IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 21, 2020, accepted May 8, 2020, date of publication May 18, 2020, date of current version June 4, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2995152*

# Mobile Privacy: Scalable Ensemble Matching for User Identification Attacks

**LUOYANG FANG[1], HAONAN WANG[2], XIANG CHENG[3], (Senior Member, IEEE), LIUQING YANG[1], (Fellow, IEEE), AND SHUGUANG CUI[4], (Fellow, IEEE)**

[1]Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523, USA
[2]Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA
[3]State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[4]Shenzhen Research Institute of Big Data and Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen 518172, China

Corresponding author: Xiang Cheng (xiangcheng@pku.edu.cn)

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province Project under Grant 2018B030338001, in part by the Natural Science Foundation of China under Grant NSFC-61629101, in part by the Natural Science Foundation under Grant DMS-1737795, Grant DMS-1923142, and Grant CNS-1932413, in part by the Open Research Fund from Shenzhen Research Institute of Big Data under Grant 2019ORF01006, in part by the National Key Research and Development Program of China under Grant 2018YFB1800800, and in part by the Guangdong Zhujiang Project under Grant 2017ZT07X152.

**ABSTRACT** Mobile privacy is broadly concerning in the mobile big data era, as user mobility behaviors are privacy-sensitive and unique. User identification attacks consist of one of the most critical privacy concerns on mobile big data. In this paper, we study mobile privacy in terms of user identifiability from the perspective of privacy adversaries. User identification in two datasets from the same data source or two different data sources is generally formulated as a linear assignment problem (LAP), in which the cost matrix of users is generated by a single distance measure. However, user identification via one single distance measure may lead to a large portion of false matches, especially when only a few users coexist across these two datasets. In addition, the cubic computational complexity of LAP limits the scale of user identification analysis. In this paper, we propose a multi-feature ensemble matching framework to improve the user identification precision based on a majority voting rule, by integrating multiple distance measures. The computational complexity of the proposed ensemble matching algorithm is an order of magnitude less than that of the single-distance based approach, which results from solving an LAP on a highly sparse matrix rather than a dense matrix. Experiments demonstrate the superior performance of our proposed scalable ensemble matching framework with respect to matching precision as well as the vulnerability of mobile network subscribers' privacy.

**INDEX TERMS**

## I. INTRODUCTION

With the explosive growth of mobile phone users, mobile big data collected by mobile network operators start to attract remarkable attention from various research communities [1]–[3]. At the same time, the privacy of mobile big data is primarily concerning, as human mobility is highly regularized and highly predictable [4], [5]. Mobile big data with spatiotemporal information may need to be released to third parties or even to the public, to facilitate various mobile data-driven applications and services. However, data publishing may lead to subscriber privacy leakage threats and risks [1], immediately resulting in data availability issues. For subscribers' privacy protection, the common practice is to anonymize the dataset by replacing subscribers' identifiers (e.g., name, social security number, etc.) with pseudo identifiers. Moreover, the anonymized identifiers are replaced frequently (e.g., every other month) as a data management practice for further privacy protection. However, these practices may not be able to effectively protect subscriber's privacy [6]–[12], due to the uniqueness of human spatiotemporal mobility trajectory. Such uniqueness of subscriber mobility behaviors can lead to another significant concern on subscribers' privacy risk, user identification [10], [13]. In this work, we study the mobile privacy in terms of user identity linkage across two datasets from the privacy attacker's point of view based on their spatiotemporal behaviors, where these two datasets could be collected at different two time periods

The associate editor coordinating the review of this manuscript and approving it for publication was Angelos Antonopoulos.

from the same data source or at the same time period yet from different data sources. The primary purpose of this work is to evaluate subscribers' privacy leakage risk in terms of user identifiability across two datasets. User identification was studied in [10] in terms of a linear assignment problem (LAP) formulation, with the prior knowledge that at most one trace can be exclusively generated by one user in a dataset (termed as exclusiveness). It has been proved [14] that the exclusiveness prior can effectively improve user identification recall performance. The success of the LAP-based user identification relies on effective quantitative distance measures between two spatiotemporal traces. Hence, most work in the literature aim to improve the user identification recall performance with advanced user mobility behavior modeling and effective distance measures between two users.

On the contrary, this paper is aimed to address two other issues of LAP-based user identification with one single distance measure. On the one hand, user identification with one single distance measure may lead to a large portion of false matches, especially when the number of coexisting users in two datasets is small. In this work, we argue that a privacy adversary not only concerns the *recall* performance of a user identification algorithm—how many user pairs out of the total ground-truth user pairs can be identified, but also seriously considers the *precision* performance of a user identification algorithm—how many user pairs out of the total pairs declared by the algorithm are correct. Note that the latter indicates the reliability of a user identification algorithm. In fact, to discover as many as possible correct pairs (i.e., improving the recall performance) has been the primary objective, but with false matches not considered before. Although correctly matched pairs are mostly identified in declared matching results, user's privacy could still be maintained to some extent. That is, correctly matched user pairs could be hidden under large amount of false positives, especially when the number of coexisting users across two datasets is small. On the other hand, the solution to LAP-based user identification problem is computationally expensive ($\mathcal{O}(N^3)$), which cannot handle the large-scale user identification analysis. In addition, the classic LAP algorithms on dense cost matrices are difficult to be horizontally scaled by parallelization, due to the sequential nature of LAP algorithms [15], [16].

To address the issues described previously, a *scalable multi-feature ensemble matching framework* is proposed in this paper, which is aimed to reduce false positives and improve the precision from the perspective of privacy attackers with the computational complexity significantly reduced. The intuition underlying the proposed ensemble matching mechanism is to cross validate matched candidates generated by different semantic spatiotemporal user modeling and their associative distance measures. As a result, a match candidate with minority votes will be considered as a false positive so that the precision of the proposed multi-feature ensemble matching framework can be significantly enhanced, though the recall performance could be slightly compromised. Our proposed ensemble matching approach acts as an information/result fusion inspired by the ''stacking'' approach [17].

The ensemble matching framework is divided into two phases, namely the *vote generation* phase and *final matching* phase. Vote generation is to collect the matching candidates as a vote matrix generated based on different distance measures, while final matching is to produce the final matching result based on the obtained vote matrix. In this work, the proposed ensemble matching framework tackles the high computational complexity problem in both the vote generation and final matching phases. Instead of solving the LAP, a dual-selection strategy in the vote generation phase is proposed by relaxing the exclusiveness constraints in LAP, which can significantly reduce the complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$. The exclusiveness enforcement is moved to the final matching phase. In the final matching phase, a bipartite *partitioning and matching (P&M)* algorithm is proposed in the final matching phase by taking advantage of the extremely high sparsity of the vote matrix. The bipartite P&M algorithm is to first segment the bipartite graph to subgraphs, by solving LAP on which the final user identification result could be generated. As a result, the computational complexity in the final matching phase is significantly reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N \log N)$. The ensemble matching algorithms, *dual-selection ensemble partitioning & matching (DS-Ensemble P&M)*, is proposed by taking advantage of the approaches described previously. It is worthing noting that the proposed ensemble matching framework could be adapted to the general modern big data processing paradigm (e.g., Map-Reduce [18]) to deal with large-scale user identification analysis. The *dual selection* in the voting phase can be paralleled in terms of the matrix rows and columns, while the bipartite P&M could be paralleled with respect to bipartite subgraphs.

Experiment results suggest that the proposed ensemble matching framework can significantly reduce false positives, while the maximal recall performance is slightly compromised. Based on user identification analysis, the mobile privacy of users revealed by the mobile data is at high risk even only with the two-day data collection periods, where the result of two-day identification analysis shows that about 30% users can be identified with 70% confidence. The main contributions of this work are summarized as follows:

- In this paper, we identify the matching precision as another important dimension, on which a privacy adversary is mostly concerned. The proposed ensemble matching framework can significantly improve the precision of user identification with a slightly-compromised recall, compared with the ones based on single distance measure.
- The proposed ensemble matching framework is general and can utilize any data model, feature extraction scheme, and distance measure, so long as they are effective and better than random guesses. This is also an important requirement of weak learners in the traditional ensemble learning.

- The computational complexity of the proposed ensemble matching framework is an order of magnitude less than the one based on the classical LAP formulation, although the proposed ensemble matching framework uses multiple distance measures rather than a single distance measure. The proposed ensemble matching framework can be easily adapted to the modern big data processing paradigm, which can facilitate large scale user identification analysis.

The rest of this paper is organized as follows. In Section II, related works in the literature will be reviewed. Problem description and formulation is addressed in Section III. In Section IV, the ensemble matching framework with the computational complexity analyzed is studied and discussed. Experiment results are presented in Section V to demonstrate the good performance of our proposed framework. Finally, concluding remarks are made in Section VI.

## II. RELATED WORK

Generally, privacy protection is highly concerning in any personal-data-related services and applications. *k-anonymity* is a common metric to evaluate the effectiveness of privacy preservation [19], which requires any record in a database to be indistinguishable to at least $k-1$ other records in the database. The most common anonymization technique is to replace critical identifiers (e.g., phone number, IMEI, etc.) with random pseudo identifiers. However, such identifier anonymization fails for the mobile data with which the subscriber spatiotemporal behavior is recorded, due to the uniqueness of human mobile trajectories [7].

In [20], Zang and Bolot studied a large-scale nationwide dataset with more than 30 billion call records corresponding to 25 million users with different spatial granularities (i.e., cell sector, cell, zip code, city, state). The spatiotemporal footprint of each user is represented by the $N$ most visited places within a pre-defined time (e.g., day, week, month, ect.), based on which the privacy leakage risk could be evaluated. The authors concluded that the spatiotemporal data sharing or publishing that is only anonymized by pseudo identifiers leads to a severe privacy leakage risk. The potential privacy-preserving solution is at least to coarsen the temporal resolution, which restricts the accuracy of extracting $N$ most visited locations from the dataset. However, the privacy protection mechanism, including detail-reduction [21] and obfuscation [22] may primarily reduce the utility of the data. It is concluded in [7] that spatiotemporal resolution curtailments may not be useful as expected, based on a human mobility study with 15-month mobile data and 1.5 million people in a country. That is, the uniqueness reduction is orders of magnitude slower than the resolution coarsening.

Therefore, a generalized scheme on the spatiotemporal privacy preserving based on $k$-anonymity was proposed in [11]. Based on such uniqueness of user mobility behavior, a data-driven spatiotemporal routing generator is developed in [23] to simulate mobility trajectories of users. In addition, it is demonstrated in [24] that the aggregated mobility dataset (e.g., the number of subscribers covered by a cell at a specific time) may also lead to a privacy breach of individual mobility trajectory. In [25], a visualization method is developed to infer one's living address based on twitter check-in data.

The user identification (or user reconciliation) [6], [10], [13], [26]–[34] is another critical problem in privacy protection, which is to link the spatiotemporal records generated by the same user in two datasets. The user identification is closely related to "de-anonymization" attacks. A typical example is the Netflix prize task that is aimed to de-anonymize user identities by public user reviews [26]. Two types of user identification can be roughly categorized, namely matching users from different data domains but in the same time [13], [27], [28] and matching the users from the same data domains in different time spans [10]. In addition, two types of location information, namely actual GPS coordinates [27], [31], [33] and base station location [6], [10], [32], are mainly studied in the literature.

In [6], De Mulder *et al.* studied the user identification based on the location update dataset from GSM networks, which records the phone's network location with geographical information periodically. The mobility Markovian model of each user is constructed based on their spatiotemporal history. However, such a Markovian model requires the dataset with subscribers' transitions among cells to be recorded, whereas such data is not widely adopted or collected by mobile network operators. In [10], [13], user identification is formulated as the minimum (maximum) cost bipartite matching with two vertex sets representing users in two datasets, respectively, where the edge weight is obtained by the distance (similarity) measure between any pair of nodes in the bipartite graph. In [10], Naini *et al.* suppress the temporal information of users' spatiotemporal trajectories and represent the user fingerprint as the histogram of visited location for a given time length. The distance between the two histograms is calculated by the Jensen-Shannon divergence. Instead of temporal information suppression, Riederer *et al.* in [13] models the number of spatiotemporal appearances of a given spatial and temporal bins by a Poisson process for each dataset, based on which the similarity scores could be generated. However, the task of [13] is to identify the user of two datasets from different domains during the same time. In [31], the user matching based on vehicle trajectories is investigated based on the improved term frequency and inverse document frequency (ITF-IDF) mobility feature. The modified Hausdorff distance between two GPS traces has been studied in [27] to distinguish users from different domains. In [29], a privacy risk assessment is studied by assuming that the privacy adversarial has a small portion of the information on users' trajectories, based on which the assessment is aimed to match the prior knowledge with the full record. The privacy leakage assessment is evaluated based on the identification rate—the reciprocal of the total number of users that matched the prior information. In [34], a partition-and-group framework

is proposed to prevent user identification attacks from the adversarial with random prior knowledge.

Although a similar bipartite matching (LAP) formulation for user identification is adopted in this work, the unique contributions of this work stand out from previous works in the following aspects. The ensemble concept or cross-validation via different spatiotemporal features is studied to enhance the robustness of user identification. Accordingly, a scalable ensemble matching algorithm with user grouping and bipartite graph partitioning is proposed, which can reliably re-identify users. Although we study the user identification in the same data domain with different time spans in this work, the proposed ensemble matching framework can be easily extended to the user identification in heterogeneous domains, as it provides an effective and scalable approach to integrate the matching results by diverse distance measures.

## III. PROBLEM STATEMENT

Assume that a spatiotemporal dataset $\mathcal{X}$ is collected by a mobile network operator during a specific time period, in which the $i$-th subscriber with his/her corresponding mobility trace $X_i$ is represented as a tuple, i.e, $(i, X_i) \in \mathcal{X}$. The mobility trace $X_i$ is a sequence of timestamped location points (time $t_h$ and location $x_h$). That is,

$$X_i = [(t_1, x_1), \cdots, (t_h, x_h), \cdots], \quad x_l \in \mathcal{A}, \tag{1}$$

where $\mathcal{A}$ denotes the discrete location point set (i.e., base stations) covered by the mobile network. A typical example of such data is the commonly studied *call detail records (CDR)* [1], which are voice or text event logs collected by network operators for service charging. A privacy attacker can access two of such datasets, $\mathcal{X}$ and $\mathcal{Y}$, collected in two time periods.

### A. USER IDENTIFICATION PROBLEM [10]

Without loss of generality, the true user identity information of dataset $\mathcal{Y}$ is assumed to be known to the privacy attacker. The attacker attempts to connect the spatiotemporal information generated by the same user in datasets $\mathcal{X}$ and $\mathcal{Y}$ based on their attributes, despite these mobility traces are associated with different anonymized IDs within these two datasets. By the assumption that each user can have at most one record in a dataset [10], [14], the user identification problem can be formulated into a *k-cardinality linear assignment problem* (kLAP) as in [10], that is,

$$\begin{aligned} &\underset{c_{ij}}{\text{minimize}} \sum_{i}^{N}\sum_{j}^{M} c_{ij} w_{ij} \\ &\text{subject to} \sum_{j}^{M} c_{ij} \le 1, \ \sum_{i}^{N} c_{ij} \le 1, \ c_{ij} \in \{0, 1\} \\ &\qquad \sum_{i}^{N}\sum_{j}^{M} c_{ij} = k, \quad \forall j \in [M], \ \forall i \in [N], \end{aligned} \tag{2}$$

where $N = |\mathcal{X}|$ and $M = |\mathcal{Y}|$ denote the number of users in datasets $\mathcal{X}$ and $\mathcal{Y}$, respectively.[1] $k$ denotes the number of users coexisting in both the datasets with different anonymized IDs, i.e., $k = |\mathcal{X} \cap \mathcal{Y}|$. It is worth noting that such a kLAP formulation would mainly take advantage of the prior knowledge that one user can generate at most one record in one dataset, termed as *exclusiveness* in this paper. The classic solution to the LAP problem is the *Kuhn-Munkres (Hungarian)* algorithm [15] and the *Jonker-Volgenant (JV)* algorithm [16], both of which the complexity is $\mathcal{O}(N^3)$.

The weight $w_{ij}$ in (2) denotes the distance between user $i$ from $\mathcal{X}$ and user $j$ from $\mathcal{Y}$, i.e.,

$$w_{ij} = \Delta(X_i, Y_j), \tag{3}$$

where $\Delta(X_i, Y_j)$ is a distance measure between two traces, $X_i$ and $Y_j$. In fact, user identification performance will be determined by how well the weight $w_{ij}$ can measure two mobility traces, as kLAP is deterministic once weights are decided.

### B. ISSUES OF SINGLE-DISTANCE-BASED USER IDENTIFICATION

To evaluate the performance of distance measures (a.k.a. user identification performance), the criterion—how many correct pairs out of the ground truth across two datasets are identified (a.k.a., *recall*)—is mostly concerning as stated in [10], [14]. As the nature of bipartite matching problem, one incorrect matched user pair may lead to multiple incorrect matched user pairs in the matching result. However, we argue that a privacy adversarial not only concerns about the recall performance but also want a robust user identifier, which is evaluated by *precision* — the ratio of the number of correctly identified pairs over the total number of declared pairs.

The argument results from the reality that the privacy adversarial does not have the prior knowledge of how many users coexists in two datasets. In other words, $k$ in (3) is unknown. As a common practice, one would assume a maximum coexisting user number $k$ (i.e., $k = \min(N, M)$ [10]) in (2) to extract as many correct pairs as possible, regardless of inevitable false positives. Nevertheless, such assumption will definitely lead to inferior precision performance (many falsely matched user pairs are included in the matching), especially when $k$ is less than $\min(N, M)$. User identification based on single distance measure could lead to a large amount of false positives, as the mobility behaviors of two users could be similar because of details reduction during mobility behavior modeling. On the other hand, the high computational complexity of LAP ($O(N^3)$) prevents the single-distance-measure-based user identifier from dealing with large-scale user identification.

### C. ENSEMBLE MATCHING FOR USER IDENTIFICATION

It is intuitive that cross validation based on multiple sources can effectively eliminate false positives. In this paper, we pro-

[1] Without loss of generality, we assume $N \le M$ in the rest of this paper.

pose a *scalable ensemble matching framework* to cross validate and identify users, taking advantage of multiple distances modeled in different aspects. Generally, each distance measure is modeled and extracted from a perspective of user mobility. As a result, the proposed multi-feature ensemble matching framework is aimed to cross validate the identified candidates by diverse semantic features and eventually determine the final matching via majority voting strategy. Accordingly, false positives could be eliminated.

Moreover, our proposed ensemble matching framework can reduce the computational complexity from $O(N^3)$ to $O(N^2)$, compared with the one based on single-distance-measure matching. The complexity reduction of the proposed ensemble matching framework results from that one only needs to solve LAP on a highly-sparse matrix rather than a dense matrix in the proposed ensemble matching framework. Thus, the proposed ensemble matching has the capability of dealing with large scale user identification analysis.

## IV. ENSEMBLE MATCHING FRAMEWORK

Let $\mathcal{W}$ denote the set of $G$ user distance matrices,

$$\mathcal{W} = \left\{ \boldsymbol{W}^1, \boldsymbol{W}^2, \cdots, \boldsymbol{W}^G \right\} , \tag{4}$$

where the element of each distance matrix is the pair-wise distance $w_{ij}$ generated by a specific distance measure between user $i$ from $\mathcal{X}$ and user $j$ from $\mathcal{Y}$. Accordingly, the proposed ensemble matching is aimed to generate reliable user pairs between datasets $\mathcal{X}$ and $\mathcal{Y}$ based on the distance matrix set $\mathcal{W}$, consisting of two phases:

- *Vote Generation:* the vote generation phase is to identify the matched candidates corresponding to each distance matrix so that a sparse vote matrix could be obtained;
- *Final Matching:* the final matching phase is to generate the final matching result on the generated sparse vote matrix with majority voting and exclusiveness property ensured.

Based on different vote generation strategies, we first propose two ensemble matching algorithms in this paper, namely *matching-filtered ensemble (MF-Ensemble) matching* and *dual-selection ensemble (DS-Ensemble) matching*. For the final matching phase, we further propose a low-complexity *Partitioning and Matching (P&M)* algorithm by taking advantage of the extreme sparsity of vote matrix.

### A. MATCHING-FILTERED ENSEMBLE MATCHING

Each distance matrix can produce $k$ matched pairs by (2) from the perspective of its underlying spatiotemporal modeling and user representation. Such matching based on kLAP (2) can be regarded as a filter to select $k$ matched candidates out of massive $\binom{N}{k}\binom{M}{k}k!$ possibilities. Therefore, the first phase of the proposed ensemble matching mechanism—vote generation—is fulfilled based on kLAP matching, which termed as matching-filtered ensemble (MF-Ensemble) matching.

With distance matrix set $\mathcal{W}$, let matrix $\boldsymbol{C}^{(g,k)} \in \{0,1\}^{N\times M}$ denotes the matching result by (2) based on the $g$-th distance measure with the assumption of $\hat{k}$ coexisting user number,

$$\boldsymbol{C}^{(g,\hat{k})} = \text{kLAP}(\boldsymbol{W}^g, \hat{k}) \, .$$

Let vote matrix $\boldsymbol{V}_{\text{MF}}^{\hat{k}} \in \mathcal{Z}^{N\times M}$ collect the matching results by total $G$ distance measures on each possible matching pair, that is,

$$\boldsymbol{V}_{\text{MF}}^{(\hat{k})} = \sum_{g=1}^{G} \boldsymbol{C}^{(g,\hat{k})}. \tag{5}$$

Therefore, by the strategy of majority votes, the proposed MF ensemble matching algorithm is aimed to maximize the sum vote by solving following combinatoric optimization problem,

$$\begin{aligned} \underset{z_{ij}^{\hat{k}}}{\text{maximize}} \ & \sum_{i}^{N}\sum_{j}^{M} z_{ij}^{\hat{k}} v_{ij,\text{MF}}^{(\hat{k})} \\ \text{subject to} \ & \sum_{j}^{M} z_{ij}^{\hat{k}} \le 1, \ \sum_{i}^{N} z_{ij}^{\hat{k}} \le 1, \ z_{ij}^{\hat{k}} \in \{0,1\} \\ & z_{ij}^{\hat{k}}(v_{ij,\text{MF}}^{(\hat{k})} - \tau) \ge 0, \quad \forall i \in [N], j \in [M]. \end{aligned} \tag{6}$$

where $\boldsymbol{Z}^{\hat{k}} \in \{0,1\}^{N\times M}$ denotes the final result generated by the proposed MF-Ensemble algorithm. The first two conditions in (6) are the same as kLAP (2), which guarantee the exclusiveness property. The $\tau$ denotes the *vote threshold* that ensures that the final result is voted by majority, whose typical value is $\tau = \lceil G/2 \rceil$. Thus, the third condition, $z_{ij}^{\hat{k}}(v_{ij,\text{MF}}^{(\hat{k})} - \tau) \ge 0$, is designed to ensure the solution voted by majority. The objective function in (6) is aimed to maximize total votes generated by multiple distance measures without any specific restriction on the cardinality of final results, as the cardinality restriction condition has already been enforced in (2) before ensemble matching.

In fact, the intuition behind sum vote maximization in (6) is to choose the one with more votes when the selection of certain two candidate pairs violates the exclusiveness property, e.g.,

$$\max(v_{ij}, v_{il}), \quad v_{ij} \ge \tau, \ , \ v_{il} \ge \tau.$$

Moreover, we reformulate (6) into the classical linear assignment problem (LAP) as follows,

$$\begin{aligned} \underset{z_{ij}}{\text{minimize}} \ & \sum_{i}^{n}\sum_{j}^{m} z_{ij} \left( G - v_{ij,\text{MF}}^{\hat{k},\tau} \right) \\ \text{subject to} \ & \sum_{j}^{M} z_{ij} = 1, \ \sum_{i}^{N} z_{ij} \le 1 \\ & \forall i \in [N], \ \forall j \in [M], \ z_{ij} \in \{0,1\}. \end{aligned} \tag{7}$$

**Algorithm 1** Matching-Filtered Ensemble Matching

1: **Input**: $\mathcal{W} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, \cdots, \boldsymbol{W}^G\}, k, \tau$
2: **Output**: $\boldsymbol{Z}$
3: Initiating vote collection matrix $\boldsymbol{V} = \boldsymbol{0}$
4: **for** $g \in \{1, 2, \cdots, G\}$ **do** ▷ for each distance measure
5: $\quad \boldsymbol{C}^{(g,k)} \leftarrow \text{kLAP}(\boldsymbol{W}^g, k)$ ▷ solve (2) on $\boldsymbol{W}^g$
6: $\quad \boldsymbol{V} \leftarrow \boldsymbol{V} + \boldsymbol{C}^{(g,k)}$
7: **end for**
8: $\boldsymbol{Z} \leftarrow \text{LAP}(\boldsymbol{V}, G, \tau)$ ▷ solve (7)
9: **for** $i, j \in [N] \times [M]$ **do**
10: $\quad$ **if** $z_{ij} <> 0$ and $v_{ij}^{\hat{k}} == 0$ **then**
11: $\quad\quad z_{ij} \leftarrow 0$ ▷ remove non-major-voted
12: $\quad$ **end if**
13: **end for**

where $v_{ij,\text{MF}}^{\hat{k},\tau} = \xi(v_{ij,\text{MF}}^{(\hat{k})}, \tau)$ denotes the votes after thresholding as follows,

$$\xi(v, \tau) = \begin{cases} v, & v \geq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Via the Hungarian or JV algorithm of the classic LAP, $N$ pairs are generated (Line 8, Algorithm 1), from which final results are determined by removing the matched pairs whose votes do not satisfy $v_{ij,\text{MF}}^{\hat{k},\tau} > 0$. Details of the proposed ensemble matching framework are demonstrated in Algorithm 1.

### B. DUAL-SELECTION ENSEMBLE MATCHING

The proposed MF-Ensemble matching needs to solve kLAP $G$ times in the vote generation phase and solve the LAP in the final matching phase, both of which the computational complexity is $\mathcal{O}(N^3)$. Such high computational complexity may make the proposed MF-Ensemble matching infeasible when user size is large. In fact, the MF-Ensemble algorithm enforces the exclusiveness property in both the vote generation phase and the final matching phase, which may not be necessary. Therefore, we propose a dual selection strategy in the vote generation phase by relaxing the exclusiveness constraint in the vote generation phase, termed as *dual-selection ensemble (DS-Ensemble) matching.*

For each distance matrix $\boldsymbol{W}^g$, matched candidates can be generated based on the minimum distance in terms of each user in both the datasets. For example, each user in dataset $\mathcal{X}$ would select the most similar user from dataset $\mathcal{Y}$ in terms of distance measure $g$, i.e.,

$$\mathcal{C}^{(g,\mathcal{X})} = \left\{ (i,j) \,\middle|\, j = \arg\min_{j \in [M]} w_{ij}^g, i \in [N] \right\} \tag{9}$$

Similarly, each user in dataset $\mathcal{Y}$ can again identify their candidates, i.e.,

$$\mathcal{C}^{(g,\mathcal{Y})} = \left\{ (i,j) \,\middle|\, i = \arg\min_{i \in [N]} w_{ij}^g, j \in [M] \right\} \tag{10}$$

Therefore, such procedure is termed as *dual selection* (Line 7&10, Algorithm 2). By regarding each pair via the dual selection procedure as one vote, the candidate matrix takes the form as follows,

$$\boldsymbol{C}_{ij}^{(g,\{\cdot\})} = \begin{cases} 1 & (i,j) \in \mathcal{C}^{(g,\{\cdot\})} \\ 0 & \text{otherwise,} \end{cases}$$

where $\{\cdot\}$ denotes dataset $\mathcal{X}$ or $\mathcal{Y}$. Hence, the final candidate matrix can be obtained by superimposing these two candidate matrix (Line 12 in Algorithm 2), i.e.,

$$\boldsymbol{C}^g = \boldsymbol{C}^{(g,\mathcal{X})} + \boldsymbol{C}^{(g,\mathcal{Y})}. \tag{11}$$

It is worth noting that each true pair can get two votes for one distance matrix in the ideal case. Also, an incorrect selection in one dataset does not impact the selection of the other, as the selection in two datasets are independent of each other. The computational complexity of dual selection is $\mathcal{O}(NM)$, an order of magnitude less than matching filtering based vote generation.

The vote collection can be achieved according to (5), i.e., $\boldsymbol{V}_{\text{DS}} = \sum_g \boldsymbol{C}^g$. In the DS-Ensemble matching algorithm, the final matching phase needs to ensure $k$-cardinality condition and exclusiveness property, in order to determine the final matching. Similar to (7), the DS-Ensemble matching algorithm is to solve the assignment problem with the constraint of majority voting (Line 14, Algorithm 2) as follows,

$$\begin{aligned} \underset{z_{ij}}{\text{minimize}} \quad & \sum_i^n \sum_j^m z_{ij}[2G - v_{ij,\text{DS}}^{\tau}] \\ \text{subject to} \quad & \sum_j^M z_{ij} < 1, \sum_i^N z_{ij} \leq 1, z_{ij} \in \{0,1\} \\ & \sum_i^N \sum_j^M z_{ij} = \hat{k}, \quad \forall i \in [N], \ \forall j \in [M], \end{aligned} \tag{12}$$

where $v_{ij,\text{DS}}^{\tau} = \xi(v_{ij,\text{DS}}, \tau)$. It is worth noting that the maximum votes for one pair $(i,j)$ through dual selection procedure is $2G$. Thus, the majority voting threshold is $\tau = G$. Details of the proposed DS-Ensemble matching can be found in Algorithm 2.

### C. DUAL-SELECTION ENSEMBLE PARTITIONING & MATCHING

The DS-ensemble matching can reduce the computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM)$ in the vote generation phase, compared with the MF-ensemble matching algorithm. Nonetheless, the scalability issue of our proposed ensemble matching framework still exists due to the high-complexity LAP-based approach in the final matching phase (i.e., (7) and (12)). In this subsection, we aim to tackle the scalability issue in the final matching phase by reducing the average time complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N \log N)$.

It is worth noting that the vote matrix $\boldsymbol{V}$ is an extremely sparse matrix, as each candidate matrix by a distance measure (i.e., $\boldsymbol{C}^g$) has at most $2N$ non-zero elements. Besides, the superimposition of a total $G$ candidate matrices will further reduce the number of non-zero elements in the vote

**Algorithm 2** Dual-Selection Ensemble Matching

1: **Input**: $\mathcal{W} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, \cdots, \boldsymbol{W}^G\}, k, \tau$
2: **Output**: $\boldsymbol{Z}$
3: Initiating vote collection matrix $\boldsymbol{V} = \boldsymbol{0}$
4: **for** $g \in \{1, 2, \cdots, G\}$ **do** ▷ candidate dual selection
5: $\boldsymbol{C}^{(g,\mathcal{X})} \leftarrow 0, \boldsymbol{C}^{(g,\mathcal{Y})} \leftarrow 0$
6: **for** $i \in [N]$ **do**
7: $j \leftarrow \arg\min_j \boldsymbol{W}^g_{ij}, \boldsymbol{C}^{(g,\mathcal{X})}_{ij} \leftarrow 1$ ▷ solve (9)
8: **end for**
9: **for** $j \in [M]$ **do**
10: $i \leftarrow \arg\min_i \boldsymbol{W}^g_{ij}, \boldsymbol{C}^{(g,\mathcal{Y})}_{ij} \leftarrow 1$ ▷ solve (10)
11: **end for**
12: $\boldsymbol{V} \leftarrow \boldsymbol{V} + \boldsymbol{C}^{(g,\mathcal{X})} + \boldsymbol{C}^{(g,\mathcal{Y})}$
13: **end for**
14: $\boldsymbol{Z} \leftarrow \text{kLAP}(\boldsymbol{V}, G, \tau)$ ▷ solve (12)
15: **for** $i, j \in [N] \times [M]$ **do**
16: **if** $z_{ij} <> 0$ and $v_{ij} < \tau$ **then**
17: $z_{ij} \leftarrow 0$ ▷ remove non-major-voted
18: **end if**
19: **end for**

matrix $\boldsymbol{V}$. In the worst case, the nonzero element number of vote matrix $\boldsymbol{V}$ will be at the level of $\mathcal{O}(GN)$, where $G \ll N$. Also, the majority voting strategy can further reduce the nonzero element number of vote matrix $\boldsymbol{V}$, as each element of $\boldsymbol{V}$ less than vote threshold $\tau$ will be set to zero. One can regard the vote matrix as the *adjacency matrix* of a bipartite graph, $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \boldsymbol{V})$, whose nonzero elements can be regarded as weighted edges between two vertex sets in the bipartite graph. The intuition to resolve the scalability issue in the final matching phase is to first partition the bipartite graph into subgraphs and then conduct matching on the subgraphs to generate final matching results, by taking advantage of the high sparsity of $\boldsymbol{V}$.

#### 1) BIPARTITE GRAPH PARTITIONING WITHOUT LOSS

The sparsity of vote matrix indicates that the entire bipartite graph may be partitioned without loss of votes, as most of the matrix elements are already zero. In other words, vote matrix $\boldsymbol{V}$ may be rearranged into the block diagonal form by shuffling rows and columns as follows,

$$\boldsymbol{V} = \text{diag}\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_r\}, \tag{13}$$

where $\boldsymbol{V}_i$ denotes a submatrix of $\boldsymbol{V}$ that cannot be further diagonalized without loss of nonzero elements. As a result, one could perform bipartite matching (i.e., (12)) on each submatrix $\boldsymbol{V}_i$ to generate final matching results. Rearranging the vote matrix $\boldsymbol{V}$ into a block diagonal form is equivalent to searching the connected components of the bipartite graph $\boldsymbol{V}$. Hence, an efficient tree-based data structure in the literature, *union find* or *disjoint set* [35, Chapter 1], can be easily employed to find the connected components with the time complexity $\mathcal{O}(GN)$.

#### 2) BIPARTITE GRAPH PARTITIONING WITH LOSS

Although the sparsity of the vote matrix may reduce the size for bipartite matching without loss of nonzero elements, the size of each submatrix cannot either be controllable nor be guaranteed to be small enough. In the worse case, the bipartite graph $\boldsymbol{V}$ cannot be partitioned at all, especially when users have very similar mobility behaviors, while vote matrix $\boldsymbol{V}$ still remains extremely sparse as previously discussed. As a result, we propose to partition the bipartite graph with the minimum nonzero loss, where the size of submatrices could be controllable to a certain degree.

Starting from binary partitioning (i.e., each vertex set of a bipartite graph is partitioned into two subsets), the minimization of *normalized cut* is commonly employed as an objective function for bipartite graph partitioning [36], [37]. Let vote matrix $\boldsymbol{V}$ be expressed in a block format as follows,

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}.$$

where $\boldsymbol{V}_{ij}$ corresponds to vertex subsets $X_i$ and $Y_j$, $i, j \in \{1, 2\}$. Thus, the normalized cut is defined as follows,

$$\text{NCut} = \frac{\text{Cut}}{2\boldsymbol{1}^T\boldsymbol{V}_{11}\boldsymbol{1} + \text{Cut}} + \frac{\text{Cut}}{2\boldsymbol{1}^T\boldsymbol{V}_{22}\boldsymbol{1} + \text{Cut}}, \tag{14}$$

where $\text{Cut} = (\boldsymbol{1}^T\boldsymbol{V}_{12}\boldsymbol{1} + \boldsymbol{1}^T\boldsymbol{V}_{21}\boldsymbol{1})$ denotes the loss of elements due to bipartite graph partitioning. It is worth noting that the normalized cut minimization is not only aimed to minimize the loss of elements due to graph partitioning, but also designed to balance the partitioning (i.e., the cardinality difference between two vertex subsets should approach to zero).

It has been shown in [37] that the normalized cut minimization based bipartite graph partitioning can boil down to finding the *second largest singular vectors ($\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$)* of $\tilde{\boldsymbol{V}} = \boldsymbol{D}_X^{-1/2}\boldsymbol{V}\boldsymbol{D}_Y^{-1/2}$

$$\tilde{\boldsymbol{V}}\tilde{\boldsymbol{u}} = \sigma_2\tilde{\boldsymbol{v}},$$

where $\boldsymbol{D}_X = \text{diag}\{\boldsymbol{V}\boldsymbol{1}\}$ and $\boldsymbol{D}_Y = \text{diag}\{\boldsymbol{V}^T\boldsymbol{1}\}$ denote the degree of each vertex in $X$ and $Y$, respectively. As a result, both user set $X$ and $Y$ can be segmented as follows,

$$\mathcal{X}_1 = \{i|\boldsymbol{u}_i \geq 0\} \text{ and } \mathcal{Y}_1 = \{j|\boldsymbol{v}_j \geq 0\}, \tag{15}$$

where $\boldsymbol{u} = \boldsymbol{D}_X\tilde{\boldsymbol{u}}$ and $\boldsymbol{v} = \boldsymbol{D}_Y\tilde{\boldsymbol{v}}$. Furthermore, $\mathcal{X}_2$ and $\mathcal{Y}_2$ can be obtained by finding the complement of $\mathcal{X}_1$ and $\mathcal{Y}_1$, respectively.

#### 3) DS-ENSEMBLE PARTITIONING AND MATCHING

Based on the previous discussions on vote matrix sparsity and bipartite graph partitioning, we propose a recursive *DS-Ensemble partitioning and matching (P&M)* algorithm with a much lower computational complexity, compared with the DS-Ensemble matching. First, the bipartite graph will be partitioned without loss based on the *union find* (Algorithm 3 Line 4). For each subgraph, a recursive partitioning and matching algorithm (Algorithm 3 Line 12-29) is employed to further segment the graph into multiple subgraphs, whose

**Algorithm 3** DS-Ensemble Partitioning&Matching

1: **Input**: $\boldsymbol{V}_{\mathrm{DS}}, X, Y, \hat{k}, G, \tau, t$
2: **Output**: $\boldsymbol{Z}$
3: $\boldsymbol{V}_{\mathrm{DS}}^{\tau} = \xi(\boldsymbol{V}_{\mathrm{DS}}, \tau)$ ▷ vote thresholding
4: $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_R \leftarrow \mathrm{UnionFind}(\boldsymbol{V}_{\mathrm{DS}}^{\tau})$ ▷ partitioning w/o loss
5: Initialize $\mathcal{Z} \leftarrow \emptyset$
6: **for** $r \in [R]$ **do** ▷ partitioning w/ loss
7: $\boldsymbol{Z}_r \leftarrow \mathrm{PartitionAndMatch}(\boldsymbol{V}_r, X_r, Y_r)$
8: $\mathcal{Z} \cap \{\boldsymbol{Z}_r\}$
9: **end for**
10: $\boldsymbol{Z} \leftarrow \mathrm{diag}(\mathcal{Z})$
11: Find top $\hat{k}$ pair based on $\boldsymbol{Z}$ and $\boldsymbol{V}_{\mathrm{DS}}^{\tau}$
12: **function** PartitionAndMatch($\boldsymbol{V}, X, Y$)
13: **if** $|X| > t$ or $|Y| > t$ **then** ▷ partitioning
14: $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}} \leftarrow \mathrm{Lanczos}(\boldsymbol{D}_X^{-1/2}\boldsymbol{V}\boldsymbol{D}_Y^{-1/2})$
15: $X_1 \leftarrow \{i|(\boldsymbol{D}_X\hat{\boldsymbol{u}})_i \geq 0\}$ and $X_2 \leftarrow X - X_1$
16: $Y_1 \leftarrow \{j|(\boldsymbol{D}_Y\hat{\boldsymbol{v}})_j \geq 0\}$ and $Y_2 \leftarrow Y - Y_1$
17: $\boldsymbol{Z}_{11} \leftarrow \mathrm{PartitionAndMatch}(\boldsymbol{V}_{11}, X_1, Y_1)$
18: $\boldsymbol{Z}_{22} \leftarrow \mathrm{PartitionAndMatch}(\boldsymbol{V}_{22}, X_2, Y_2)$
19: $\boldsymbol{Z} \leftarrow \mathrm{diag}\{\boldsymbol{Z}_{11}, \boldsymbol{Z}_{22}\}$
20: **else** ▷ matching
21: $\boldsymbol{Z} \leftarrow \mathrm{LAP}(\boldsymbol{V}, 2G)$
22: **end if**
23: **for** $i, j \in [N] \times [M]$ **do** ▷ clean via majority voting
24: **if** $z_{ij} <> 0$ and $v_{ij} < \tau$ **then**
25: $v_{ij} \leftarrow 0$
26: **end if**
27: **end for**
28: **return** $\boldsymbol{Z}$
29: **end function**

size is not greater than the size threshold ($t$ in Algorithm 3). In each final subgraph, the Hungarian or JV algorithm (Algorithm 3 Line 21) will be employed to obtain the final matching result (7) with the majority voting ensured (Algorithm 3 Line 23-27). After collecting all the matching pairs from all the subgraphs, one can output the top-$\hat{k}$ matched pairs. It is worth noting that the DS-Ensemble P&M algorithm is a suboptimal algorithm, compared with the DS-Ensemble matching. For Details of the algorithm, refer to Algorithm 3.

In the DS-Ensemble P&M algorithm, the heaviest computational load of bipartite graph partitioning is the second largest singular vector calculation, where the full singular value decomposition is computationally intensive (i.e., $\mathcal{O}(N^3)$). However, thanks to the high sparsity of $\tilde{\boldsymbol{V}}$, the computational complexity of the second largest singular vector calculation can be reduced to $\mathcal{O}(\mathrm{nnz}(\tilde{\boldsymbol{V}}))$ based on Lanczos method in [37] and [38, Chapter 8], where $\mathrm{nnz}(\tilde{\boldsymbol{V}})$ denotes the number of nonzeros in matrix $\tilde{\boldsymbol{V}}$. As a result, the time complexity of binary bipartite graph partitioning is $\mathcal{O}(GN)$. The recursive number of bipartite graph partitioning depends on the size threshold $t$. By roughly assuming that each bipartite graph partitioning can exactly divide the graph into two equal-size subgraphs, the recursive number is $\mathcal{O}(\log(N/t))$, and each recursive layer is on the complexity of $\mathcal{O}(GN)$. Thus, the computational complexity of the proposed P&M algorithm is $\mathcal{O}(Nt^2 + \log(N/t)GN)$, where $\mathcal{O}(Nt^2)$ originates from $N/t$ matchings on subgraphs with the size less than $t$. As $t$ and $G$ are fixed and predefined, the average computational complexity of the DS-Ensemble P&M algorithm could be simplified to $\mathcal{O}(NM + N\log N)$, where $\mathcal{O}(NM)$ originates from the dual selection procedure in the vote generation phase.

## V. EXPERIMENTS

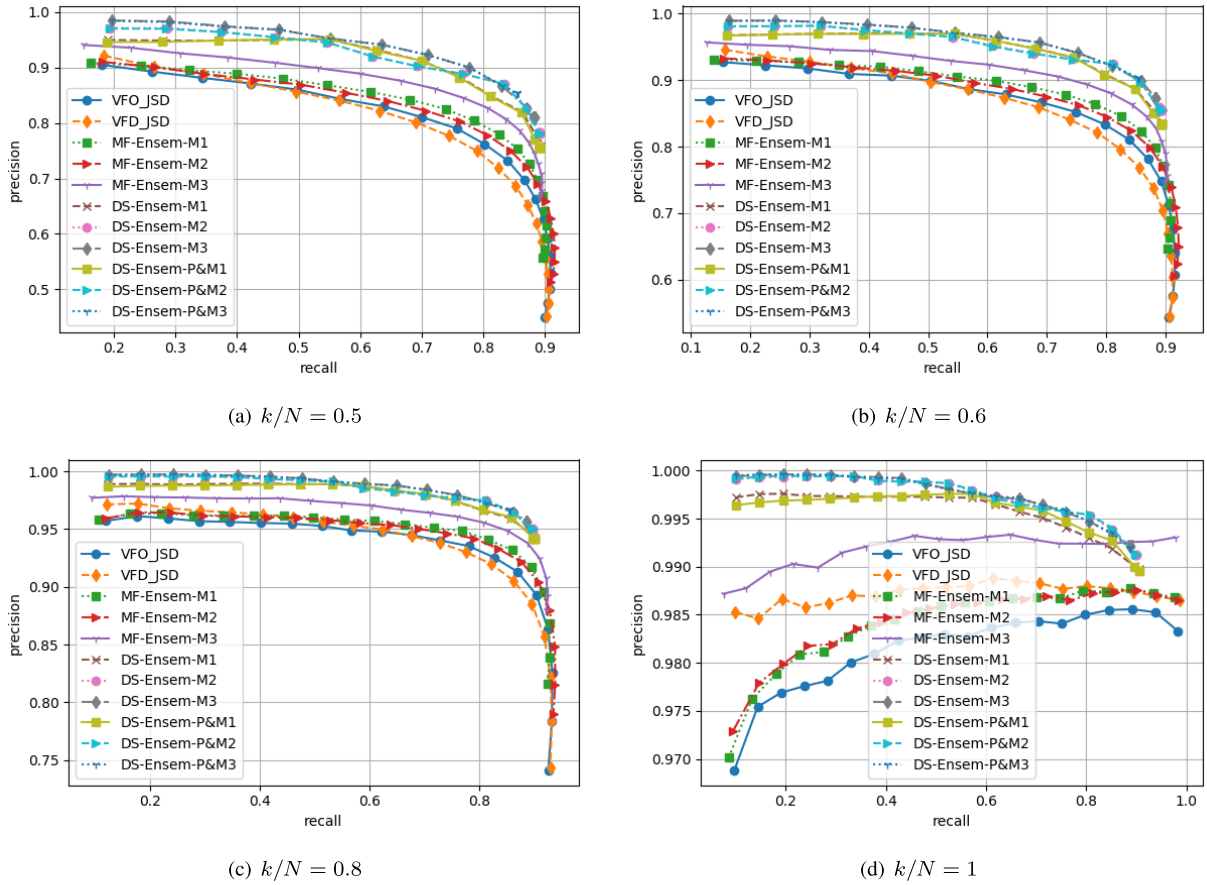
In this section, we validate our proposed ensemble matching via experiments on a real-world signaling dataset [39] collected in a mobile network, which is an extension of the commonly studied call detail record (CDR) dataset.

### A. STUDIED DATASET

The signaling data is a typical example of control-plane data collected from mobile networks [1], which is collected at the mobility management entity of LTE networks. The signaling dataset records every communication/location update event of all active subscribers in a mobile network. Data fields of the signaling data include *subscriber's anonymized identifier*, *time stamp*, *location coordinates* (i.e., the longitude and latitude of the base station), *event type*, and *cell type* (i.e., small cell or macro cell). In addition, the signaling data logs event type as well as the direction of the event (e.g., initiating a call or being called). Compared with the commonly used call detail record (CDR) data, the signaling data does not record the duration information of voice services. However, the signaling data further logs two types of location update events in addition to the regular event types (calls or texts), namely the regular location update and the periodic location update. In cellular networks, location updating is a fundamental technique of idle mobile device mobility management. Regular location updates are triggered by tracking area crossing, while periodic location updates are prompted by a timeout event when no event occurs for a subscriber within a predefined time period. In the studied dataset, the timeout interval is about 1 hour, which can guarantee that any active subscribe has at least one observation per hour in the dataset.

To mimic the data publish process, three scenarios generated from the studied dataset are tested to evaluate the proposed multi-feature ensemble matching framework, in comparison with the existing methods in the literature.

- In *Scenario 1*, a two-week dataset of total 15, 000 users is recorded from July 1st, 2016 to July 14th, 2016. This scenario is employed to mimic that the network operator publishes a different subset of subscribers during different periods. In the experiments, 10, 000 users in the first week and the second week are randomly sampled out of total 15, 000 as dataset $\mathcal{X}$ and $\mathcal{Y}$, respectively, while the number of overlapping users over two datasets $k$ will be




(a) $k/N = 0.5$ (b) $k/N = 0.6$

(c) $k/N = 0.8$ (d) $k/N = 1$

**FIGURE 1.** **Performance comparison in Scenario 1.**

specified later. The experiment results shown later are average of 10 such tests.

- In *Scenario 2*, a two-week dataset of total 5976 users are extracted based on their mobility behaviors in specific regions of interest for certain applications. This scenario is to mimic that the network operator publishes a region-based dataset. Although the total user number of this scenario is less than the one in Scenario 1, the user identification in this scenario is much more challenging, for users in this scenario have much more similar mobility behavior. Similar to Scenario 1, 2, 500 users are randomly sampled out of total 5976 users, while the number of coexisting user $k$ is controlled in different test cases. Again, dataset $\mathcal{X}$ and $\mathcal{Y}$ covers the first week and the second week, respectively.
- *Scenario 3* is aimed to illustrate privacy risk via large-scale user identification analysis on 150, 000 users with different data collection periods, ranging from 2 days to 7 days. It is worth noting that single-distance-based LAP user identification cannot deal with such a scale of users.

### B. USER IDENTIFICATION PERFORMANCE

To evaluate the user identification performance, the classical *precision-recall* is employed. The precision metric is to assess how accurate a user identification algorithm is, which is defined as

$$\text{precision} = \frac{\text{correct identified pairs \#}}{\text{total declared pairs \#}}. \tag{16}$$

The recall metric is to evaluate how many user pairs can be correctly identified out of the total number of coexisting users $k$,

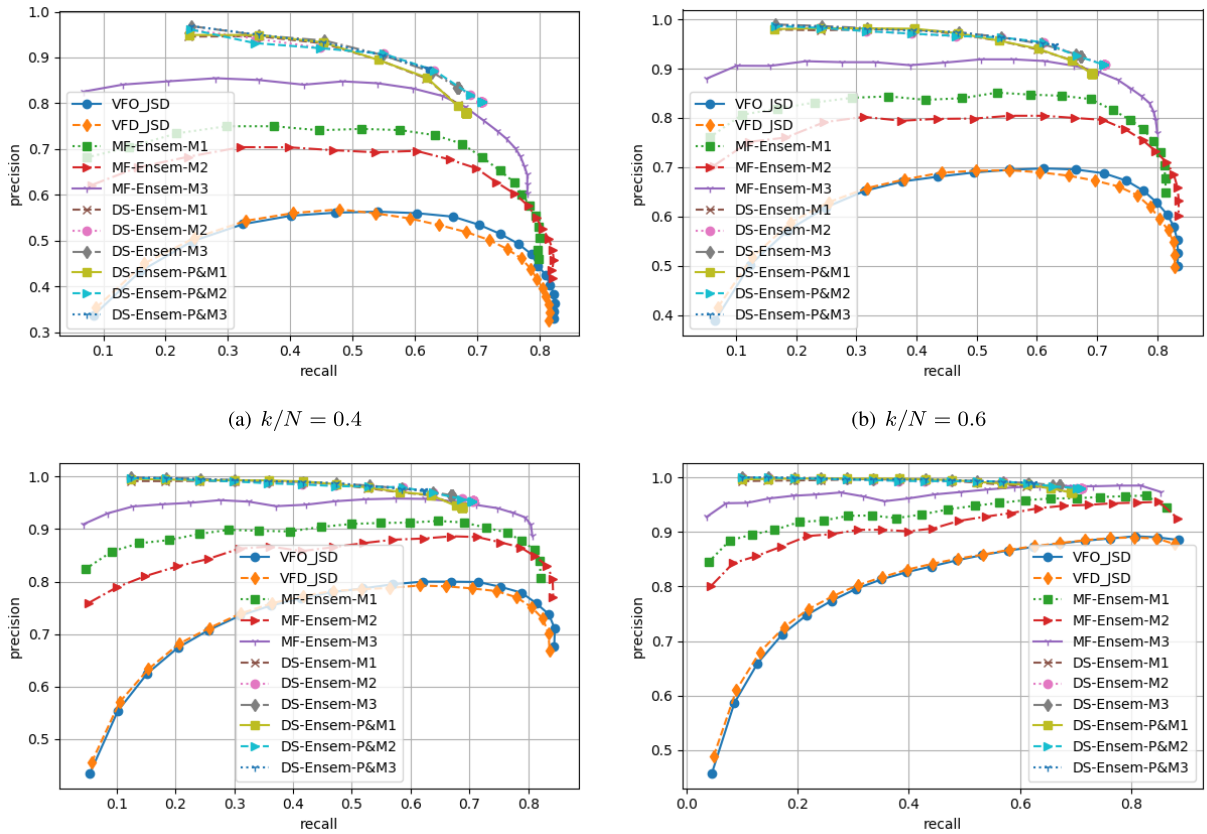
$$\text{recall} = \frac{\text{correct identified pairs \#}}{\text{total existed pairs \#}}. \tag{17}$$

In general, the tradeoff between the precision and recall could be controlled by a preset parameter $\hat{k}$. In the experiments, we use $\hat{k} \in [0.01 \times N, 0.02 \times N, \cdots, N]$ to generate precision-recall points in the figures.

Figs. 1 and 2 show the precision-recall comparison between the best single matching (VFO-JSD & VFD-JSD) [40], MF-Ensemble Matching (Algorithm 1), DS-Ensemble

**TABLE 1.** **Distance measures for each ensemble.**

| | |
|---|---|
| Ensemble 1 | VFO-L1, VFO-JSD, VFD-JSD |
| Ensemble 2 | VFO-L1, VFO-JSD, VFD-JSD, DHR-IOU |
| Ensemble 3 | VFO-L1, VFO-JSD, VDO-JSD, VFD-JSD, DHR-IOU |




(a) $k/N = 0.4$ (b) $k/N = 0.6$

(c) $k/N = 0.8$ (d) $k/N = 1$

**FIGURE 2.** Performance comparison in Scenario 2.

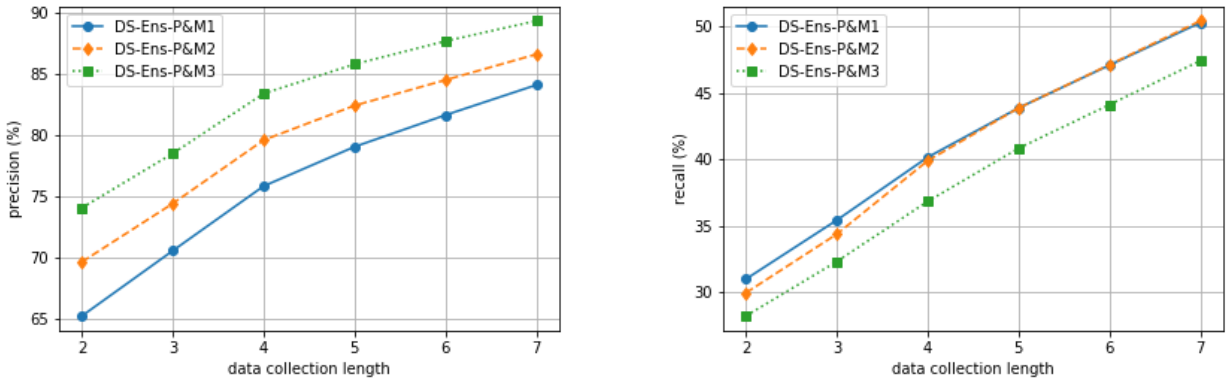
Matching (Algorithm 2), and DS-Ensemble P&M (Algorithm 3) algorithms. The distance measures involved in each ensemble can be found in Table 1, and their details can be found in Appendix. It can be observed in both scenarios that the ensemble can broadly outperform the best individual matching in terms of the precision. The performance gain is more significant as one distance measure is not capable of distinguishing certain user pairs, due to their similar residency areas. However, the proposed MF-Ensemble algorithm can effectively take advantage of multiple diverse distance measures.

The tradeoff between the maximum recall and precision rates can also be observed in Figs. 1 and 2. In other words, the proposed ensemble matching framework can achieve much higher precision at the same recall, while it has a smaller maximum recall compared with the individual ones (i.e., the absolute user pairs that the ensemble matching can discover is less than that of individual matchings). However, the maximum recall gap between the ensemble and the individual is shown negligible compared with the precision performance gain. Besides, the more distance measures involved in the ensemble can result in a better precision performance but with maximum recall performance slightly compromised.

It can be observed that the DS-Ensemble algorithm can achieve higher precision and trade off more maximum recall performance. The DS-Ensemble can achieve almost 100% precision at low recall rates and more than 95% at high recall rates. As a result, the DS-Ensemble algorithm can be viewed as the most reliable user identification algorithm. However, the reliability of DS-Ensemble comes at the cost of the maximum recall performance, especially when users are similar to each other, as shown in Scenario 2. The proposed low-complexity DS-Ensemble P&M algorithm has a similar performance as the DS-Ensemble matching, as the size of most subgraphs after partitioning without loss is less than the threshold ($t = 1,000$ in Algorithm 3).

### C. USER GROUPING

The complexity of each distance measure between two users depends on the support of histogram or the number of points in two convex hulls. However, one needs to calculate $N \times M$ distances across two datasets so that a distance matrix $\boldsymbol{W}^g$ can be generated. The complexity of distance matrix generation, $\mathcal{O}(\psi NM)$, may lead to scalability issue when the user size of two datasets is tremendous. Here, $\psi$ denotes the computational complexity of distance measures per pair. Inspired by the graph partitioning concept employed in the DS-Ensemble




(a) Precision (b) Recall

**FIGURE 3.** Large-scale user identification analysis in Scenario 3.

P&M algorithm, we propose to first cluster users into small groups so that the distance matrix generation and the ensemble matching within each group could be conducted.

The mobility feature of user $i$ on each base station takes the form as follows,

$$\boldsymbol{f}_i = [f_{i1}, f_{i2}, \cdots, f_{iL}]^T \ . \tag{18}$$

In fact, $f_{il}$ characterizes the mobility behavior of user $i$ on location $l$, $f_{il} = \hat{\Pi}_{il} \times \log(N/n_l)$, where $n_l$ is the number of users visiting location $l$ out of total user number $N$. The term $\log(N/n_l)$ is similar to inverse document frequency in the field of document clustering [41], which is designed to depict the importance of location points for clustering. In other words, if most users visit one location, the value of $\log(N/n_l)$ will be small, meaning that such location is less important to distinguish users.

Hence, one can obtain a feature matrix by stacking the feature of all the users from both datasets as follows,

$$\boldsymbol{F} = [\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{(N+M)}] \in \mathcal{R}^{L\times(N+M)}, \tag{19}$$

where each column represents the mobility behavior of a user. Two characteristics of feature matrix $\boldsymbol{F}$ could be observed: 1) the number of base stations could be very large, up to 6, 500 in the studied dataset, due to a large geographic area studied; 2) the mobility feature of users $\boldsymbol{f}_i$ could be very sparse, as one user can most likely visit a small portion out of all the base stations. It is worth recalling that the objective of user grouping is to reduce the user set size for matching, whose complexity is $\mathcal{O}(NM)$. In other words, the computational complexity of the clustering algorithm cannot be as high as $\mathcal{O}(NM)$. Otherwise, direct user identification on the entire user set would be more meaningful. Besides, the high dimension of user mobility feature can lead to the uselessness of the commonly employed low-complexity $k$-means clustering algorithm.

As a result, the clustering algorithm based on non-negative matrix factorization (NMF) [41] is employed to cluster users in this work. The NMF is essentially to minimize the Frobenius norm of the difference between the original matrix and the multiplication of two non-negative factorized matrices as follows,

$$\begin{aligned} &\underset{\boldsymbol{P},\boldsymbol{Q}}{\text{minimize}} \ \|\boldsymbol{F} - \boldsymbol{P}\boldsymbol{Q}\|_{\mathcal{F}} \\ &\text{subject to } \boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_R] \in \mathcal{R}_+^{L\times R} \\ &\qquad\qquad \boldsymbol{Q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_{(M+N)}] \in \mathcal{R}_+^{R\times(N+M)}, \end{aligned} \tag{20}$$

where $R$ denotes the factorization rank and also the number of user groups. Based on the non-negativity of both matrices, each user can be represented by the non-negative weights $\boldsymbol{q}_i$ on group representations $\boldsymbol{P}$ as follows,

$$\boldsymbol{f}_i = \sum_{r=1}^{R} q_{ir}\boldsymbol{p}_r \ ,$$

where $q_{ir}$ denotes the weight on group $r$ of user $i$. As $q_{ir}$ is non-negative, the user group for each user could be determined by finding the maximum user group weight as follows,

$$r_i = \arg\max_r q_{ir} \ . \tag{21}$$

Therefore, one could obtain user grouping results with the complexity of $\mathcal{O}(R(N+M))$, once the feature matrix is factorized. In the literature, the multiplicative update method [41], [42] is commonly employed for NMF with the complexity of $\mathcal{O}(iR(N+M))$, where $i$ denotes the overall iteration. As a result, by combining with the DS-Ensemble P&M algorithm, the complexity of the entire user identification procedure could be significantly less than $\mathcal{O}(NM)$.

### D. PRIVACY EVALUATION

Figs. 3 shows the precision and recall performance of large-scale user identification analysis on a 150, 000 user set, based on the proposed DS-Ensemble P&M algorithm. In Figs. 3, the length of user data collection ranges from 2 to 7 days (x axis). For complexity reduction, the entire user set is first partitioned by user grouping, as discussed in Section V-C.

The user grouping would lead to the loss of recall performance, since some users may be clustered into different groups. The incorrect clustering rate ranges from 16.36% (7-day data collection) to 19.64% (2-day data collection). It can be observed that both the recall and precision performance can be improved as the data collection length grows, which suggests the reduction of data collection could lead to privacy protection to some degree. Overall, the subscriber privacy is vulnerable in terms of user identifiability across two datasets, if the dataset is released only with ID anonymization. In Figs. 3, it shows that the user privacy is still at high risk, as the proposed DS-Ensemble P&M algorithm still can recognize almost half of the user pairs at very high confidence (up to 90%).

## VI. CONCLUSIONS

In this paper, we studied the privacy attack in terms of user identifiability across two datasets based on the spatiotemporal data collected from mobile networks. By integrating multiple distance measures, a scalable ensemble matching framework was proposed to reduce false positives significantly. Taking advantage of the extreme sparsity of the vote matrix, the computational complexity of the proposed ensemble matching framework was an order of magnitude lower. In addition, user grouping was studied to further reduce the overall computational complexity so that large-scale user identification can be facilitated. Experiments demonstrated that our proposed multi-feature ensemble matching achieves superior performance (up to 100% precision), which also suggested the vulnerability of mobile network subscribers' privacy.

## APPENDIX A
## VISITING FREQUENCY ONLY (VFO) MODELING [10]

The location visiting frequency only (VFO) was proposed in [10], [14] to distinctly characterize a user.

### A. DATA MODELING

With the location point set (base station set) being abstracted as an alphabet set $\mathcal{A} = \{a_1, \cdots, a_L\}$, the raw mobility trace (1) can be first modeled as a string with length $T$ by discarding the time information,

$$X_i = x_{i1}, x_{i2}, \cdots, x_{iT}. \tag{22}$$

Every element $x_{it} \in \mathcal{A}$ in the string is assumed to be i.i.d. from the alphabet set $\mathcal{A}$ based on an unknown location visiting *probability mass function* $\Pi_i$.

### B. REPRESENTING FEATURE

Based on the i.i.d assumption of the string generation, the location visiting probability $\Pi_i$ could be estimated by the empirical probability distribution or histogram $\widehat{\Pi}_i$, i.e.,

$$\widehat{\Pi}_{i,l} = \frac{\mathcal{N}_i(a_l)}{T}, \quad a_l \in \mathcal{A}\,, \tag{23}$$

where $\mathcal{N}_i(a_l) = \sum_{x_{it}=a_l} 1$ denotes the number of appearance of letter $a_l$ in the string $X_i$, counting the number of visits of user $i$ at location $a_l$. Thus, the spatiotemporal behaviors of a user could be represented by the *histogram*, characterizing his/her *visiting frequency* over location point set $\mathcal{A}$.

### C. DISTANCE MEASURES

The intuitive yet heuristic $L_1$ distance function could be employed to assess the distance between two histograms as follows,

$$\Delta_{\text{VFO-L1}}(X_i, Y_j) = \frac{1}{2}\sum_{a_l \in \mathcal{A}} \left|\widehat{\Pi}_{i,l} - \widehat{\Pi}_{j,l}\right|\,. \tag{24}$$

Based on the multi-hypothesis test framework discussed in [14], to determine the optimal hypothesis using the likelihood test is equivalent to solving the kLAP with distance generated by the Jensen-Shannon divergence (JSD). Thus, the JSD could serve as a distance measure on the histograms as follows [10],

$$\Delta_{\text{VFO-JSD}}(X_i, Y_j) = \text{JSD}(\widehat{\Pi}_i, \widehat{\Pi}_j)\,. \tag{25}$$

where

$$\text{JSD}(p, q) = \text{KL}\,(p \,\|(p+q)/2) + \text{KL}\,(q \,\|(p+q)/2)\,. \tag{26}$$

## APPENDIX B
## VISITING FREQUENCY AND DURATION (VFD) MODELING

In the literature, the visiting frequency only (VFO) proposed in [10], [14] can effectively capture the traces generated by two users. However, the VFO captures one spatial aspect of the available mobility traces, while neglecting the potential temporal information valuable for user identification. Though the collected dataset may be an event log with users' spatiotemporal trajectory sporadically sampled, the temporal information could still be employed to characterize users. In this subsection, we propose a visiting frequency and duration (VFD) feature to jointly capture the distance in both the spatial and temporal aspects.

### A. DATA MODELING

Atop the string model (22) described in Appendix A, the raw spatiotemporal attribute (1) could be modeled as a tuple string with size $P_i$ as follows:

$$X_i = (x_{i1}, t_{i1}), (x_{i2}, t_{i2}), \cdots, (x_{iP_i}, t_{iP_i})\,, \tag{27}$$

where $x_{ip} \in \mathcal{A}$ denotes the $p$-th recorded location of user $i$, and $t_{ip}$ denotes the corresponding duration between the current event and the next one.

Based on the spatiotemporal tuple string modeling, we also assume that each tuple is i.i.d. generated by an unknown probability distribution, where the duration of a user at a given location $a_l \in \mathcal{A}$ is modeled as an exponential (EXP) distribution conditioned upon the location $a_l$,

$$f(t|a_l; \lambda_{i,l}) = \lambda_{i,l}\exp(-\lambda_{i,l}t), \quad t > 0, \tag{28}$$

where $\lambda_{i,l}$ denotes the reciprocal of the average duration of user $i$ at location point $a_l$.

### B. REPRESENTING FEATURE

Assume that duration generated at locations are uncorrelated, the likelihood of $X_i$ takes the form

$$\mathcal{L}(X_i) = \prod_l \mathcal{L}(X_i; a_l)\Pi_i(a_l) \tag{29}$$

where $\mathcal{L}(X_i; a_l)$ denotes the likelihood of $X_i$ observed at location $a_l$ as in (28).

As a result, one can obtain two representations to characterize users in both the spatial and temporal aspects. The spatial representation is the visiting frequency by the empirical probability distribution $\widehat{\Pi}_i$ calculated via (23). The temporal representation could be obtained by the location-dependent exponential distribution parameter set $\widehat{\Lambda}_i = \{\widehat{\lambda}_{i,l}\}$, where each element $\lambda_{i,l}$ can be estimated at each $a_l \in \mathcal{A}$,

$$\hat{\lambda}_{i,l} = N_l(a_l) / \sum_{x_{ip}=a_l} t_{ip}. \tag{30}$$

### C. DISTANCE MEASURES

With the similar multi-hypothesis test framework in [14], a distance measure between two users in terms of both $\widehat{\Pi}_i$ and $\widehat{\Lambda}_i$ can be derived. With respect to the likelihood function (29), the derived distance measure can be decomposed into two components, namely visited frequency only (VFO) and visited duration only (VDO), as follows,

$$\Delta_{\text{VFD-WD}}(X_i, Y_j) = \Delta_{\text{VFO-WD}}(X_i, Y_j) + \Delta_{\text{VDO-WD}}(X_i, Y_j) \tag{31}$$

Here, the "WD" is short for weighted divergence, which is a generalization of Jensen-Shannon divergence. The $\Delta_{\text{VFO-WD}}$ is originated from (26) with weighted divergence employed, while the $\Delta_{\text{VDO-WD}}$ is obtained based on the divergence between two exponential distributions on their corresponding visited durations as follows,

$$\Delta_{\text{VDO-WD}}(X_i, Y_j) = \sum_{a_l \in \mathcal{A}} \Big[ q_i \widehat{\Pi}_{i,l} \text{KL}\left(\hat{\lambda}_{i,l} \| \hat{\lambda}_{ij,l}\right) + q_j \widehat{\Pi}_{j,l} \text{KL}\left(\hat{\lambda}_{j,l} \| \hat{\lambda}_{ij,l}\right) \Big] \tag{32}$$

where $\text{KL}(\lambda_1 \| \lambda_2)$ denotes the KL divergence on two EXP distributions, i.e., $\text{KL}(\lambda_1 \| \lambda_2) = \log(\lambda_1/\lambda_2) + (\lambda_2/\lambda_1) - 1$. And $\hat{\lambda}_{ij,l}$ is the weighted harmonic average over $\hat{\lambda}_{i,l}$ and $\hat{\lambda}_{j,l}$. Assume that the string length of each user and the number of each observation are the same, the JSD could be easily obtained as well as the L1 distance,

$$\Delta_{\text{VFD-L1}}(X_i, Y_j) = \sum_{a_l \in \mathcal{A}} \left| \frac{\widehat{\Pi}_{i,l}}{\hat{\lambda}_{i,l}} - \frac{\widehat{\Pi}_{j,l}}{\hat{\lambda}_{j,l}} \right|. \tag{33}$$

## APPENDIX C
## DAILY HABITAT REGION (DHR) MODELING

The previously discussed spatiotemporal features abstract discrete location points as independent and unrelated letters in an alphabet set $\mathcal{A}$. Such modeling discards the critical geospatial information. The geospatial information may help combat the information loss due to the sporadic sampling of users' spatiotemporal trajectories. Thus, a heuristic spatiotemporal feature is employed for user identification [39], *daily habitat regions* (DHR), as well as its corresponding distance measures, based on the geospatial information in this subsection. The daily habitat regions capture the daily spatial coverage of a subscriber, which are expected to be consistent to some degree and may serve as the subscriber's mobility fingerprints.

### A. DATA MODELING

The spatiotemporal attribute (1) is first formulated into sets of location points:

$$X_i = \{\mathcal{X}_{i1}, \mathcal{X}_{i2}, \cdots, \mathcal{X}_{iQ_X}\}, \tag{34}$$

where each set $\mathcal{X}_{iq} \subseteq \mathcal{A}$ denotes a set of location points that the user visits during a calendar date $q$ and $Q_X$ and $Q_Y$ denote the number of days collected in dataset $\mathcal{X}$ and $\mathcal{Y}$, respectively.

### B. REPRESENTING FEATURE

Here, we employ a classical computational geometry concept, convex hull, to approximate the spatial coverage that a user visits daily. By approximating a small region of geo-surface as an Euclidean space, the convex hull of a given point set $\mathcal{X}_{iq}$ in a 2-dimensional surface is defined as the set of the convex combination of the given finite point set. Thus, the daily convex hull, $C_{iq}$, is employed to represent the spatiotemporal behaviors of a user for a given day. Hence, the mobility traces of user $i$ is represented as a set of daily convex hulls,

$$\mathcal{C}_i = \left\{C_{i1}, C_{i2}, \cdots, C_{iQ_i}\right\}, \tag{35}$$

where each convex hull is again assumed to be i.i.d. generated from an unknown probability distribution.

### C. DISTANCE MEASURES

With the convex hull set representing users' spatiotemporal behaviors, we first define two distance measures between two convex hulls based on the cosine distance and the intersection-over-union (IoU), respectively,

$$\begin{aligned} \delta_{\cos}(C_p, C_q) &= 1 - \frac{\text{area}(C_p \wedge C_q)}{\sqrt{\text{area}(C_p) \times \text{area}(C_q)}}, \\ \delta_{\text{iou}}(C_p, C_q) &= 1 - \frac{\text{area}(C_p \wedge C_q)}{\text{area}(C_p \vee C_q)}, \end{aligned} \tag{36}$$

where $C_p \wedge C_q$ and $C_p \vee C_q$ denote the intersection and union of the two convex hulls, respectively, and the operator area$(\cdot)$ is to calculate the area of a polygon. Therefore, a distance measure between two convex hull sets is proposed based on (36) to evaluate the similarity of two subscribers as follows,

$$\begin{aligned} \Delta_{\text{DHR-COS}}(X_i, Y_j) &= \frac{1}{Q_i \times Q_j} \sum_{C_{ip} \in X_i} \sum_{C_{iq} \in Y_j} \delta_{\cos}(C_{ip}, C_{iq}), \\ \Delta_{\text{DHR-IOU}}(X_i, Y_j) &= \frac{1}{Q_i \times Q_j} \sum_{C_{ip} \in X_i} \sum_{C_{iq} \in Y_j} \delta_{\text{iou}}(C_{ip}, C_{iq}). \end{aligned} \tag{37}$$

Intuitively, the distance measure between two convex hull sets is to calculate the average distance between any two convex hulls in two respective sets. When the convex hull cannot be obtained because the number of distinct visited location points within a day is less than 3, the daily habitat region would be omitted. If no convex hull could be generated, the user will be labeled as non-identifiable.

## REFERENCES

[1] X. Cheng, L. Fang, X. Hong, and L. Yang, "Exploiting mobile big data: Sources, features, and applications," *IEEE Netw.*, vol. 31, no. 1, pp. 72–79, Jan. 2017.
[2] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 124–161, 1st Quart., 2016.
[3] X. Cheng, L. Fang, and L. Yang, "Mobile big data based network intelligence," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4365–4379, Dec. 2018.
[4] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.
[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.
[6] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *Proc. 7th ACM Workshop Privacy Electron. Soc.*, Alexandria, VI, USA, 2008, pp. 23–32.
[7] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, p. 1376, Mar. 2013.
[8] A. Cecaj, M. Mamei, and N. Bicocchi, "Re-identification of anonymized CDR datasets using social network data," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PERCOM WORKSHOPS)*, Budapest, Hungary, Mar. 2014, pp. 237–242.
[9] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with GLOVE," in *Proc. 11th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Heidelberg, Germany, Dec. 2015, p. 26.
[10] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016.
[11] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Atlanta, GA, USA, May 2017, pp. 1–9.
[12] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surveys*, vol. 42, no. 4, pp. 1–53, Jun. 2010.
[13] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, Montreal, QC, Canada, Apr. 2016, pp. 707–719.
[14] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 452–468, Jan. 2015.
[15] R. Jonker and T. Volgenant, "Improving the Hungarian assignment algorithm," *Oper. Res. Lett.*, vol. 5, no. 4, pp. 171–175, Oct. 1986.
[16] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, Dec. 1987.
[17] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.
[18] J. Dean and S. Ghemawat, "MapReduce: A flexible data processing tool," *Commun. ACM*, vol. 53, no. 1, pp. 72–77, Jan. 2010.
[19] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.
[20] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, Las Vegas, NV, USA, Sep. 2011, pp. 145–156.
[21] Y. Song, D. Dahlmeier, and S. Bressan, "Not so unique in the crowd: A simple and effective algorithm for anonymizing location data," in *Proc. 37th Annu. Int. ACM SIGIR Conf.*, Gold Coast, QLD, Australia, Jul. 2014, pp. 19–24.
[22] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Limits of location privacy under anonymization and obfuscation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 764–768.
[23] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining Knowl. Discovery*, vol. 32, no. 3, pp. 787–829, May 2018.
[24] Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin, "A new privacy breach: User trajectory recovery from aggregated mobility data," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1446–1459, Jun. 2018.
[25] I. Liccardi, A. Abdul-Rahman, and M. Chen, "I know where you live: Inferring details of People's lives by visualizing publicly shared location data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, San Jose, CA, USA, May 2016, pp. 1–12.
[26] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Oakland, CA, USA, May 2008, pp. 111–125.
[27] L. Rossi, J. Walker, and M. Musolesi, "Spatio-temporal techniques for user identification by means of GPS mobility data," *EPJ Data Sci.*, vol. 4, no. 1, pp. 1–16, Aug. 2015.
[28] W. Cao, Z. Wu, D. Wang, J. Li, and H. Wu, "Automatic user identification method across heterogeneous mobility data sources," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, Helsinki, Finland, May 2016, pp. 978–989.
[29] R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale, "A data mining approach to assess privacy risk in human mobility data," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 1–27, Feb. 2018.
[30] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, 2018, pp. 1–15.
[31] S. Chang, C. Li, H. Zhu, T. Lu, and Q. Li, "Revealing privacy vulnerabilities of anonymous trajectories," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12061–12071, Dec. 2018.
[32] A. Pyrgelis, N. Kourtellis, I. Leontiadis, J. Serra, and C. Soriente, "There goes wally: Anonymously sharing your location gives you away," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 1218–1227.
[33] D. Kondor, B. Hashemian, Y.-A. de Montjoye, and C. Ratti, "Towards matching user mobility traces in large-scale datasets," *IEEE Trans. Big Data*, early access, Sep. 24, 2018, doi: 10.1109/TBDATA.2018.2871693.
[34] F. Xu, Z. Tu, H. Huang, S. Chang, F. Sun, D. Guo, and Y. Li, "No more than what i post: Preventing linkage attacks on check-in services," in *Proc. World Wide Web Conf. (WWW)*, San Francisco, CA, USA, May 2019, pp. 3405–3412.
[35] R. Sedgewick and K. Wayne, *Algorithms*, 4th Ed. Reading, MA, USA: Addison-Wesley, 2011.
[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[37] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Bipartite graph partitioning and data clustering," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, Atlanta, GE, USA, Oct. 2001, pp. 25–32.
[38] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 2012.
[39] L. Fang, X. Cheng, L. Yang, and H. Wang, "Location privacy in mobile big data: User identifiability via habitat region representation," *J. Commun. Inf. Netw.*, vol. 3, no. 3, pp. 31–38, Sep. 2018.
[40] L. Fang, "Data mining and spatiotemporal analysis of modern mobile data," Ph.D. dissertation, Colorado State Univ., Collins, CO, USA, 2019.
[41] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Informaion Retr. (SIGIR)*, Toronto, ON, Canada, Jul. 2003, pp. 267–273.
[42] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257. Boca Raton, FL, USA: CRC Press, 2014.

**LUOYANG FANG** received the Ph.D. degree from the Department of Electrical and Computer Engineering, Colorado State University, in 2019. He is currently a Software Engineer with Google Inc. His research interests include big data, mobile data, location privacy, data mining, distributed systems, and network-aware job scheduling.

**HAONAN WANG** received the Ph.D. degree in statistics from the University of North Carolina, Chapel Hill, NC, USA, in 2003. He is currently a Professor of statistics with Colorado State University, Fort Collins, CO, USA. His research interests include object-oriented data analysis, statistical analysis on tree-structured objects, functional dynamic modeling of neuron activities, and spatio-temporal modeling.

**XIANG CHENG** (Senior Member, IEEE) received the Ph.D. degree from Heriot-Watt University, Edinburgh, U.K., and the University of Edinburgh, Edinburgh, U.K., in 2009.

He is currently a Professor with Peking University. His general research interests include channel modeling, wireless communications, and data analytics, subject on which he has published more than 200 journals and conference papers, six books, and holds ten patents.

Prof. Cheng was a recipient of the IEEE Asia Pacific Outstanding Young Researcher Award, in 2015; and a co-recipient of the 2016 IEEE JSAC Best Paper Award: Leonard G. Abraham Prize, the NSFC Outstanding Young Investigator Award, and the First-Rank and Second-Rank Award in Natural Science, Ministry of Education, in China. He has also received the Best Paper Awards at the IEEE ITST'12, ICCC'13, ITSC'14, ICC'16, ICNC'17, GLOBECOM'18, ICCS'18, and ICC'19. He has served as the symposium leading chair, the co-chair, and a member of the Technical Program Committee for several international conferences. He is currently an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and the *Journal of Communications and Information Networks. He is an IEEE Distinguished Lecturer.*

**LIUQING YANG** (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Minnesota, Minneapolis, in 2004.

She is currently a Professor with Colorado State University. Her general interests include communications and networking—subjects on which she has published more than 330 journal articles and conference papers, four book chapters and five books.

Dr. Yang became an IEEE Fellow in 2015. She was a recipient of the ONR Young Investigator Program (YIP) Award, in 2007, and the NSF Faculty Early Career Development (CAREER) Award, in 2009, the Best Paper Award at IEEE ICUWB'06, ICCC'13, ITSC'14, GLOBECOM'14, ICC'16, WCSP'16, GLOBECOM'18, ICCS'18 and ICC'19. She is the Editor in Chief of *IET Communications*, Executive Editorial Committee (EEC) Member of the IEEE TRANSACTIONS ON COMMUNICATIONS, and Senior Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She has also served as editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE *Intelligent Systems*, and *PHYCOM: Physical Communication, and as a Program Chair, Track/Symposium or a TPC Chair for many conferences.*

**SHUGUANG CUI** (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, California, USA, in 2005. Afterwards, he has been working as an Assistant, Associate, Full, Chair Professor of electrical and computer engineering with the University of Arizona, Texas A&M University, UC Davis, and CUHK with Shenzhen, respectively. He has also been the Executive Vice Director with the Shenzhen Research Institute of Big Data. His current research interests include data driven large-scale system control and resource management, large data set analysis, the IoT system design, energy harvesting-based communication system design, and cognitive network optimization. He is a member of the Steering Committee for IEEE TRANSACTIONS ON BIG DATA and the Chair of the Steering Committee for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He was also a member of the IEEE ComSoc Emerging Technology Committee. He was an IEEE ComSoc Distinguished Lecturer, in 2014 and IEEE VT Society Distinguished Lecturer, in 2019. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the Worlds' Most Influential Scientific Minds by ScienceWatch, in 2014. He was a recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He has served as the General Co-Chair and TPC co-chairs for many IEEE conferences. He has also been serving as the Area Editor for the *IEEE Signal Processing Magazine*, and an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JSAC SERIES ON GREEN COMMUNICATIONS AND NETWORKING, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has been the elected member of the IEEE Signal Processing Society SPCOM Technical Committee, from 2009 to 2014 and the elected Chair of the IEEE ComSoc Wireless Technical Committee, from 2017 to 2018.