**: It computes the maximal geographic travel distance between two consecutive spatial-temporal records of a user.
- **Radius of Gyration (RG)**: The radius of gyration is the characteristic distance a user traveled during a trip, i.e., the distance between the origin and the geographic mass center of the spatial-temporal records during the trip. We choose the maximal one from trips during the measurement period.
- **Waiting Times**: It computes the time difference between any two temporal consecutive spatial-temporal records in the trajectory of a user. In particular, we use the mean (MWT) and standard deviation (SWT) of the waiting times as two mobility features of a user.
- **Jump Lengths**: It computes the spatial distance between any two temporal consecutive spatial-temporal records in the trajectory of a user. In particular, we use the mean (MJL) and standard deviation (SJL) of the jump lengths as two mobility features of a user.

3.4 Deep Transfer Learning

TransRisk has two phases, i.e., the feature embedding phase and the knowledge transferring phase.

3.4.1 Feature Embedding. In the feature embedding phase, TransRisk is given 1) a massive labeled training data $D_s = (X_i^s, T_i^s, y_i^s, L_i^s)$ from the source dataset, where X_i^s is the mobility feature vector of user i; T_i^s is his/her ST tensor; y_i^s is the corresponding short-term label (privacy risk) in source dataset; L_i^s is the corresponding long-term label (privacy risk) in source dataset; 2) a few labeled training data $D_t = x_i^t, y_i^t$ from the target dataset, where X_i^t, T_i^t, and y_i^t are the mobility feature vector, ST tensor, and short-term label of user i in target dataset. TransRisk utilizes the same deep learning layers to embed X_i and T_i to obtain an embedding feature vector as shown in figure 4. Formally, we design a embedding function g to obtain the embedding feature vector, i.e.,

$$g = concat(g_f(X_i), g_c(T_i)) \tag{3}$$

The the concat function in above equation is a function to concatenate the outputs of $g_f(X_i)$ and $g_c(T_i)$, where $g_f(X_i)$ is a fully connected layer (input dimension is 14 and output dimension is 8) followed by a relu layer; $g_c(X_i)$ is two attention-based convolutional layers[3] and a flatten layer where each attention-based convolutional layer followed by a pooling layer. The input channel, output channel, and kernel side of first attention-based convolutional layer are 24, 8, and 3. The input channel, output channel, and kernel side of first attention-based convolutional layer are 8, 4, and 3. We use an attention-based convolutional layer instead of a traditional convolutional layer to capture the global information of the ST-tensor.

3.4.2 Transfer Learning. The goal of the transfer learning phase is to learn a function f that predicts the label of unlabeled users from the target dataset based on the assumption that there exists a covariate shift[21] between embedding features from source dataset D_s and target dataset D_t. To achieve this, we define the function f as

$$f = f_s(h(g^s)) + f_{sa}(d(g^s, g^t)), \tag{4}$$

where h is a fully connected layer to predict a privacy risk given g^s and f_s is a $L2$ loss function, where its label is L_i^s. $f_{sa}(d(g^s, g^t))$ is a separation loss function[21]. $f_{sa}(d(g^s, g^t))$ calculates the loss for pairs of users (g^s and g^t) from source dataset and target dataset. Here d is a distance measure function defined as

$$d(g^s, g^t) = \frac{1}{2}||g^s - g^t||^2, \tag{5}$$

where $||\cdot||$ is the Frobenius norm. The label for f_{sa} is the absolute value of the difference between the short-term labels of users from source data and target data.

Data Flow: (1) Given the trajectory of a user from source dataset D_s and the trajectory of the target user from D_t, TransRisk first converts each trajectory data into a mobility feature vector X and a spatial-temporal tensor T. **(2)** In a sequence of embedding layers, the aggregated mobility features are fed into a fully connected layer, and the spatial-temporal tensor is fed into the attention-based convolutional layers followed by a flatten layer. **(3)** The outputs of the above two layers are concatenated into one feature vector. **(4)** In transfer learning, there are two data streams, i.e., the embedding feature g^s from the source input and the embedding feature g^t from the target input. **(5)** The model feeds g^s into h, a fully connected layer, for the training of the regression loss f_s. **(6)** The model feeds the corresponding g^s and g^t into f_{sa} to calculate the pairwise distance for the training of the separation loss. **(7)** The goal of the transfer learning module is to minimize the total loss, i.e., $f_s + f_{sa}$.

3.5 TransRisk+: an Ensemble Predictor

Given a target dataset and a source dataset, they might have a significant difference (might be kilometers level) either on spatial granularity or temporal continuity or both. This might cause low transfer accuracy. To explore if we can use multiple source datasets to improve the accuracy of prediction, we also design an ensemble learning version of TransRisk to predict the privacy risk based on the output of multiple TransRisk models, where each one is trained by the target dataset with a source dataset. The model is called TransRisk+. Therefore, TransRisk+ utilizes multiple corresponding base TransRisk modules for these multiple source datasets and obtain multiple possible privacy risks for users from a target dataset. Then TransRisk+ feeds the possible labels into an ensemble function to obtain the final predicted label. TransRisk+ utilizes a classic gradient-boosted-decision-tree (GBDT) regression function[13] as the ensemble learning module to predict the privacy risk for the target user, where the loss function is the Friedman MSE [13].

4 EVALUATION

In this section, we first introduce the data management, ground truth, evaluation setting, including the metrics, and the baselines. We then show the result of the comparison of TransRisk and baselines. Then, we evaluate

TransRisk+ with a baseline by using all source datasets as input. Finally, we show a heatmap about the privacy risk for the regions in Shenzhen.

4.1 Data Management and Processing

To manage and process five massive datasets, we employ a high-performance cluster with two open-source data processing frameworks, i.e., Hadoop and Spark. In particular, the cluster includes: (i) 12 Hewlett-Packard machines with 2 Tesla K80c each; (ii) 10 Dell machines with 4 Tesla K80c each; (iii) 4 Xeon E5-2650 with a half TB memory each; (iv) A series of 800GB SSD and 15TB of spinning-disk spaces; (v) 2 PB additional disk space.

4.2 Ground Truth

We use five datasets used for the evaluation of TransRisk, including four datasets already defined in Section 2.1, i.e., CDR, Connection, ETC, Vehicle, and a target dataset, i.e., Camera. In particular, we apply the attack model to the data and obtain the privacy risk. With the measured privacy risk and the extracted mobility features, we obtain the training data and the testing data. In our evaluation, all involved data only includes records on weekdays.

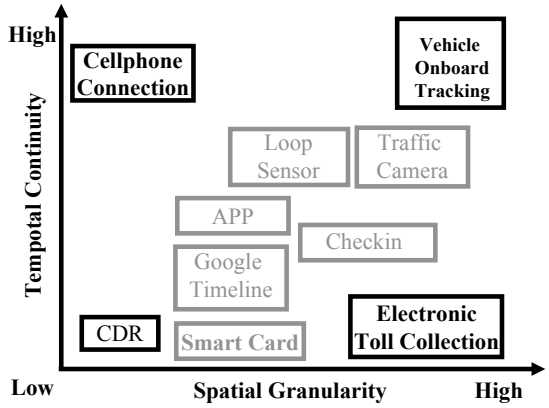


Fig. 6. Study Spectrum of Mobility Datasets

4.2.1 Data Granularity Spectrum. A key question to our approach is which dataset is appropriate to be used as the source dataset. Given various existing mobility datasets, it is very challenging, if not impossible, to try all the popular datasets as the source datasets. To address this issue, we quantify mobility datasets by two themes, i.e., spatial granularity and temporal continuity, as shown in Figure 6. For example, the vehicular GPS data for usage-based insurance [20] in the top right corner has the highest spatial granularity and the highest temporal continuity (e.g., GPS data with 10-second uploading intervals). In contrast, the electronic toll collection system in the bottom left corner has the lowest spatial granularity and temporal continuity. We argue these two themes are fundamental because they quantify the interactions of urban service systems with their users by (i) how fine-grained the locations we have (e.g., at the level of GPS or a cellular tower) and (ii) how fast we update the interactions (e.g., periodically every 10 seconds, or on-demand event-based updating such as a phone call). With these two themes, TransRisk utilizes four mobility data sets in the four corner cases (the bold boxes in Figure 6) to explore which source dataset could achieve the best prediction accuracy given on short-term data, which might cover the spatial granularity and temporal continuity of most urban mobility datasets.

4.2.2 Target Dataset. In our setting, we need to choose the camera dataset as the new mobility dataset. Hence, we utilize two-month traffic camera data from Shenzhen. The traffic cameras in Shenzhen concentrate on the downtown of Shenzhen. The road network in Shenzhen contains 73 thousand intersections and 101 thousand road segments, of which 248 intersections are equipped with traffic cameras. We use two-month vehicle data generated from the camera surveillance records at these 248 intersections as our target dataset.

4.3 Evaluation Setting

4.3.1 Metrics. We use the root-mean-square error **RMSE** and mean-absolute error **MAE** as two criteria to measure the accuracy of TransRisk and baselines with a range from 0 to 1 inclusively, where 0 represents the best, and 1 represents the worst.

4.3.2 Baselines. We compare TransRisk (TR in the rest figures) with four baselines as follows.

- **Random Forest (RF)**: [25] extracts mobility features from the spatial-temporal trajectory of users to represent the profile of users. Given the mobility feature vector of a user, [25] utilizes a Random Forest prediction model to predict the privacy risk of the user. In our evaluation, we use 500 estimators and use log_2 of the number of attributes as the maximal number of attributes for each estimator.
- **K Nearest Neighbor (KNN)**: K Nearest Neighbor[1] provides a non-parameter learning method for class prediction. Given the mobility feature vector of a user, KNN searches neighbors with the most similar mobility features and predicts the label of the user, the number of neighbors is set to 5.
- **Decision Tree Learning (DT)**: Decision Tree Learning [30] is a classic prediction model that has been used in many fields. Given the mobility feature vector of a user, the decision tree learning uses a decision tree to check each mobility features and concludes the privacy risk of the user. In this evaluation, we also set log_2 of the number of attributes as the maximal number of attributes for the decision tree.
- **TrAdaboost (TA))**: TrAdaboost is a classic multiple source transfer learning regression model [24], which is based on Adaboost algorithm. Compared to the above baselines, the input of this method is the mobility feature vectors of the users from source datasets and the mobility feature vector of the user from the target dataset. We use 20 estimators in TrAdaboost and set the number of iteration steps to 10.
- **TransRisk- (TR-)**: TransRisk- is a variant version of TransRisk wherein the feature embedding we only utilize the ST tensor as input without aggregated mobility features. This version of TransRisk is a variant of the classic deep transfer learning model for image data.
- **TransRisk+ (TR+)**: To explore if multiple source datasets could improve the accuracy, we also evaluate TransRisk+ with multiple combinations of source datasets.

In the evaluation, we regard 10% of target data as training data and 90% as test data since there might be only a few labeled target data for new urban services. On the other hand, for one source dataset, we use all records for training. Given the cluster, the training time of TransRisk is less than 24 hours, and the prediction time of one record is less than 1 second. Compared with the high time complexity of privacy risk measurement, prediction with models could reduce the latency significantly.

4.4 Evaluation

We first compare TransRisk with different baselines on the Camera dataset and study the impact of different factors, i.e., data volume, and source dataset. Then we show the effectiveness of TransRisk on different mobility datasets and the performance of TransRisk+.

4.4.1 Comparison on Camera Dataset. Figure 7 shows the RMSE and MAE of TransRisk and baselines under different numbers of K (defined in section 2.2) where TransRisk and TransRisk- are trained by using vehicles data as source data. We find TransRisk and TransRisk- have very similar performances with different values of K.

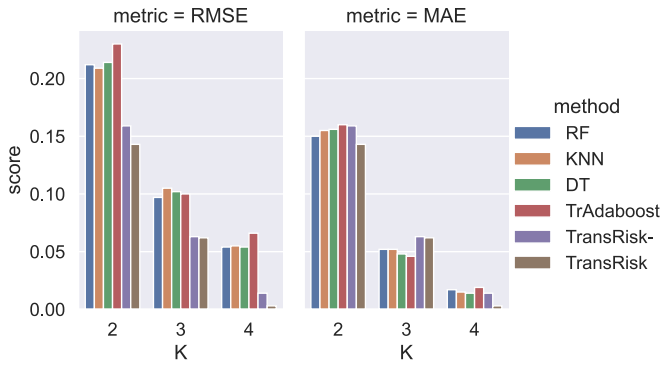
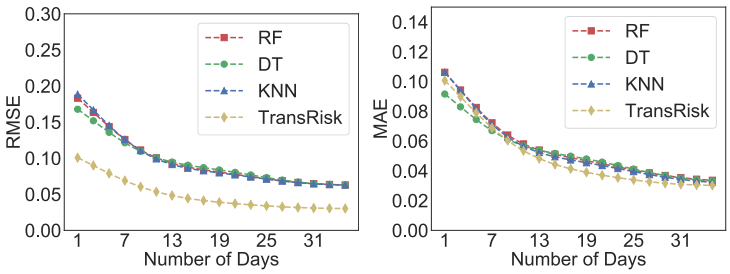


Fig. 7. Performance of One Week Data

TransRisk has the best performance on $K = 2$ and $K = 3$ while TransRisk- has a better performance on $K = 4$ compared to TransRisk. The baselines have slightly worse performance. In particular, TrAdaboost has the lowest accuracy. One possible reason is that TrAdaboost regards the mobility feature vectors of all source datasets in the same way, which may deteriorate its performance. On the other hand, we also utilize 35 days of camera data with $K = 3$ to train three baselines (including RF, KNN, and DT), where 80% of data are used for training. The performances of these three baselines are much better than TransRisk, where their RMSE is around 0.05 and MAE is around 0.03. However, TransRisk only requires the short-term and small amount of data for training, which can greatly reduce the budget of evaluating the privacy level of a specific data collection policy.



(a) RMSE of 7 Weeks (b) MAE of 7 Weeks

Fig. 8. RMSE and MASE of All Methods on 7 Weeks Data (K=3)

4.4.2 Impact of Training Data Volume. With the massive daily uploaded data, there will be more training data fed to the predictive privacy measurement models. Therefore, the performances of TransRisk and baselines may be improved with increasing training data. To study the impact of training data with different lengths of days, we generate training data from 1 day to 35 days (only including the weekday data, i.e., around seven weeks). Figure 8a and Figure 8b show the evolution of the RMSE and MAE for TransRisk and three baselines when $K = 3$. We find that, with the increasing data, the performances of all methods are improving. In particular, TransRisk has a significantly better performance in terms of RMSE compared to other baselines. While in terms of MAE TransRisk is only slightly better than others. One possible reason is the distribution of privacy risk of camera data focused on a narrower range of values.

Figure 9 shows the performances of TransRisk, TransRisk-, and TrAdaboost on camera data with different source datasets. We find all source datasets have similar performance for TransRisk. In particular, TransRisk- has similar performance compared to TransRisk except the case using vehicles data as source data. One possible reason might be that when Vehicles data and Camera data have similar mobility features, their privacy risk is

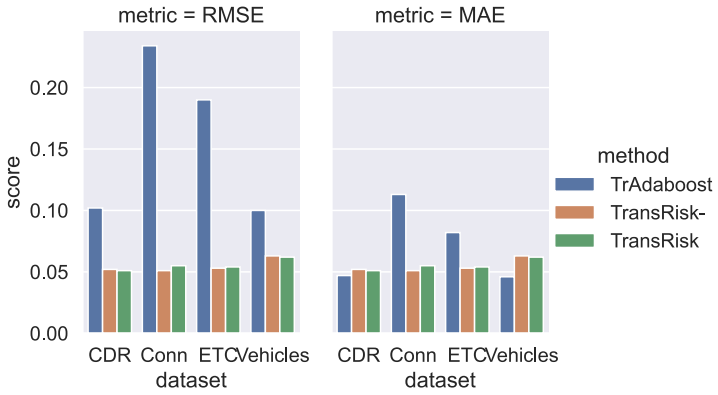


Fig. 9. Performance of Different Data Sources

quite different. The result in Figure 9 shows TransRisk is robust for the Camera dataset with different source datasets with $k = 3$ and one-week training data.

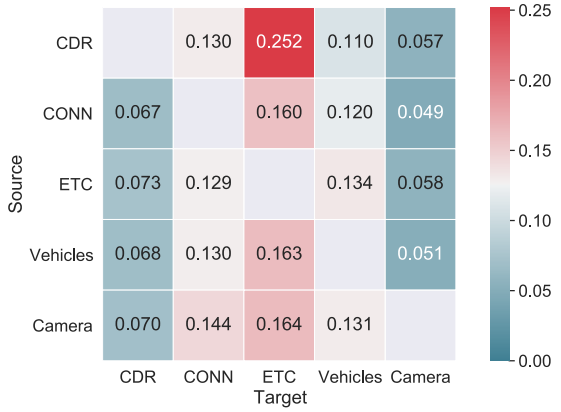


Fig. 10. RMSE Grid of TransRisk with K=2

4.4.3 Effectiveness on Different Datasets. To study the effectiveness of TransRisk on different mobility datasets, we evaluate TransRisk with different source and target pairs paired by the five datasets, i.e., CDR, CONN, ETC, Vehicles, and Cameras. Figure 10 shows the RMSE grids for these pairs with $K = 3$ and three-week training data. We find when using CDR as source data, ETC as target data, TransRisk has a very bad performance. While using Camera and CDR as target data, transRisk always has a good performance (under 0.08) in terms of RMSE. This might be caused by the fact that ETC data has a significantly coarser spatial granularity compared to others.

4.4.4 Distribution of Predicted Privacy Risk and Groundtruth. We show the CDF for the predicted privacy risk and the ground truth for using vehicles data as source dataset and camera data as target dataset with $K = 3$ and one-week data. We found the predicted privacy risk concentrated on the range from 0.8 to 1, which is the most portion of the ground truth value.

4.4.5 Predicted Label Distribution. In addition to privacy risk, identifying if a user is a "high risk" user or not is also important. Therefore, we also change the regressor into a classifier and divide the privacy risk into three

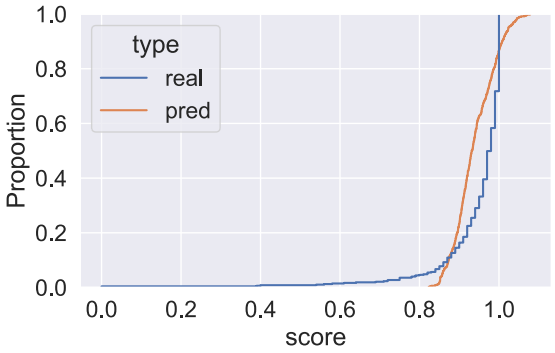


Fig. 11. CDF of Predicted Privacy Risk and Groundtruth

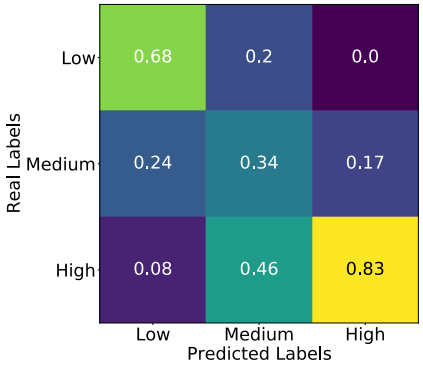


Fig. 12. Predicted Labels Grid

levels, i.e., Low risk, Medium risk, and High risk. We use the F1 score as the criteria to measure the accuracy of our model with a range from 0 to 1 inclusively, where 0 represents the worst and 1 represents the best. The metric will be used in section to show the true label distribution. Figure 12 shows a grid of true label distribution for each predicted label with $K = 3$ and one-week training data. For example, the Low column (the most left one) is for the users who are marked by TransRisk as the low-risk user, and the number 0.68 in the Low row indicates there are 68% users whose true label is low risk. From the result, we find that most high-risk users are predicted as they have a high mobility privacy risk. In particular, for the medium risk users, there are 46% of users are mistakenly predicted as high-risk users.

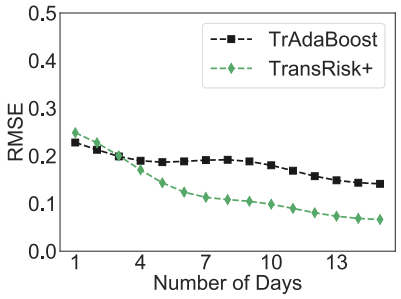


Fig. 13. RMSE of 3 Weeks

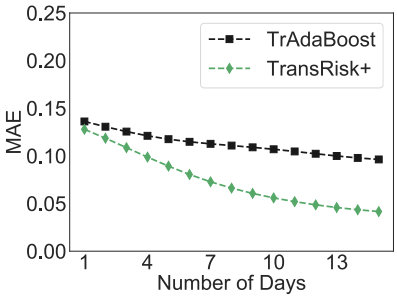


Fig. 14. MAE of 3 Weeks

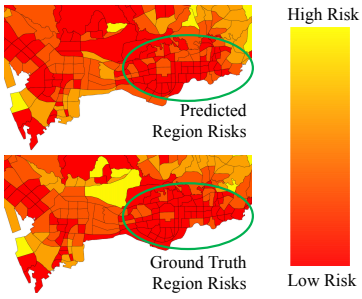


Fig. 15. Region Risk for Camera Data

4.4.6 Impact of Multiple Sources. To explore the performance of using all sources datasets for prediction, we design a variant version of TransRisk, i.e., TransRisk+. To evaluate its performance on using all source datasets (i.e., except Camera data), we compare TransRisk+ with TrAdaboost on data with $K = 3$ and different lengths of days (three weeks). Figure 13 and Figure 14 shows TransRisk+ has a better performance than that of TrAdaboost. Also, we find the performances of TransRisk+ improve with the increase of the number of days. Figure 13 and Figure 14 show TransRisk could be applied to multiple sources scenario. Compared to TransRisk, TransRisk+ has worse performance than TransRisk at the beginning with very few data. This is because the mobility pattern of CDR and ETC is quite different from other datasets. However, with more days of data for training being used, the performance of TransRisk+ is very close to the performance of TransRisk. TransRisk+ could be used for the case that the mobility pattern of all the source datasets are quite different from the target dataset.

4.5 Visualization of Privacy Risk

To provide the visualization of privacy risk of camera data, we predict the privacy risk of users in the camera dataset, and assign the users risk to 491 regions in Shenzhen (the census blocks), e.g., the risk of a region is the average risk of users who have visited the region. We visualize the region risk in Figure 15 where the red color indicates lower risk; the yellow color indicates a higher risk. We find that most regions have a correct prediction at the regional level. Also, as shown in the heatmap of ground truth region risk from Figure 15, we find the record density and population density impact the privacy risk, e.g., the regions in the CBD area (circled area) have a much lower privacy risk, which means the spatial-temporal records collected at these regions have less risk to be exposed to adversaries.

5 RELATED WORK

5.1 Matching-based Privacy Risk Assessment

The privacy risk assessment for mobility datasets has been widely studied. Most of them propose a re-identification attack model and utilize a corresponding matching function to measure the privacy risk of users from the mobility dataset. [9] found that it is able to identify a specified resident from 1.5 million people by only four spatial-temporal records with 95% accuracy. [10] confirmed this with the study of a 3-month credit card transactions dataset of 1.1 million people. [35] showed the capability to identify a resident from aggregated mobility data by correlating the mobile application data and cellular data. Similarly, [17] demonstrates how easy to re-identify people in a nation wise dataset. [32] measured the probability of physical location privacy leak of people based on the online service access data via static and mobile devices. [7] measured the capability to identify people when they reveal the nearby POI by analyzing the uniqueness of POI in five different large cities in the world. [19] offers a pretty work for the re-identification of the identity based on the cab traces. These works focus on one or multiple mobility datasets, and their result reveals the vulnerability of anonymized mobility datasets. However, due to the high time complexity of their matching algorithm, these assessments might not be applicable to large-scale mobility datasets given the massive daily uploaded data.

5.2 Prediction-based Privacy Risk Assessment

Recently, some prediction-based privacy assessment methods have been presented to solve efficiency issues for these daily updated large-scale mobility datasets. [33] showed the feasibility to predict locations of people by learning the regularity and conformity of heterogeneous mobility datasets, e.g., vehicles and check-in data. [4] demonstrates that the socioeconomic status of people is predictable, which is based on the features extracted from the mobile phone data and reconstruct the distribution of wealth of a nation. [25] presents a decision tree-based random forest predictive model to efficiently measure the privacy risk by the extracted mobility features of the users from the vehicular tracking system. [29] utilizes a generative model to measure the probability of identifying users from the crowd given incomplete datasets. [22] trains classifiers to capture the relation between individual mobility patterns and the privacy risk of individuals. With training with massive measured mobility data, these models could efficiently predict the privacy risk of users when the mobility dataset is updated. However, for the newly deployed urban services, collecting massive data contradicts the purpose of privacy risk prediction. The operators may want to set up the configuration for data collection within a short time.

5.3 Transfer Learning in Privacy Study

Many works in the privacy and security community have been proposed based on transfer learning. [15] designs a transfer learning model to identify persons from images, which utilizes the convolution neural network for feature extraction. [26] proposes an asymmetric multitask learning approach for image data to transfer a view-invariant representation from the source dataset to an unlabelled target dataset, which people in the target dataset do not need to be in the source dataset. [18] also presents a deep domain adaptation model to perform the cross-domain

person re-identification, which utilizes the domain-invariant and specific features of images. These works all focus on the person's re-identification from image data. Different from the above works, our work focus on the re-identification study for mobility data. Our method predicts the user mobility privacy risk based on the knowledge learned from a source mobility dataset, given the ST tensor and mobility feature vector. Compared to the existing prediction-based privacy assessment, our work reduces the requirement of the amount of labeled data for measuring the user privacy risk of a new urban service. To our best knowledge, TransRisk is the first work to focus on the efficient mobility privacy risk prediction based on transfer learning and explore the impact of different source mobility datasets with diverse spatial granularity and temporal continuity.

6 DISCUSSIONS

Limitations. In this paper, we only use the data from one Chinese city Shenzhen to study the privacy risk, which may not be generalized to other counties due to different mobility patterns and cultural diversity. As for other cities in China, we believe our results can be generalized to other major cities with similar large-scale urban infrastructures, e.g., Beijing and Shanghai. For smaller cities, it is hard to predict given their limited infrastructure to capture mobility and lifestyle differences, e.g., a shorter commuting distance leads to different features. Further, we only select four representative systems in Figure 6 as examples to study combinations of spatial granularity and temporal continuity for privacy study. Thus, our results may not be generalized to other systems with different combinations. Lastly, the proposed method helps estimate the privacy risk, while it does not have a privacy guarantee, e.g., security risks brought via transfer learning, which is out of the scope of this work.

Potential Implications. Through the evaluation of TransRisk on the camera dataset, we show the effectiveness of our work on assisting the traffic camera surveillance system of a city to adjust the data granularity in a short time. In general, our work has some potential implications for different aspects of the application: **1** For the government or institute who plan to deploy an urban service under the specific privacy policy, e.g., General Data Protection Regulation [14], TransRisk offers an efficient privacy risk evaluation method for various urban data collection applications with very few and short-term data to protect privacy before setting up the service on large scale. Therefore, TransRisk is able to greatly reduce the budget of evaluating the privacy level of a specific data collection policy. **(2)** For people who concern about the privacy risk or who want to balance their benefits of involvement in an urban service, TransRisk has the potential to assist the decision-making about the participation or the involvement amount with rigorous theoretical analysis and empirical validation[8][6][16].

Privacy and Ethics. All the data sets are collected under the consent of the urban service users. Our partners have been removing the user ID and their detailed personal data, so our results did not focus on individual users.

7 CONCLUSIONS

In this work, We conduct the first comprehensive study on transfer-learning-based mobility privacy risk prediction. We design a novel model to predict the privacy risk of a user based on transferring knowledge from other mobility data sources. We evaluate our work on real-world mobility data from multiple mobility datasets. Our work has potential implications for future endeavors in privacy risk prediction, and mobility data protection in the research community and industries. Going forward, the location data collected from urban services will play an increasingly important role in smart city development. Therefore, identifying the privacy risk of different data collecting methodologies in a fast way is crucial for future policy making and privacy preserving technology development.

ACKNOWLEDGMENTS

This work is partially supported by NSF 1849238, 1932223, 1951890, 1952096, 2003874, 2047822, 1920462, 1943100, 2002985 and the Facebook Faculty Fellowship. We thank all the reviewers for their insightful feedback to improve this paper.

REFERENCES

[1] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.
[2] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.
[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.
[4] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.
[5] Antoine Boutet, Sonia Ben Mokhtar, and Vincent Primault. Uniqueness assessment of human mobility on multi-sensor datasets. 2016.
[6] Dan Calacci, Alex Berke, Kent Larson, et al. The tradeoff between the utility and risk of location data and implications for public good. *arXiv preprint arXiv:1905.09350*, 2019.
[7] Hancheng Cao, Jie Feng, Yong Li, and Vassilis Kostakos. Uniqueness in the city: Urban morphology and location privacy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):62, 2018.
[8] Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Robert Birke, Sara Bouchenak, Lydia Y. Chen, Nicolas Marchand, and Bogdan Robu. Pulp: Achieving privacy and utility trade-off in user mobility data. In *2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS)*, pages 164–173, 2017.
[9] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.
[10] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.
[11] Vision System Design. Deep learning system powers traffic enforcement system, 2021.
[12] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
[13] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
[14] GDPR. General data protection regulation https://gdpr-info.eu, 2018.
[15] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.
[16] Mehmet Emre Gursoy, Ling Liu, Stacey Truex, Lei Yu, and Wenqi Wei. Utility-aware synthesis of differentially private and attack-resilient location traces. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 196–211, 2018.
[17] Xiaojun Hei, Chao Liang, Jian Liang, Yong Liu, and Keith W Ross. A measurement study of a large-scale p2p iptv system. *IEEE transactions on multimedia*, 9(8):1672–1687, 2007.
[18] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–178, 2018.
[19] Chris YT Ma, David KY Yau, Nung Kwan Yip, and Nageswara SV Rao. Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM transactions on networking (TON)*, 21(3):720–733, 2013.
[20] Robert John McMillan, Alexander Dean Craig, and John Patrick Heinen. Motor vehicle monitoring system for determining a cost of insurance, August 18 1998. US Patent 5,797,134.
[21] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017.
[22] Luca Pappalardo, Gianni Barlacchi, Roberto Pellungrini, and Filippo Simini. Human mobility from theory to practice:data, models and applications. In *Companion Proceedings of The 2019 World Wide Web Conference on*, pages 1311–1312, 2019.
[23] Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1-2):75–92, 2016.
[24] David Pardoe and Peter Stone. Boosting for regression transfer. In *ICML*, 2010.
[25] Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, and Anna Monreale. A data mining approach to assess privacy risk in human mobility data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(3):1–27, 2017.

[26] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1306–1315, 2016.

[27] Zhou Qin, Fang Cao, Yu Yang, Shuai Wang, Yunhuai Liu, Chang Tan, and Desheng Zhang. Cellpred: A behavior-aware scheme for cellular data usage prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–24, 2020.

[28] Zhou Qin, Zhihan Fang, Yunhuai Liu, Chang Tan, and Desheng Zhang. A measurement framework for explicit and implicit urban traffic sensing. *ACM Transactions on Sensor Networks (TOSN)*, 17(4):1–27, 2021.

[29] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.

[30] Lior Rokach and Oded Z Maimon. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2008.

[31] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[32] Huandong Wang, Chen Gao, Yong Li, Zhi-Li Zhang, and Depeng Jin. From fingerprint to footprint: Revealing physical world privacy leakage by cyberspace cookie logs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1209–1218. ACM, 2017.

[33] Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. Regularity and conformity: Location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284. ACM, 2015.

[34] Xiaoyang Xie, Yu Yang, Zhihan Fang, Guang Wang, Fan Zhang, Fan Zhang, Yunhuai Liu, and Desheng Zhang. cosense: Collaborative urban-scale vehicle sensing based on heterogeneous fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–25, 2018.

[35] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1241–1250. International World Wide Web Conferences Steering Committee, 2017.

[36] Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. Vemo: Enabling transparent vehicular mobility modeling at individual levels with full penetration. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.