Discovery of contrast corridors from trajectory data in heterogeneous dynamic cellular networks

Li Li
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
lli10@student.unimelb.edu.au

Sarah Erfani
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
sarah.erfani@unimelb.edu.au

Chien Aun Chan
Department of Electrical and Electronic Engineering
The University of Melbourne
Melbourne, Australia
chienac@unimelb.edu.au

Christopher Leckie
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
caleckie@unimelb.edu.au

***Abstract*—In 5G networks, the deployment and management of mobile edge computing infrastructure is a major challenge for mobile operators. Most recent work focused on extracting static movement patterns of mobile users from the trajectories generated during a specific time period to help with the management and orchestration of network resources. However, movement patterns of mobile users are not static over time. Understanding significant differences in mobile users' movement during different time periods can provide insights for mobile operators to dynamically reconfigure the network in response to the changes in traffic flows by time of day. Therefore, in this paper, we propose a framework based on contrast data mining to identify significantly different movement patterns, which we model as corridors, during different time periods. To measure the difference, an improved distance measure based on a modified Hausdorff distance and Earth Movers' distance is proposed to calculate the dissimilarity between the identified corridors, which considers the spatial heterogeneity of mobile networks. To further extract the significantly different corridors, we formulate the definition of contrast corridors of mobile users' movement. Experimental results on synthetic datasets as well as real-life datasets collected by China Mobile show that our method can effectively and robustly detect contrast corridors from trajectories generated from different time periods in mobile networks by improving the F1 score by 20% on average.**

***Index Terms*—contrast data mining, corridor identification, mobile networks.**

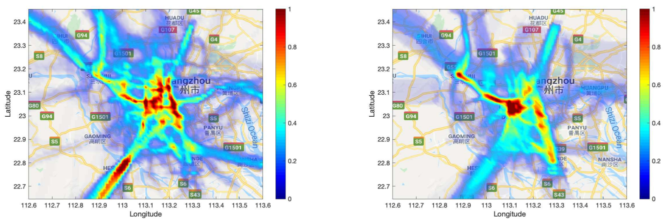
I. Introduction

With the rapid development of wireless networks (from 2G/3G/4G to 5G) in recent years, internet-connected mobile devices are penetrating every aspect of life, work and entertainment. A major challenge in the management of mobile networks is how to provide high bandwidth coverage to large numbers of mobile users. In particular, understanding and discovering significant changes in the movement patterns of users in mobile networks can help service providers in the deployment of networks and base stations, and the management of network resources. For example, the ability to identify changes in users' movement patterns can help the orchestration of 5G network resources through network function virtualization [1] and network slicing [2].

In a cellular network, each cell tower (base station) covers a small geographical area. If phone users access the Internet, their positions can be passively detected by the cell towers that provide Internet data to them. Thus, a mobile phone user's trajectory can be represented as a sequence of cell tower IDs with corresponding timestamps. This trajectory acquisition technique provides a good way for mobile operators to understand mobile users' movement patterns. Some studies have focused on identifying the underlying geographical corridors of users, which can be treated as pathways that are frequently traversed by a considerable number of mobile users [3], [4], [5]. However, most studies tried to find the pathways based on the data during one specific time period and treated the network as temporally homogeneous in their analyses.

Therefore, in our work, we focus on the problem of identifying what are those significant changing corridors, which we model as contrast patterns that can be used to identify targets for network reconfiguration. We consider two challenges from real-life data. The first one is that mobile trajectories are coarse and their granularity varies due to the non-uniform spatial distribution of cell towers [6], [7]. Thus, it is necessary to propose a distance measure that can deal with the heterogeneous spatial scales when measuring the dissimilarity between trajectories. The second challenge is that identifying static corridors plays an important role in managing networks for the long term design of network, but with the introduction of new generations of cellular network technology, such as 5G, there is a great opportunity for dynamically reconfiguring the network in response to changes in traffic flows by time of day. For example, users' movement patterns might be different in the morning to the patterns in the afternoon, as shown in Fig. 1.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE



(a) morning

(b) afternoon

Fig. 1: Heat maps of mobile phone users' traffic density in mornings and afternoons in a city of China. (The results are obtained based on the methods introduced in [4])

In this paper, we propose to use contrast data mining on trajectories generated in mobile networks to analyze the changes in phone users' movement patterns during different time periods. To the best of our knowledge, this is the first paper that focuses on the temporal changes of human trajectories generated from heterogeneous mobile data networks. Our main contributions are: (1) we propose a modified Hausdorff distance to measure the dissimilarity between corridors in heterogeneous mobile networks; (2) we propose a contrast corridor mining algorithm based on Earth Movers' Distance to detect the differences/changes in movement patterns during different time periods; (3) we conduct experiments over the real-life data of mobile users in a southern province in China as well as a synthetic data to evaluate our approach, which shows improvements in both interpretability and detection ability.

The remainder of the paper is organized as follows. Section II provides a review of the related work. Section III introduces some definitions and formulates the problem we focus on. In Section IV, we present the main methodology we propose for mining contrast corridors, which includes the distance measure for corridors, the distance measure for subtrajectories and contrast corridor mining. The details of our experiments on synthetic data as well as real life data and a discussion of the results of our methods are shown in Section V. Section VI concludes our work and proposes some future challenges.

II. Related Work

In this section, we briefly review the most related work, i.e., corridor identification and contrast mining on trajectory data.

A. Corridor Identification

The problem of corridor/pathway identification has been widely studied in the literature, mostly using subtrajectory clustering [8] to address the challenges. For example, Lee et al. [9] proposed a partition-and-group framework for clustering trajectories, TRACLUS, which enables the discovery of common sub-trajectories, based on a trajectory partitioning algorithm that uses the MDL (Minimum Description Length) principle. A three-phase approach was proposed in [3] to discover trajectory corridors (i.e., frequently followed paths) using the discrete Fréchet distance. Trajectories are segmented into subtrajectories (short polygonal curves, not line segments) using meshing-grids, and then the sub-trajectories in each grid cell are clustered separately using hierarchical clustering. In [10], a trajectory clustering method based on motifs (frequently occurring substrings) was proposed. Trajectories are simplified first and partitioned according to some predefined motion patterns, such as wide left turn and short left turn. Then the algorithm computes motifs and maps subtrajectories corresponding to motifs into some feature space. Finally, DBSCAN (Density-Based Spatial Clustering of applications with Noise) is applied and representative trajectories are obtained using the method mentioned in [9]. Recently, Zygouras et al. [5] proposed a method for detecting a set of corridors from GPS trajectories using the MDL principle. However, most studies focus on the analysis of GPS data. For the trajectories in mobile networks, in [11], the authors proposed a method based on the Apriori algorithm to find frequent hotspot sequences. In our previous work [4], a large-scale mobile network dataset was studied and we found that mobile network data differs from GPS data due to two inherent properties. Specifically, the temporal resolution of mobile network data is normally much lower than GPS data, and the spatial resolution of mobile network data can vary from several hundred meters to a few meters according to the cell tower density of different areas. Therefore, we proposed a multi-scale trajectory clustering algorithm for corridor identification in heterogeneous mobile networks. The experimental results demonstrated that the proposed method can better deal with mobile network data than other methods proposed for GPS data. However, in our previous work the temporal heterogeneity in users' movement patterns was not considered.

B. Contrast Mining

Contrast patterns are often defined as patterns whose supports differ significantly among the datasets that are under contrast [12]. These contrast patterns can describe discriminative behavior between classes or emerging trends between datasets with respect to a property of interest by means of an understandable representation [13], [14]. In the problem of contrast mining on trajectory data, the aim is to find discriminative patterns (e.g., sequences, graphs, matrices, tensors) that occur frequently in one dataset and infrequently in another. For example, in the work of Wang et al. [15], a framework for discovering the impact of road closures on traffic flows was proposed. By computing the growth rate of traffic flows on n-Edgesets, the emerging n-Edgesets were selected by using the LOF (Local Outlier Factor).

In our work, we focus on finding the discriminative corridors, which are represented as directed weighted graphs, between two trajectory data sets generated in different time periods.

III. Overview of Problem

In this section, we introduce some definitions and formulate our problem.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

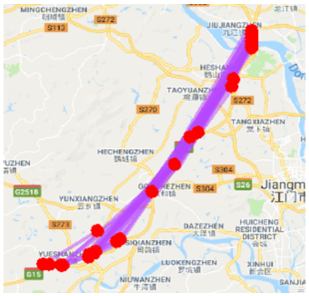


Fig. 2: Illustrative example of a corridor.

A. Preliminaries

Suppose that in a mobile network, there are N cellular towers, which are represented as $C = \{c_1, c_2, ..., c_N\}$, and each cell tower has a unique identifier and has known coordinates.

Trajectory A trajectory can be represented as a sequence of states in a given period of time: $Traj = \{s_1, s_2, ..., s_l\}$, where l is the number of cells visited by the user. A state is defined as $s = (c, t, stay)$, where c is the cell ID, t is the time when the user entered the current cell, and $stay$ is the stay time in the current cell.

Tracklet A tracklet, T, is a directed fragment of a trajectory $Traj$, i.e., $T = \{s_a, ..., s_b\}$, where $1 \leq a < b \leq l$. Tracklets can be extracted from trajectories by using the method proposed in [4], which considers movement direction and stay time $stay$.

Corridor A corridor can be treated as a pathway that is frequently traversed by a considerable number of mobile users [4], which is a cluster of similar tracklets. It can be represented as a graph, denoted as $cor = \langle V, E \rangle$, where V is the set of cells in the corridor, E is the set of all the edges in the graph and the weight of each edge represents the traffic load between the corresponding two nodes. Fig. 4a shows an example of one corridor, where small circles here represent nodes/cell towers, and the line width of each edge represents the weight of the corresponding edge.

B. Problem Statement

In this paper, we focus on characterizing the major differences between these corridors in different time periods. Specifically, given the historical trajectories of M mobile users during two different time periods, i.e., the positive trajectory data set $TRAJ^+ = \{Traj_1^+, Traj_2^+, \ldots, Traj_M^+\}$ and the negative trajectory data set $TRAJ^- = \{Traj_1^-, Traj_2^-, \ldots, Traj_M^-\}$, and the identified corridor sets $COR^+ = \{cor_1^+, cor_2^+, \ldots\}$ and $COR^- = \{cor_1^-, cor_2^-, \ldots\}$, our research question is how to detect the significantly different corridors, called *contrast corridors*, between COR^+ and COR^-. In order to answer this question, we study the following two sub-problems:

(1) How to measure the dissimilarity between two corridors, i.e., how can we calculate the distance between corridors?

(2) How to define and mine contrast corridors?

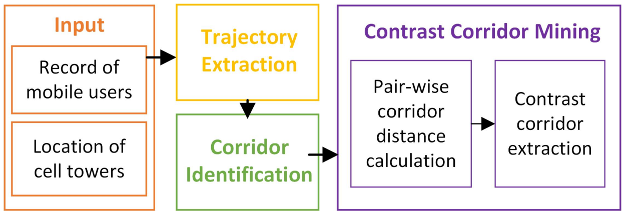


Fig. 3: Framework of our proposed method.

To address the first question, we should consider the heterogeneous distribution of corridors in order to distinguish corridors located in dense areas (e.g., central areas) and avoid missing similar corridors in sparse areas, such as corridors located in rural areas. The challenge of the second sub-problem is how to measure the change of support in this specific problem, because it differs from general contrast mining problems, which are mainly focused on transactional data.

C. Framework

The framework of our proposed method is illustrated in Fig. 3. The inputs are the record of mobile users and the location of cell towers (longitude and latitude). Each record includes user ID, cell ID and the corresponding timestamp. There are three main steps to identify the contrast corridors in different time periods. The first step is trajectory extraction. In the first step, data are preprocessed by data aggregation, data cleaning and oscillation resolution [16]. Then the trajectory of each user is extracted according to our definition in Section III-A. In the second step, corridors are identified by using the method proposed in [4]. The final step is to identify significantly different corridors by using contrast mining. In particular, the pair-wise distances between corridors are calculated and contrast corridors are formally defined and extracted.

IV. Methodology for Contrast Corridor Mining

In the following subsections, we describe the details about our methodology for contrast corridor mining.

A. Distance Measure for Corridors

In order to measure the distance between two corridors, we propose an algorithm that is based on EMD (Earth Mover's Distance) [17]. The EMD is a method that is applied to evaluate the dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, which we call the ground distance, is given. It is proportional to the minimum amount of *work* required to change one distribution into the other, where a unit of work is defined as the amount of work needed to move a unit of weight by a unit of *ground distance*.

Suppose there are two corridors $cor_p = G_p \langle V_p, E_p \rangle$ and $cor_q = G_q \langle V_q, E_q \rangle$, where V denotes the vertices and E represents the edges in the corridors. For each non-zero edge in a corridor, the weight of it, w, is defined as the weight of the edge, which indicates the traffic volume between two cells. Each corridor can be treated as a distribution, which has a set of edges with their corresponding weights, i.e., a cor can be represented as $\{(e_1, w_{e_1}), (e_2, w_{e_2}), ...\}$, where e is an edge in the graph and w_e is the weight of edge e. $\{(e_1, w_{e_1}), (e_2, w_{e_2}), ...\}$ is also called the *signature* of the distribution.

Then the ground distance matrix between two corridors is defined as $D = [d_{kl}]$ $(k = 1, ..., m, l = 1, ..., n)$, where m and n are the number of edges in corridor p and q respectively, and d_{kl} is the ground distance between edge $e_k \in E_p$ and edge $e_l \in E_q$. Considering the non-uniform distribution of cell towers, we propose to calculate the ground distance between two edges based on a modified version of the distance measure proposed in [4], which is described in Section IV-B. Then we can formulate the problem as a linear programming problem.

Given the corresponding signatures of two corridors and the ground distance matrix, our aim is to find a flow $F = [f_{kl}]$, where f_{kl} is the flow between edges e_k and e_l, that minimizes the overall cost defined as:

$$WORK(p, q, \mathbf{F}) = \sum_{k=1}^{m} \sum_{l=1}^{n} f_{kl} d_{kl}, \qquad (1)$$

subject to the following constraints:

1) $f_{kl} \geq 0$;
2) $\sum\limits_{l=1}^{n} f_{kl} \leq w_{e_k}$, $1 \leq k \leq m$;
3) $\sum\limits_{k=1}^{m} f_{kl} \leq w_{e_l}$, $1 \leq l \leq n$;
4) $\sum\limits_{k=1}^{m} \sum\limits_{l=1}^{n} f_{kl} = min(\sum\limits_{k=1}^{m} w_{e_k}, \sum\limits_{l=1}^{n} w_{e_l})$.

The first constraint means that only the weight from cor_p can be moved to cor_q. The second constraint ensures that the total weight moved from an edge in cor_p to cor_q should be no more than its own weight, and the third constraint ensures that the total weight moved to any edges in cor_q should be no more than their own weight. The last constraint requires that the total weight moved should be equal to the total weight of the lighter corridor.

Then the EMD is defined as the work normalized by the total flow:

$$EMD(cor_p, cor_q) = \frac{\sum\limits_{k=1}^{m} \sum\limits_{l=1}^{n} f_{kl} d_{kl}}{\sum\limits_{k=1}^{m} \sum\limits_{l=1}^{n} f_{kl}}. \qquad (2)$$

B. Multi-scale Hausdorff Distance

In order to calculate the ground distance between two edges, in this paper we adopt the distance measure proposed in [4] with additional modifications by considering the direction. Here we brief describe it and then introduce our modification on it.

Hausdorff distance is the maximum distance of a set to the nearest point in the other set, which represents the maximum mismatch level between two point sets [18]. We modify it from two aspects. First, instead of using the distance between points, the distance between a point and a line segment is adopted to measure the distance between a point in one set to the other set. Second, to overcome the problem caused by cell heterogeneity, the distance is normalized by a scaling factor, which is calculated based on the cell density. The definition is given as follows.

Modified Hausdorff Distance Given two tracklets $T_1 = \{c_1^1, c_2^1, ..., c_m^1\}$ and $T_2 = \{c_1^2, c_2^2, ..., c_n^2\}$, the MHD $dist(T_1, T_2)$ is defined as:

$$dist(T_1, T_2) = max(\Delta(T_1, T_2), \Delta(T_2, T_1)), \qquad (3)$$

$$\Delta(T_i, T_j) = max(d^{\rho}_{c_1^i, T_j}, d^{\rho}_{c_2^i, T_j}, \dots, d^{\rho}_{c_m^i, T_j}). \qquad (4)$$

In Equation 4, $\Delta(T_i, T_j)$ is the distance between two tracklets. $d^{\rho}_{c_i, T_j}$ is the distance between a point c_i and a tracklet T_j, which can be calculated as:

$$d^{\rho}_{c_i, T_j} = \alpha_{c_i} \cdot d_{c_i, T_j}, \qquad (5)$$

where α is the normalization factor and $d_{c,T}$ is the distance between cell c and tracklet T, i.e., the shortest distance from the cell center to all line segments in tracklet T. The equation for calculating α is given as:

$$\alpha_{c_i} = \frac{N \cdot \rho(c_i)}{\sum_{k=1}^{N} \rho(c_k)}, \qquad (6)$$

where $\rho(c_i)$ is the density of the cell center of c_i. The density contributed by a cell is assumed to be a Gaussian distribution. The mean vector is the center of the cell tower, and that three times the standard deviation is equal to the coverage radius of the cell. Therefore, the accumulated density at each cell in the network is considered as the density of the cell in a network, which is:

$$\rho(c_i) = \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_j} exp(-\frac{d^2_{c_i, c_j}}{2\sigma_j^2}), \qquad (7)$$

where d_{c_i, c_j} is the spherical distance between centers of two cells c_i and c_j, and $\sigma_j = r_j/3$, r_j is the coverage radius of c_j.

This measurement can be used to find the closeness of two sets of points. However, it ignores the direction information of sequences. Therefore, we propose to add another factor that considers the direction of movement. Here we use the accumulated direction of all the segments in one tracklet as the direction of movement of the tracklet. Given a tracklet $T_i = \{c_1, c_2, ..., c_m\}$, the direction of movement of it can be calculated as:

$$dir_{T_i} = \sum_{i=1}^{m-1} dir(s_i, s_{i+1}), \qquad (8)$$

where $dir(\cdot, \cdot)$ is the movement direction of two consecutive states. Then the direction difference between two tracklets T_i and T_j is:

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

```
Algorithm 1: $\Delta(T_1, T_2)$
Input: Two tracklets $T_1$ and $T_2$
Output: $\Delta(T_1, T_2)$
$\Delta(T_1, T_2) = 0$ ;
for each $c_i^1$ in $T_1$ do
    $d_{c_i^1 T_2} = \infty$ ;
    for each line segment $c_j^2 c_{j+1}^2$ in $T_2$ do
        calculate the distance $d_{c_i^1 (c_j^2 c_{j+1}^2)})$ between a
         point $c_i^1$ and a line segment $c_j^2 c_{j+1}^2$ ;
        $d_{c_i^1 T_2} = min(d_{c_i^1 T_2}, d_{c_i^1 (c_j^2 c_{j+1}^2)})$ ;
    end
    calculate $\alpha_{c_i^1}$ using Equation 6 and 7 ;
    $\Delta(T_1, T_2) = max(\Delta(T_1, T_2), \alpha_{c_i^1} \cdot d_{c_i^1 T_2})$ ;
end
calculate $\beta$ using Equation 8, 9, 10 ;
$\Delta(T_1, T_2) = \beta\Delta(T_1, T_2)$ .
```

$$dir(T_i, T_j) = \frac{dir_{T_i} \cdot dir_{T_j}}{|dir_{T_i}| \cdot |dir_{T_j}|}. \tag{9}$$

The direction distance will be used to obtain the second normalization factor β, which can be calculated as:

$$\beta = \begin{cases} 1/dir(T_i, T_j), & dir(T_i, T_j) > 0 \\ \infty. & dir(T_i, T_j) \leq 0 \end{cases} \tag{10}$$

Then the distance between two tracklets after normalized is:

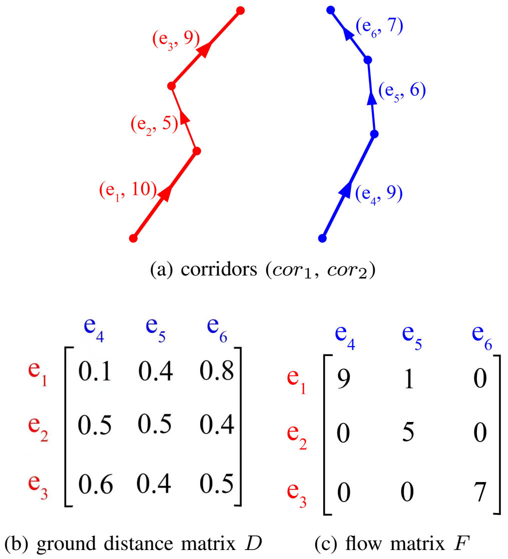
$$\Delta(T_1, T_2) = \beta \cdot \Delta(T_1, T_2). \tag{11}$$

The pseudocode is provided in Algorithm 1. The time complexity of the Modified Hausdorff distance is the same as the Hausdorff distance, which is $O(n^2)$, where n is the average number of sample points in the trajectories under comparison.

Here we provide an example to illustrate how to calculate the distance between two corridors. As shown in Fig. 4a, we have two corridors, which have been represented as two graphs. The signatures of these two corridors are $cor_1 = \{(e_1, 10), (e_2, 5), (e_3, 9)\}$ and $cor_2 = \{(e_4, 9), (e_5, 6), (e_6, 7)\}$ respectively. The ground distance matrix between these 6 edges can also be calculated by using the distance measure described in Section IV-B. Suppose that the ground distance matrix is equal to the matrix shown in Fig. 4b, by solving the linear programming problem we can obtain the flow matrix as shown in Fig. 4c, and the final EMD between these two corridors is 0.2636.

C. Contrast Corridor Mining

Given two different datasets, two sets of corridors $P = \{cor_{p_1}, ..., cor_{p_i}, ..., cor_{p_{N_1}}\}$ and $Q = \{cor_{q_1}, ..., cor_{q_j}, ..., cor_{q_{N_2}}\}$ can be found respectively using our proposed corridor identification method, where N_1 and N_2 are the numbers of corridors identified in these two datasets, and cor_{p_i} $(1 \leq i \leq N_1)$ and cor_{q_j} $(1 \leq j \leq N_2)$ represent each individual corridor in these two datasets.



(a) corridors (cor_1, cor_2)

$$\begin{array}{c|ccc} & e_4 & e_5 & e_6 \\ \hline e_1 & 0.1 & 0.4 & 0.8 \\ e_2 & 0.5 & 0.5 & 0.4 \\ e_3 & 0.6 & 0.4 & 0.5 \end{array} \qquad \begin{array}{c|ccc} & e_4 & e_5 & e_6 \\ \hline e_1 & 9 & 1 & 0 \\ e_2 & 0 & 5 & 0 \\ e_3 & 0 & 0 & 7 \end{array}$$

(b) ground distance matrix D (c) flow matrix F

Fig. 4: An illustration of distance calculation between corridors.

```
Algorithm 2: Contrast corridor mining
Input: a positive corridor set $P$ and a negative
       corridor set $Q$
Output: a set of contrast corridors $Cor_{con}$
$Cor_{con} = \varnothing$ ;
for each $cor_{p_i}$ in $P$ do
    $common = 0$ ;
    for each $cor_{q_j}$ in $Q$ do
        calculate the EMD distance
         $EMD(cor_{p_i}, cor_{q_j})$ and the flow matrix $F$
         between $cor_{p_i}$ and $cor_{q_j}$ ;
        $supp = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} f_{kl}}{\sum w_{p_i}}$ ;
        if $EMD(cor_{p_i}, cor_{q_j}) < dis_{thres}$ and
         $supp > supp_{thres}$ then
            $common = 1$ ;
        end
    end
    if $common \neq 1$ then
        $Cor_{con} = Cor_{con} \cup \{cor_{p_i}\}$ ;
    end
end
```

Then we define a contrast corridor as follows:

Contrast Corridor A corridor cor_{p_i} in P is defined as a contrast corridor if it satisfies any of these two following conditions:

1) **Distance** If the distance between corridor cor_{p_i} and any other corridors in Q is greater than a threshold dis_{thres}, then the corridor is treated as a contrast corridor in P, which indicates that cor_{p_i} is sufficiently different from

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

any corridors in Q.

$$EMD(cor_{p_i}, cor_{q_j}) \geq dis_{thres}, \forall cor_{q_j} \in Q; \quad (12)$$

2) **Support** If two corridors are similar to each other but a certain amount of earth of one corridor cannot move to the other corridor, then it will be treated as a contrast corridor. Satisfaction of this condition means that the corridors under contrast are similar but their supports are significantly different.

$$\frac{\sum_{k=1}^{m}\sum_{l=1}^{n} f_{kl}}{\sum w_{p_i}} \leq supp_{thres}, \forall cor_{q_j} \in \{cor_q | EMD(cor_{p_i}, cor_q) \leq dis_{thres}, cor_q \in Q\}, \quad (13)$$

where $\sum w_{p_i}$ is the sum of the weights of all the edges in corridor cor_{p_i}.

The pseudocode is provided in Algorithm 2. The time complexity of contrast corridor mining is $O(N_1 N_2)O(x)$, where $O(x)$ is the time complexity of solving the linear programming problem, which can normally be solved in polynomial time using the simplex algorithm.

V. Experiments and Results

In this section, we first evaluate our proposed algorithm on synthetic data in terms of accuracy, F1 score, precision and recall. Then two real-life case studies are conducted to further evaluate the effectiveness of our method.

A. Contrast Mining on Synthetic Data

To validate the effectiveness of our proposed method, we compare it with two other baseline methods:

1. **Non-density** This method is generally similar to our proposed method except for the distance metric between two edges, i.e., the original Hausdorff distance is adopted. (The non-density method can be any other state-of-the-art method that does not take the heterogeneity into account, such as DFM [19], DTW [20], LCSS [21], EDR [22] and SSPD [23]. However, since the results in [4] have shown that they are not suitable for dealing with mobile network data, we have included them in our empirical analysis.)

2. **Node-based** The ground distance is defined as the spherical distance between two nodes, and for each node in a corridor, its weight is defined as the degree of the node, which is the number of edges that are incident to the node.

Synthetic data are generated by utilizing the real corridors identified by the method proposed in [4]. We generate positive and negative datasets based on the data of Foshan City. Given a set of corridors $COR = \{cor_1, cor_2, \ldots, cor_n\}$, we first divide the corridor set into two equal-sized subsets (here we assume n to be even, i.e., $COR_1 = \{cor_1, \ldots, cor_{n/2}\}$ and $COR_2 = \{cor_{n/2+1}, \ldots, cor_n\}$). The corridors in COR_2 are considered as the negative dataset, i.e., $COR^- = COR_2$. Then the corridors in the positive dataset COR^+can be generated by considering the following four scenarios:

1) $n/2$ corridors have same distribution as the corridors in COR_2;
2) $n/2$ corridors have same distribution as the corridors in COR_1;
3) $n/2$ corridors have same cell set as the cell set of corridors in COR_2 but with distinct distributions;
4) $n/2$ corridors have same distribution as the corridors in COR_2 but with significantly different amount of traffic volume.

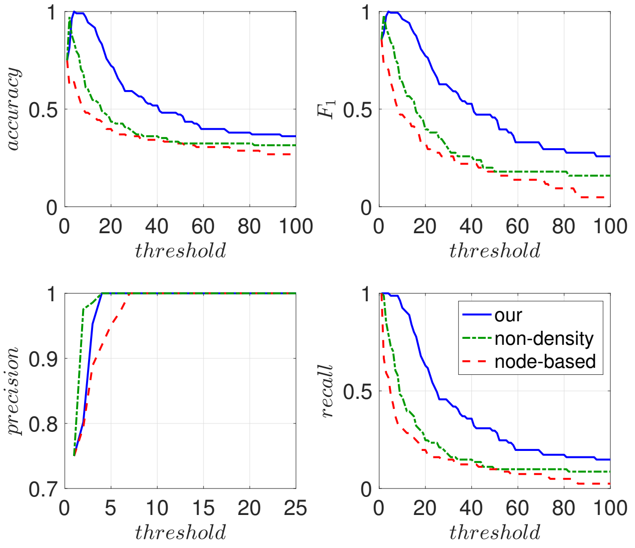


Fig. 5: Performance comparison in contrast pattern finding between our proposed method and two reference methods.

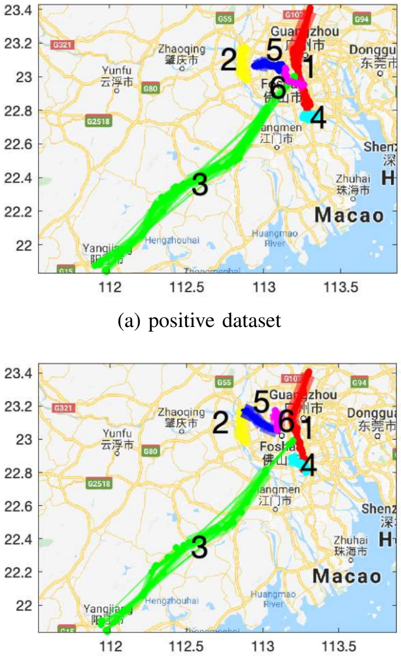
Corridors generated by scenario 1 are treated as common corridors for both positive and negative datasets, whereas corridors generated by scenarios 2, 3 and 4 are contrast corridors according to our definition.

Fig. 5 shows the comparison between our proposed method and the other two baselines in terms of accuracy, F_1 score, precision and recall. The results indicate that our proposed method performs better than the other two methods with varying distance thresholds (which has been normalized for comparison). In the node-based method, the direction of the movement pattern is ignored. Therefore, some closely-located corridors may be treated as very similar even though they have different directions/connections between nodes.

B. Contrast Mining on Real Data

The real-life dataset was originally collected by China Mobile, which contains 5,000 mobile users from a province in South China. The cell locations (longitude and latitude) of each user were recorded every 5 minutes in a time period of three weeks (from 23:55 14/11/2015 to 23:50 05/12/2015).

First, we select two corridor sets to illustrate the effectiveness of our proposed method compared with two other baselines. As shown in Fig. 6, there are six corridors in each of the two corridor sets under contrast. While three of the corridors (the first three corridors in each dataset, labeled as 1, 2, and 3) are the same in both the positive and negative datasets,



(a) positive dataset

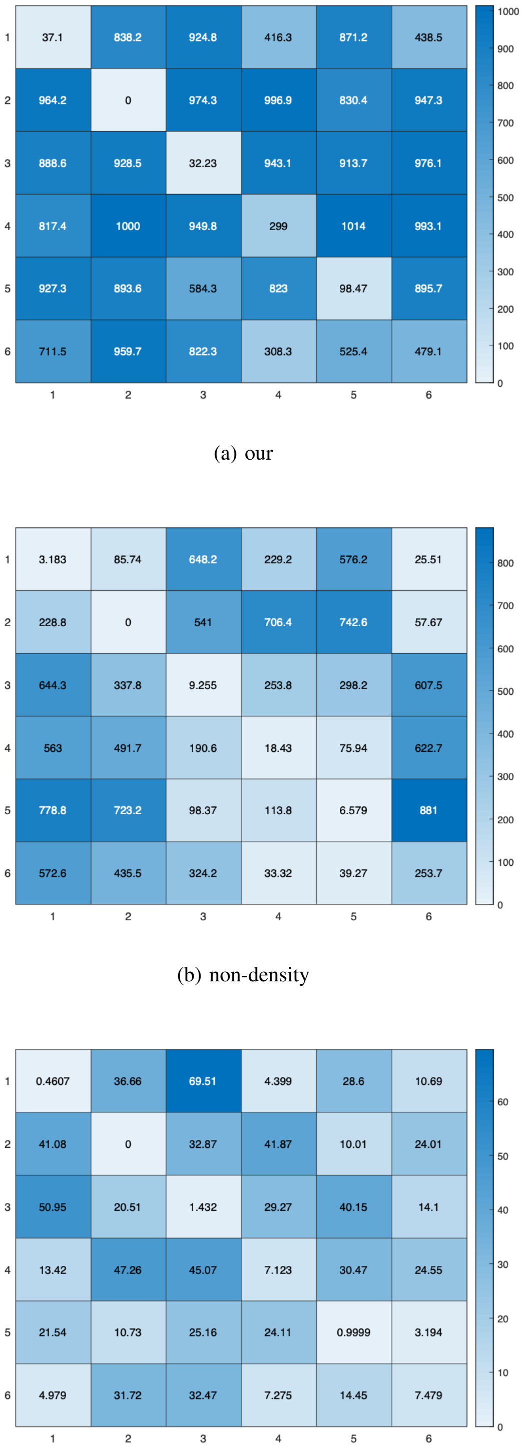
(b) negative dataset

Fig. 6: Two sets of corridors under contrast.

other corridors are different from each other. Fig. 7 shows the distance matrices we obtained by using different methods. Ideally, in the distance matrix, the value at position $(1, 1)$, $(2, 2)$ and $(3, 3)$ should be much smaller than the distances at other positions, because in our dataset there are only these three pairs of similar corridors. By using our proposed method, clearly similar corridors have smaller distances between each other. Although the non-density and node-based method can also identify similar corridors, incorrect conclusions may be obtained when calculating the distance between corridor 5 in the positive dataset and corridor 5 in the negative dataset. This is because these two corridors have similar directions and they are located close to each other in a dense area. Our method can deal with this well since it normalizes the distance by considering the heterogeneous distribution of cell towers.

Then two sets of experiments are conducted using the dataset in Foshan City, i.e., weekday (Mon-Fri) vs weekend (Sat-Sun) and morning (0:00-12:00) vs afternoon (12:00-0:00). The EMD between weekdays and weekends in Fig. 8a is 0.7632. There are more contrast corridors identified on weekdays in Fig. 8a, and most of them are located in the central area, which indicates that people would move more frequently on weekdays, especially in the central area. The contrast patterns on the weekdays may be attributed to work commutes. By contrast, some contrast patterns surrounding the urban area (e.g., beltway) are identified on weekends, as shown in Fig. 8b, and these patterns may be explained by leisure activities out of the city center on weekends.

The results of morning vs afternoon (Fig. 8c and Fig. 8d) indicate that the city is more active in the afternoon compared with in the morning, since more contrast corridors are identi-



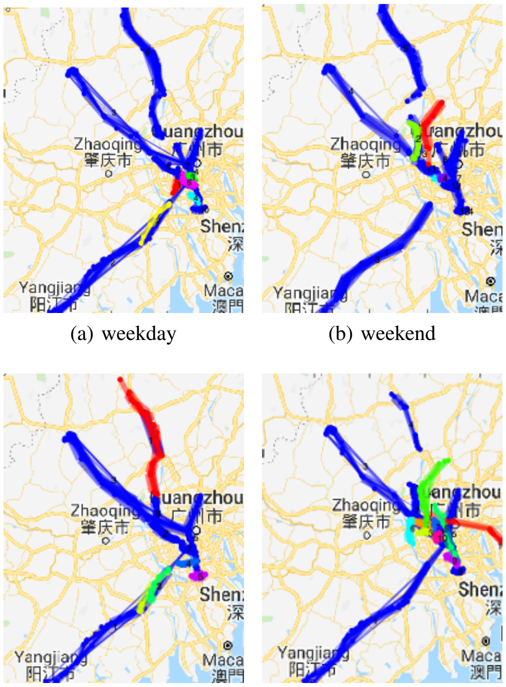
(a) our

(b) non-density

(c) node-based

Fig. 7: The distance matrix between two sets of corridors obtained by three different methods.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE



(a) weekday (b) weekend

(c) morning (d) afternoon

Fig. 8: Identified common and contrast corridors in different time periods. Corridors with blue color indicate they are common corridors in two time periods, while other colors are contrast corridors.

fied in the afternoon. Most contrast corridors in the morning are located outside the city area, while in the afternoon some contrast corridors are identified in the central area. Note that the afternoon time period starts from 12:00 to 0:00, which means the corridors are generated not only from the afternoon activity but most of the evening movement of mobile users.

VI. Conclusion

In this paper, a framework for mining the changes in the movement patterns of mobile users is proposed. We consider the non-homogeneous distribution of cell towers in the distance measure, which is more appropriate for trajectories generated in mobile networks compared to other state-of-the-art distance measures. A contrast corridor mining algorithm is also proposed to find significant changes between the corridors generated in different time periods. Both synthetic and real-life datasets are applied to validate the effectiveness of our proposed method. Contrast corridors can be effectively detected from trajectories in mobile networks, and our method outperforms others by an average 20% improvement in the F1 score. Our findings could help mobile operators to identify the key focus areas (i.e., corridors) in large-scale deployments of 5G networks for cost minimization, and the ability to identify the temporal dynamics of corridor patterns can help the management and orchestration of 5G network resources. In the future, it would be interesting to study the identification of the change points of movement patterns.

References

[1] R. Mijumbi, J. Serrat, and et al., "Management and orchestration challenges in network functions virtualization," *IEEE Communications Magazine*, pp. 98–105, 2016.

[2] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5g: Survey and challenges," *IEEE Communications Magazine*, pp. 94–100, 2017.

[3] H. Zhu, J. Luo, and et al., "Mining trajectory corridors using fréchet distance and meshing grids," in *PAKDD'10*, 2010, pp. 228–237.

[4] L. Li, S. Erfani, C. A. Chan, and C. Leckie, "Multi-scale trajectory clustering to identify corridors in mobile networks," in *CIKM'19*, 2019, pp. 2253–2256.

[5] N. Zygouras and D. Gunopulos, "Discovering corridors from gps trajectories," in *SIGSPATIAL'17*, 2017, pp. 61:1–61:4.

[6] L. Li, C. A. Chan, S. Erfani, and C. Leckie, "Adaptive edge caching based on popularity and prediction for mobile networks," in *IJCNN'19*, 2019, pp. 1–10.

[7] C. A. Chan, M. Yan, A. F. Gygax, W. Li, L. Li, I. Chih-Lin, J. Yan, and C. Leckie, "Big data driven predictive caching at the wireless edge," in *ICC Workshops*, 2019, pp. 1–6.

[8] P. K. Agarwal, K. Fox, K. Munagala, A. Nath, J. Pan, and E. Taylor, "Subtrajectory clustering: Models and algorithms," in *Proc. ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2018, p. 75–87.

[9] J. Lee and J. Han, "Trajectory clustering: A partition-and-group framework," in *SIGMOD'07*, 2007, pp. 593–604.

[10] J. Gudmundsson, A. Thom, and J. Vahrenhold, "Of motifs and goals: Mining trajectory data," in *SIGSPATIAL'12*, 2012, pp. 129–138.

[11] Y. Qiao and et al., "A mobility analytical framework for big mobile data in densely populated area," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 1443–1455, 2017.

[12] G. Dong and J. Bailey, *Contrast Data Mining: Concepts, Algorithms, and Applications*, 1st ed. Chapman & Hall/CRC, 2012.

[13] A. García-Vico and et al., "An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018.

[14] L. Li, S. Erfani, and C. Leckie, "A pattern tree based method for mining conditional contrast patterns of multi-source data," in *ICDMW'17*, 2017, pp. 916–923.

[15] X. Wang, C. Leckie, and et al., "Discovering the impact of urban traffic interventions using contrast mining on vehicle trajectory data," in *PAKDD'15*, 2015, pp. 486–497.

[16] W. Wu, Y. Wang, and et al., "Oscillation resolution for mobile phone cellular tower data to enable mobility modelling," in *MDM'14*, 2014, pp. 321–328.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *ICCV'98*, 1998, pp. 59–66.

[18] F. Hausdorff, *Grundzüge der mengenlehre*. American Mathematical Soc., 1978, vol. 61.

[19] T. Eiter and H. Mannila, "Computing discrete fréchet distance," Citeseer, Tech. Rep., 1994.

[20] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD'94*, 1994, pp. 359–370.

[21] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *ICDE'02*, 2002, pp. 673–684.

[22] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *VLDB'04*, 2004, pp. 792–803.

[23] P. C. Besse, B. Guillouet, J. Loubes, and F. Royer, "Review and perspective for distance-based clustering of vehicle trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, pp. 3306–3317, 2016.