AttractRank: District Attraction Ranking Analysis Based on Taxi Big Data

Guangqiang Xie, *Member, IEEE,* Runpeng Zhang, Yang Li, Ling Huang, Chang-Dong Wang, *Member, IEEE,* Hao Yang, and Jiahao Liang

***Abstract*—The city's district attraction ranking plays an essential role in the city's government because it can be used to reveal the city's district attraction and thus help government make decisions for urban planning in terms of the smart city. The traditional methods for urban planning mainly rely on the district's GDP, employment rate, population density, information from questionnaire surveys and so on. However, as a comparison, such information is becoming relatively less informative as the explosion of an increasing amount of urban data. What's more, there is a serious shortcoming in these methods, i.e., they are independent representations of the attraction of a district and do not take into account the interaction among districts. With the development of urban computing, it is possible to make good use of urban data for urban planning. To this end, based on taxi big data obtained from Guangzhou, China, this paper proposes a district attraction ranking approach called AttractRank, which for the first time uses taxi big data for district ranking. An application system is developed for demonstration purposes. Firstly, the entire Guangzhou city is divided into a number of districts by using Constrained K-means. Secondly, the original PageRank algorithm is extended to integrate with the taxi's OD (origin-destination) points to establish the OD matrix, whereby the attraction ranking of each district can be calculated. Finally, by visualizing the results and case studies obtained from AttractRank, we can successfully obtain the pattern of how attractions of districts change over time and interesting discoveries on urban lives, therefore it has wide applications in urban planning and urban data mining. The application system has been deployed online at http://qgailab.com/ieee/attractrank/index.html.**

***Index Terms*—PageRank, Constrained K-means, district attraction, taxi's OD points**

I. Introduction

CHINA has four megacities, namely Beijing, Shanghai, Guangzhou, and Shenzhen. Although these megacities have developed very well, the development of different districts in the city is very uneven. For example, the GDP of the Conghua District in Guangzhou can not reach one-tenth of the GDP of the Tianhe District[1]. What's more, taxis are concentrated in the Tianhe District and rarely go to the Conghua District, which means that the attraction of the Tianhe District is very high. It is an urgent need for the government to understand the attraction of different districts and identify various functional districts of the city with urban computing involved, which helps to optimize urban planning in a city and promote a comprehensive upgrade of various functional districts of the city. Therefore, it is very important to analyze the attraction of each district of the city.

The classical methods for urban planning mainly rely on traditional measures such as district GDP, employment rate, population density and information from questionnaire surveys, which have poor timeliness, low accuracy, and are unable to obtain satisfactory results. What's more, these methods are not as powerful as they used to be in the era of big data, in which urban computing with urban data is more and more popular. With the rapid development of location-aware technology, it is becoming possible to obtain a large amount of travel data, and discovering trip attractive areas from massive movement data is becoming a new research focus. Zheng $et\ al.$ [1] proposed an attractive area clustering algorithm based on grid density. Based on taxi spatio-temporal trajectory data, Mou $et\ al.$ [2] proposed a non-negative matrix decomposition model with kernel function and used this model to identify urban functions. Sun $et\ al.$ [3] proposed systematic research on urban Region-of-Interest analysis through mining the large-scale online map query logs. In addition, May $et\ al.$ [4] presented a framework for classifying districts in cities by their attractiveness to visitors, and relating Points of Interests types to district attraction patterns. However, these methods have a shortcoming that they ignore the interaction among districts, it still lacks quantitative ways to investigate district attractions in a holistic manner. The attraction of a city's district must be inextricably linked to other districts. For example, as will be shown later, the attraction of the districts near the Guangzhou Tianhe Central Business District will be strongly influenced by it. Although a method based on PageRank [5] has taken the interaction among districts into consideration [6], it is along with a shortcoming which would lead to incredible results in some cases. To this end, we propose a district attraction ranking method called AttractRank, which is the first attempt to use taxi big data for district ranking as well as to overcome some shortcomings of the existing related methods. In particular, we focus on a particular type of mobile data,

This work was supported by NSFC (61976052, 61876193), National Key Research and Development Program of China (2018YFC0809700), Guangdong Natural Science Funds for Distinguished Young Scholar (2016A030306014). Corresponding author: Yang Li.

Guangqiang Xie, Runpeng Zhang, Yang Li, Hao Yang and Jiahao Liang are with the College of Computer, Guangdong University of Technology, Guangzhou, China.
E-mail: xieqg@gdut.edu.cn, runpengzhang1998@gmail.com, liyang@gdut.edu.cn, gdhy9064@gmail.com, 1436266185@qq.com.

Ling Huang and Chang-Dong Wang are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, and with Guangdong Province Key Laboratory of Computational Science, Guangzhou, China, and with Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China.
E-mail: huanglinghl@hotmail.com, changdongwang@hotmail.com.

[1]http://210.72.4.52/gzStat1/chaxun/ndsj.jsp

namely the origin-destination pair, and propose a new way to calculate the attraction of each district of the city. To extract information from complex connections among a large number of locations, we analyze the big data of Guangzhou taxi GPS OD (origin-destination) points and use these points to divide Guangzhou into a number of districts by Constrained K-means and analyze the OD matrix in different time periods. Then, we propose a novel algorithm named AttractRank to obtain the attraction of each district in the city.

For clarity, the main contributions of this paper are summarized as follows:

- A new method of calculating district attraction ranking called AttractRank is proposed, which is the first attempt to use taxi big data for district ranking as well as to overcome some shortcomings of the existing related methods.
- A web online system is developed for demonstration purposes, which is available online at http://qgailab.com/ieee/attractrank/index.html. This visualization system can display district attraction ranking and various interactive charts over different time periods on the Google map. In addition, our taxi statistic data and codes are made publicly available on Github at https://github.com/GDUT-Rp/2020_IEEE_AttractRank for academic research usage.
- The approach is applicable to other cities, as long as the taxi big data is available.
- The results calculated with AttractRank algorithm can be adopted to further analysis and applications, helping to discover more informative patterns of the city.

The remainder of the paper is organized as follows: Section II briefly summarizes the related work. Section III introduces the framework of AttractRank and the techniques applied. Section IV introduces the data preprocessing of taxi big data. Section V describes the extension of PageRank to calculate the attraction of districts in a city. In Section VI, a comparison between algorithms is conducted. In Section VII, experiments are conducted. Finally, Section VIII concludes this paper.

II. Related Work

In the past few decades, many efforts have been made in identifying and analyzing the spatial attractive districts, points or zones. But the existing studies mainly focus on some spatial attributes based on population, economic and ethnographic structures [7]. Nowadays, as the development of network theory and the emergence of various data such as social data, migration data, taxi operating data and smart card data, statistical analysis of network and flows between locations can be conducted. Based on the aforementioned data, a research direction called smart city has emerged [8], [9]. Its goal is to use data from various sources to solve some of the problems encountered in today's urbanization process, such as air pollution [10], traffic congestion [11], and energy waste [12]. In the era of big data, data-driven application has become a feature in many areas, such as wastewater treatment process [13], fault detection [14].

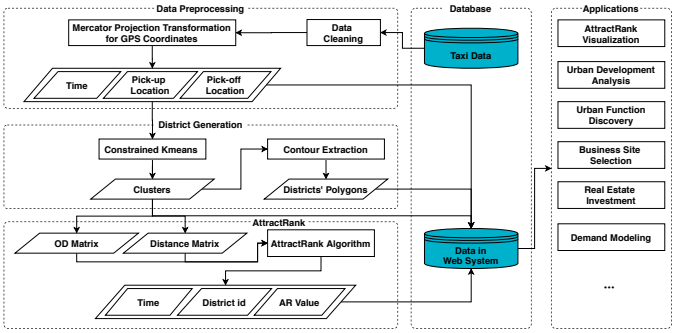


Fig. 1. The framework overview of AttractRank.

Specifically, the problem we are studying in this paper, namely the nature of the city's region [15], is one of the key challenges in this research direction. Some early efforts have been made on network analysis by means of weighted PageRank [16], which aim to explore the district attraction, such as Place Rank [17] and Geographically modified PageRank [6]. After that, with the maturity of big data technologies and location-aware computing technologies, taxi trajectory data and point-of-interest(POI) data are increasingly adopted to analyze the attraction of districts. Zheng $et\ al.$ [1] devised a grid-density-based clustering algorithm to discover travel attractive districts in different periods of a day. Mou $et\ al.$ [2] identified the functions of districts quantitatively by proposing a non-negative matrix decomposition model. Sun $et\ al.$ [3] used the method of dividing the urban area into small region grids and evaluating their attractions by their PageRank values to identify the attractions of those grids and the latent travel patterns of visitors.

Apart from the attractions of city districts, there is much research on calculating attraction indexes at different levels of the district [4], [18], [19]. May $et\ al.$ [4] presented a framework for classifying districts in cities by their attractiveness to visitors, and relating Points of Interests types to district attraction patterns. The migration data of passenger transport around the Spring Festival in China can help us to realize the difference between urban development in China [18]. In addition, a novel algorithm called RoadRank was proposed to compute the influence scores of each road segment in an urban road network [19].

III. Data Processing and Visualization Techniques

A. The Framework of AttractRank

Figure 1 shows the overview of our framework of AttractRank. In the first task, we preprocessed the taxi data and used the technique of converting the GPS coordinates into the plane rectangular coordinate system, extracting the OD pairs and saving them to the database. In the second task, Constrained K-means is adopted to divide the city into a specified number of districts using all points of OD pairs, and the contour extraction algorithm is used to generate the districts' outline by these points, which can be clearly observed in our visualization system. In the third task, by extracting the travel flows among these districts,

TABLE I
SOME SAMPLES OF TAXI OPERATING RECORDS.

Properties		Operating record #1	Operating record #2	Operating record #3	Operating record #4
Equipment ID		EID #1	EID #2	EID #2	EID #3
Plate number		No. 1	No. 2	No. 2	No. 3
Company ID		CID #1	CID #2	CID #2	CID #3
Departure time		2017-02-01 00:38:50	2017-02-01 00:27:23	2017-02-01 00:48:47	2017-02-01 00:51:08
Arrival time		2017-02-01 01:14:02	2017-02-01 00:29:34	2017-02-01 01:00:22	2017-02-01 01:03:20
Origin location	Longitude	113.295573	113.302767	113.259270	113.279844
	Latitude	23.388139	23.101922	23.122217	23.139482
Destination location	Longitude	113.337824	113.290531	113.328548	113.293651
	Latitude	23.118570	23.105612	23.134518	23.105482
Mileage (KM)		43.07	1.40	8.70	6.32
Unit price (1CNY/KM)		3.90	2.60	2.60	2.60
Price (1CNY)		127	10	27	22
Trade code		Code #1	Code #2	Code #3	Code #4

we form an OD matrix where each element denotes the flow frequency between corresponding districts and a distance matrix where each element denotes the distance between corresponding districts. After that, the AttractRank algorithm is conducted on the matrices in order to calculate the attraction value of each district. The sources of AttractRank algorithm, visualization system and data processing are available at https://github.com/GDUT-Rp/2020_IEEE_AttractRank. These sources are mainly implemented based on Python 3.6, Java 1.8, JavaScript with the Google Maps API.

B. Mercator Projection

Mercator Projection is one of the projections which transform the coordinates with longitude λ and latitude φ to the two-dimensional coordinates with x and y, the conversion formula is as follows:

$$x = \pi R \lambda \tag{1}$$

$$y = \pi R \ln\left[\tan(\frac{\pi}{4} + \frac{\varphi}{2})\right] \tag{2}$$

where R is the equatorial radius of the earth.

C. Constrained K-means

K-means is a clustering algorithm that is used to generate a specified number of clusters with data of points. Mini Batch K-means is its optimized version which is applied in clustering for a large amount of points. Constrained K-means [20] is a variant of K-means with cannot-link constraints. It is required cannot-link constraints to be provided, which is hard to obtain in reality. As an alternative, we collect the polygon boundaries in which the points can't link to others from different ones. Then Mini Batch K-means is adopted in each polygon boundary to generate clusters as Constrained K-means should have done.

D. Contour Extraction

Contour extraction is a technique to extract contours from data of points. There are various methods for different purposes, such as Convex Hull and Alpha Shape. Convex Hull always extracts the convex polygon which just fits the points, while Alpha Shape can generate the concave polygon which may maintain the shape of given points in some cases according to the setting parameters. For the sake of pretty display which is more fitting to the map, we choose the Alpha Shape as the algorithm of contour extraction. Otherwise, Convex Hull is enough to extract contours of points.

E. Compare Data Processing and Visualization Techniques

As a comparison, an innovative data analysis is demonstrated for weather using Cloud Computing, integrating both system and application Data Science services to investigate extreme weather events [21]. Chang explained the data processing architecture and the principles underlying P-Map, P-Merge and P-Reduce and proposed a greedy algorithm to complete big data processing while considering the number of nodes [22]. In contrast, in our data processing, Constrained K-means is adopted to cluster the data of the pick-up and drop-off points to divide the districts and use the Alpha Shape algorithm to draw the polygonal outline of each district for each cluster, which can be clearly displayed in our visualization application system. In addition, we develop a novel vehicle flow diagram and a comparison view that can be loaded with tens of thousands of data, which shows the flow of taxis. It can be seen that our data processing techniques and data visualization techniques are both lightweight and convenient.

IV. DATA PREPROCESSING AND ANALYSIS

A. Data Introduction

Our taxi data consists of 11,175,138 taxi operating records in 59 days (i.e., two months) in Guangzhou, China. Each record contains important information during a route, such as the coordinates of the origin and the destination, the price and the distance along the trip. For clarity, Table I lists some samples of taxi operating records. It should be noted that operating record #2 and #3 are the continuous records of the same taxi. One key information is the coordinates of the origin and the destination in every record.

B. Data Preprocessing

First, data cleaning is performed on the coordinate data by removing all data in which the coordinates of the origin or the destination are out of Guangzhou, which occupies 3.93% records.

Second, the Euclidean distance between coordinates with longitude λ and latitude φ is not too pronounced to distinguish the different distances in the corner of the world [23]. In order to process the clustering algorithm, we need to convert the coordinate (λ, φ) to the new coordinate denoted as (x, y) on the map of the Mercator projection. After the conversion, the Euclidean distance between coordinates with x and y is computed to facilitate the calculation of the clustering algorithm as will be shown in the next subsection.

C. District Generation and Analysis

To analyze the attraction in various areas in Guangzhou, we need to partition the entire Guangzhou into a number of districts first. While the traditional district generation is based on the administrative district where some districts are usually of abnormal shapes and related with administrative management, it is not applicable in the attraction analysis in our work. This is because there may exist several sub-districts in one administrative district that has diverse attraction. To this end, the clustering algorithm is adopted to generate the districts by clustering nearly 10 million coordinate points of origins and destinations collected in two months. Because of the dense spatial distribution characteristics of the traffic transportation data, the K-means clustering algorithm is adopted to cluster all the coordinates of the origins and destinations to generate districts automatically. Due to the reason that the Pearl River lies across the entire Guangzhou city, there are many rivers interconnected by a number of bridges, which naturally separate the Guangzhou city into several main lands. Therefore, Constrained K-means [20] is adopted by taking the river barriers into consideration, using the polygon boundaries consisting of rivers selected manually with the tool of an online map[2]. According to the cluster number estimate method introduced in [24], it seems to be suitable to separate the entire Guangzhou city into 90 clusters, with each cluster being a district. However, in this case, there exist some clusters that are distributed in some areas far away from the urban areas, we have to add some centers manually into these clusters so as to ensure that these remote areas contain enough cluster centers, adding up to 135 centers. In this way, a total of 134 districts are obtained by removing a district that is on an isolated island without any taxi records, as illustrated in Figure 2.

For the sake of having a deeper comprehension of the attraction of various districts in Guangzhou and preparing for the further computation, we conduct the statistics in each district based on the origin-destination data of total two-month data. And then we build the OD matrix, the element of which stores the amount of the records of the taxi transferring from one district to another with the precision of an hour. The detailed description will be elaborated later. For clarity, the amount of taxi origin-destination in each district within two months is plotted in Figure 3.

[2]http://geojson.io

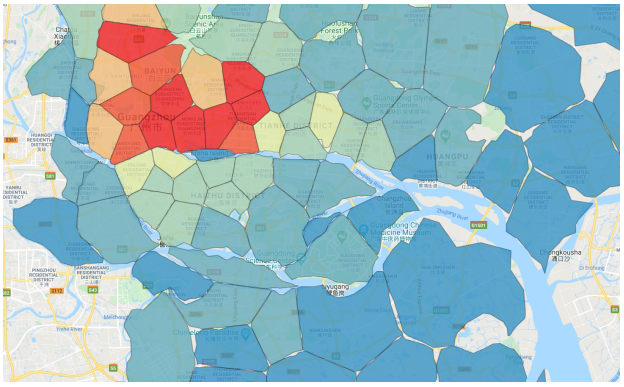


Fig. 2. Illustration of districts of Guangzhou: The boundary of each district is generated by the Alpha Shape algorithm, which maintains the shape of the OD points within every district.

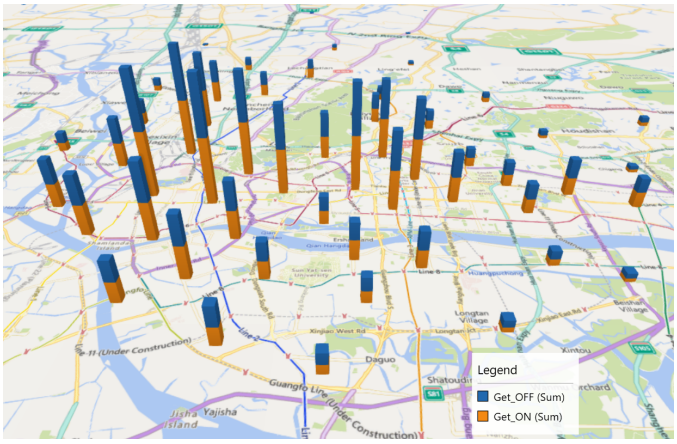


Fig. 3. The amount of taxi origins and destinations in each district respectively: Notice that taxi origins and destinations in Guangzhou are mainly concentrated in the urban districts. In addition, it is out of our subjective viewpoint that they are highly concentrated in Yuexiu District rather than Tianhe Central Business District.

V. The Proposed AttractRank Algorithm

A. PageRank and Extensions

PageRank is one of the classical methods for analyzing linkage structure of the web [5]. It can be regarded as a procedure for simulating the movement of random surfers within a web-page network connected by hyperlinks [25]. Many efforts have been made in providing a deep insight into the theory of PageRank [26], [27], [28].

The formula of PageRank is as follows:

$$PR(u) = \sum_{v \in \mathcal{B}_u} \frac{PR(v)}{L(v)} \tag{3}$$

where u, v are website pages, $PR(u)$ represents the signification of website page u, $\mathcal{B}_u$ is a set consisting of all of the webpages which are backlinks (e.g., a is a backlink of b means that a links to b) of webpage u, $L(v)$ represents the number of the external links starting from webpage v [29].

But there is a fatal shortcoming that dead-ends (i.e., the website pages without any external links) would result in the consequence that PR values converge to zero through calculation via Eq. 3 for the black-hole-like impact of dead-ends. In order to handle the problem above, Larry Page proposed the following improved formula:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in \mathcal{B}_u} \frac{PR(v)}{L(v)} \ , \ s.t.\ 0 < d < 1 \tag{4}$$

where N represents the total number of the pages, d is the ratio coefficient balancing the weight of the two parts.

Equation 4 is mainly applied to the website pages. To adapt to the geographical situation, the GPR (Geographic PageRank) was proposed in [6], which incorporates the effect of geographic proximity and attractiveness of the location. The GPR formula is as follows:

$$\begin{aligned} F_G(i,j) &= \frac{I(j)^\alpha}{d_{i,\ j}^\beta} \\ GPR(u) &= \sum_{v \in \mathcal{O}_u} GPR(v) \times \frac{F_G(v,u)}{\sum_{w \in \mathcal{D}_v} F_G(v,w)} \end{aligned} \tag{5}$$

where $I(j)$ represents the in-degree of district j, $d_{i,j}$ is the distance between district i and district j, α and β are two factors measuring the impact on the in-degree and distance, respectively. $GPR(x)$ represents the attractiveness (i.e., the signification) of x, $\mathcal{O}_u$ represents the set of origins whose destination is district u, $\mathcal{D}_v$ represents the set of destinations whose origin is district v.

It is deserved to be mentioned that although the GPR considers the distance factor, it is along with the shortcomingthat it doesn't make full use of the OD data.

B. The AttractRank Algorithm

For the sake of adapting the PageRank algorithm to our work as well as solving the shortcomings above, we propose an extension of PageRank called AttractRank. The main idea is as follows. The reciprocal value of the distance is introduced as the bias of the probability setting out from one district to another, making all the districts have a chance of accessing other districts. In other words, there is no dead-end or district with the only self-loop among all districts. Let $AR(u)$ denote the attraction of district u. The formula of AttractRank is as follows:

$$\begin{aligned} AR(u) &= \sum_{v \in \mathcal{N}} AR(v) \times P(v,u) \\ s.t. \sum_{i \in \mathcal{N}} AR(i) &= \sum_{j \in \mathcal{N}} \sum_{k \in \mathcal{N}} OD_{j,k} \end{aligned} \tag{6}$$

where $\mathcal{N}$ denotes the district set, OD is the origin-destination matrix with each $OD_{i,j}$ denoting the amount of operating records whose origin is district i and destination is district j, and P is the transferring probability matrix with each $P_{i,j}$ representing the probability that the attraction $AR(i)$ of district i transfers to the attraction $AR(j)$ of district j. In particular,

Algorithm 1 The AttractRank Algorithm

Input: The origin-destination matrix OD; The distance matrix D; The trade-off parameter r.
Output: A vector of AttractRank values of all districts AR.

1: Compute the sum of matrix OD as S;
2: Compute the distance impact matrix G between each pair in all districts via Eq. 8;
3: Compute transferring probability matrix P via Eq. 7;
4: Find the normalized principle eigenvector E of matrix P^T whose eigenvalue is 1;
5: For each district i, $AR_i \leftarrow E_i \times S$;
6: **return** AR.

the transferring probability matrix P can be computed as follows:

$$P_{v,u} = \begin{cases} \frac{G(v,u)}{\sum_{w \in \mathcal{N}} G(v,w)} & \text{if } \sum_{w \in \mathcal{N}} OD_{v,w} = 0 \\ (1-r)\frac{G(v,u)}{\sum_{w \in \mathcal{N}} G(v,w)} + r \cdot \frac{OD_{v,u}}{\sum_{w \in \mathcal{N}} OD_{v,w}} & \text{otherwise} \end{cases} \tag{7}$$

where u and v represent two districts, r is the trade-off parameter, and $G(v,u)$ represents the impact of distance on the probability from district v to district u. It is defined as follows:

$$G(v,u) = \frac{1}{D_{v,u}} \tag{8}$$

where $D_{v,u}$ represents the distance between districts v and u, i.e., the distance between the centers of the two districts. In particular, $D_{v,v}$ is defined as $\min(D_{v,u})$ where district u is not district v.

For clarity, the entire AttractRank algorithm is summarized in Algorithm 1.

C. Theoretical Analysis

To simplify the analysis of the existence and uniqueness of the AttractRank algorithm, we introduce the concept of Markov chain, which is a very particular kind of stochastic process. A Markov chain is defined as follows [30]:

Definition 1: A stochastic process $X = \{X_n : n \geq 0\}$ on a countable set $\mathcal{S}$ is a Markov Chain if, for any $i, j \in \mathcal{S}$ and $n \geq 0$,

$$P\{X_{n+1} = j | X_0, \cdots, X_n\} = P\{X_{n+1} = j | X_n\} \tag{9}$$

$$P\{X_{n+1} = j | X_n = i\} = p_{ij} \tag{10}$$

where p_{ij} is the probability that the Markov chain jumps from state i to state j which satisfies $\sum_{j \in \mathcal{S}} p_{ij} = 1, i \in \mathcal{S}$. And we use matrix $P_{MC} = [p_{ij}]$ to denote the transition matrix of the chain.

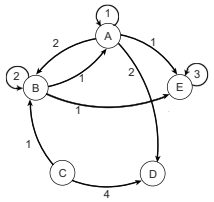
In the theory of Markov chains, one of the most significant parts, apart from the transition matrix, is stationary distributions. Stationary distributions are defined as follows [30]:

Definition 2: A row vector π is a stationary distribution for the Markov chain X_n with the transition matrix P_{MC} if

$$\pi = \pi P_{MC} \tag{11}$$

$OD_{o,d}$ Orig. \ Dest.	A	B	C	D	E	Sum
A	1	2	0	2	1	6
B	1	2	0	0	1	4
C	0	1	0	4	0	5
D	0	0	0	0	0	0
E	0	0	0	0	3	3
Sum	2	5	0	6	5	18

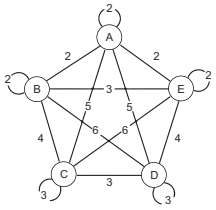
(a) Origin-destination matrix



(b) The graph of origin-destination matrix

$D_{o,d}$ Orig. \ Dest.	A	B	C	D	E
A	2	2	5	5	2
B	2	2	4	6	3
C	5	4	3	3	6
D	5	6	3	3	4
E	2	3	6	4	2

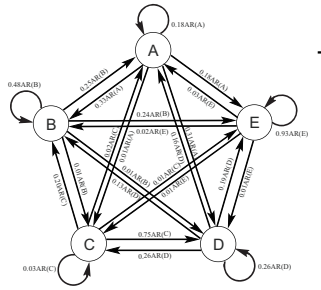
(c) Distance matrix



(d) The graph of distance between the five districts

(e) Transferring probability matrix

$P_{i,j}$ Orig. \ Dest.	A	B	C	D	E	Sum
A	0.18	0.33	0.01	0.31	0.18	1.0
B	0.25	0.48	0.01	0.01	0.24	1.0
C	0.02	0.20	0.03	0.75	0.01	1.0
D	0.16	0.13	0.26	0.26	0.19	1.0
E	0.03	0.02	0.01	0.01	0.93	1.0



(f) The graph of transferring probability matrix

Districts	AR
A	1.27
B	1.81
C	0.52
D	1.34
E	13.06
Sum	18

(g) AttractRank of the five districts

Fig. 4. Demonstration of the process of AttractRank: The matrix OD is shown in (a), the graph of which is shown in (b). Equally, the matrix D is shown in (c), and the graph of the distance between each district is exhibited in (d). With the matrix OD, matrix D and trade-off parameter r, the matrix P is computed as shown in (e) and visualized in (f). Finally, the vector AR of all districts is shown in (g).

A Markov chain may have infinite stationary distributions. There is a theorem relating to the stationary distributions of Markov chains [30]:

Theorem 1: If the Markov chain X_n is irreducible and aperiodic, π_u is a stationary distribution of X_n, then π_u is the stationary distribution for X_n unique up to multiplication by a constant.

An irreducible Markov chain means that the probability transferring from state i to state j at one or more steps is greater than 0 for any $i, j \in \mathcal{S}$, that is

$$\forall i, j \in \mathcal{S}, \exists n \in N^+, p_{ij}^n > 0 \tag{12}$$

where p_{ij}^n represents the probability transferring from state i to j at step n, particularly, $p_{ij}^1 = p_{ij}$.

An aperiodic Markov chain means that the period d_i of state i is 1 for any $i \in \mathcal{S}$, where period d_i is the greatest common divisor of all n which satisfies that the probability transferring from state i to itself at step n is greater than 0, that is

$$\begin{aligned} &\forall i \in \mathcal{S}, d_i = 1 \\ &d_i = \min_{j,k \in \mathcal{C}_i} gcd(j,k) \qquad (13) \\ &\mathcal{C}_i = \{n | p_{ii}^n > 0\} \qquad (14) \end{aligned}$$

where $gcd(j,k)$ is the greatest common divisor of integer j and k.

In accordance with Eq. 7, the transferring probability matrix P satisfies that $\sum_{j \in \mathcal{N}} P_{ij} = 1, i \in \mathcal{N}$, such that P is a transition matrix of a Markov chain. Further more, by denoting this Markov chain as X_{AR}, X_{AR} is irreducible because

$$\forall i, j \in \mathcal{N}, P_{ij} > 0 \tag{15}$$

In the meantime, X_{AR} is aperiodic because

$$\begin{aligned} &\forall i, j \in \mathcal{N}, P_{ij} > 0 \qquad (16) \\ \Rightarrow &\forall i \in \mathcal{N}, \mathcal{C}_i = N^+ \\ \Rightarrow &\forall i \in \mathcal{N}, d_i = 1 \end{aligned}$$

According to Theorem 1, it is guaranteed that X_{AR} has a stationary distribution unique up to multiplication by a constant, i.e., Eq. 11 with P as P_{MC} has a unique normalized solution π_u. In the AttractRank algorithm, Eq. 6 can also be expressed as

$$\pi_{AR} = \pi_{AR} P \tag{17}$$

where π_{AR} is a row vector of attraction AR, i.e., there exists a unique normalized solution π_u in Eq. 6, and we have

$$\pi_{AR} = (\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} OD_{i,j}) \pi_u \tag{18}$$

Thus the existence and uniqueness of the AttractRank algorithm is proved.

VI. Comparison between AttractRank and GPR

For illustration and comparison purpose, a demonstration example is conducted as follows. Assume that there are a total of 5 districts, denoted as A, B, C, D and E, respectively. The origin-destination matrix of these districts and the corresponding graphs are shown respectively in Figure 4(a) and Figure 4(b). Notice that there is a dead-end D without external links and a district E with the only self-loop external links. The distance between each other is shown in Figure 4(c) and Figure 4(d). Based on the origin-destination matrix OD and the distance matrix D, by setting $r = 0.9$ which approximates to the default value 0.85 used in the original PageRank and its variants [5], [26], [27], [28], the transferring probability matrix P can be computed as Figure 4(e) and Figure 4(f). It shows

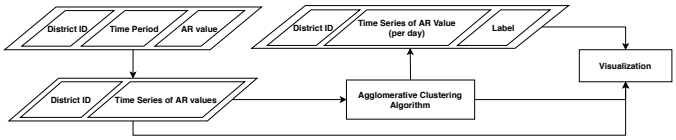


Fig. 5. Flow chart of data processing in experiment.

that all the districts are possible to access the others, without any districts alone. Based on the transferring probability matrix P, the AttractRank vector AR of the 5 districts is shown in Figure 4(g). Notice that the sum of AR of the 5 districts is equal to the sum of the OD matrix.

With the same condition, setting $\alpha = 1$ and $\beta = 1$, we calculated the GPR values as well shown as Table II via Eq. (5). It can be observed that GPR values didn't distinguish between district B and district E, as well as district C and district D for its shortcoming, despite their unlike inflows and outflows.

TABLE II
THE GPR VALUES OF FIVE DISTRICTS.

Districts	A	B	C	D	E
GPR values	5.91	5.47	5.07	5.07	5.47

VII. EXPERIMENTS

A. Experiment Setup

The experimental environment is based on Python 3.6. The web online system is developed for demonstration purposes, which is based on Alibaba Cloud Server Ubuntu 16.04, running with the environment of jdk1.8, storage component Mysql 5.7 and cache server redis 3.2. The system is deployed online at http://qgailab.com/ieee/attractrank/index.html.

To show the data processing and technologies involved in the experiment, we draw the flow chart of the experiment, as shown in Figure 5. In the experimental part, we first convert the data obtained by AttractRank to get the time series of districts' AR values. Then, the properties of the original data are shown initially using the distribution map. After that, we divide the 134 districts into eight groups by employing the Agglomerative Clustering algorithm, and we also show the visualization of the clustering results. Finally, the properties of each group are discussed by utilizing the stack area graph and cartographic visualization techniques.

B. Analysis of Aggregating Districts

The network of operating records is used to calculate the AttractRank values of each district in Guangzhou. As can be seen from Figure 6(a), the mean AttractRank values per hour of each district have a heavy-tailed distribution, as will be discussed later.

The Agglomerative Clustering, a well-performed algorithm for hierarchical clustering, is used to classify the districts for the convince of visualization and analysis. The algorithm can recursively merge the pair of clusters that minimally increase the linkage distance, and obtain the result eventually.

The metric we chose to measure the quality of clustering is a distortion function, it is defined as follows:

$$J(c, u(t)) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \|x_i(t) - u_{c^{(i)}}(j)\| \qquad (19)$$

where m is the number of the districts, n is the number of time periods, $x_i(t)$ is a time series of the AttractRank value of the i^{th} district, $c^{(i)}$ is the label of the i^{th} district, and $u_{c^{(i)}}(t)$ is a time series of mean AttractRank value of group $c^{(i)}$.

TABLE III
THE VALUES OF METRIC WITH DIFFERENT NUMBERS OF CLUSTERS.

Number of Clusters	4	5	6	7	8	9	10
Value of Metric	19.17	17.79	16.96	16.29	15.47	14.35	14.02

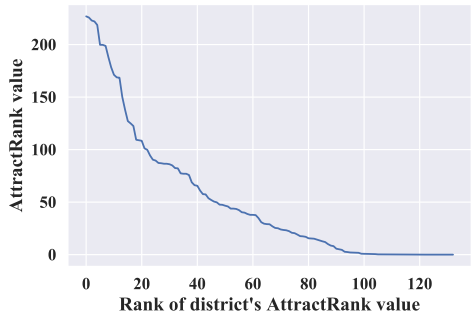
As shown in Eq. 19, the value of metric decreases as the number of clusters increases. To balance the convenience of discussion with the proper number of clusters and the metric of clustering, we performed a tuning to find the proper number of cluster centers for clustering. As shown in the table III, with the number of cluster centers increasing, the value of metric decreases. However, when the number of cluster centers is 8, the reduction speed is slow. For the convenience of discussion, we use 8 as the number of cluster centers for clustering. Then the districts of Guangzhou are divided into 8 groups with respect to their time series of AttractRank value. Figure 6(b) shows the results of clustering. The result of aggregating is consistent with the study of urban planning that cities can be divided into different sections with different functions. And their attraction can be various by the hour in a day repeatedly.

We also plot the districts on the map in different colors with respect to the labels obtained above. As shown in Figure 7(a), a strong spatial correlation within one group can be clearly observed, where the districts in the same group reveal geographical proximity between each other. What's more, as shown in Figure 7(b), some districts are much more attractive than others and influence the districts which are spatially closed to them. The districts plotted in red can strongly attract the taxis, and the districts closed to them possess roughly 50% less attraction, but it seems that most of the remaining districts possess little attraction to taxi drivers. Those are supported by the fact that Guangzhou is a highly centralized city.

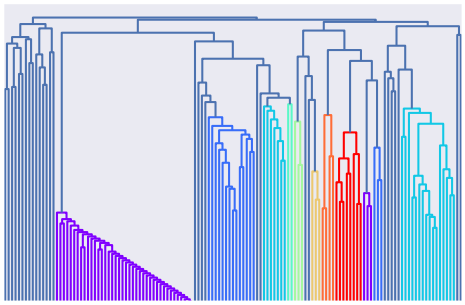
C. Exploration of Attraction Flowing Between Groups

With the purpose of studying how the attraction flows between groups, we plot a stacked area chart of group attraction. As shown in Figure 8, we can realize the the proportion of group attraction changes over time, it's possible for us to explore the patterns of how the attraction to taxis flows among those groups and to show hidden characteristics of lifestyle in Guangzhou.

Above all, let's discuss the explanation of labels. The eight groups can be divided into 5 types. Group one includes 69 districts, which consists of more than half of the districts, but none of them shows attraction to taxi. Group one can be treated as a group of residential districts because of the low attraction to taxis. Group two contains two districts, namely Guangzhou

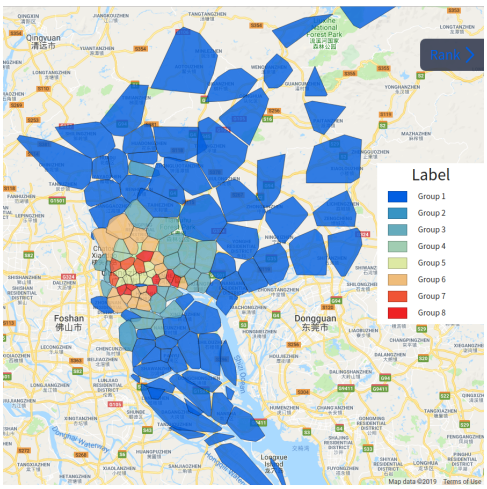


(a) Distribution of mean AttractRank values per hour of each district. Y-axis represents the mean AttractRank value per hour of a district, while x-axis is the order of districts sorted by their AttractRank value.

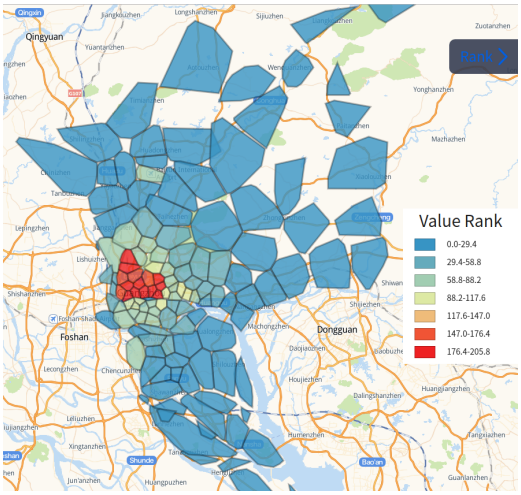


(b) Demonstration of results obtained in district hierarchical clustering.

Fig. 6. Illustration of classifying districts. (a) shows the distribution of mean AttractRank values per hour of each district. (b) is the illustration of the hierarchical clustering of districts. The features of the districts are the time series of AttractRank values of those districts. Each bottom dot represents one of the districts, and the hierarchical tree shows how the districts are recursively merged into one cluster that minimally increases the linkage distance.



(a) Distribution of labels obtained with hierarchical clustering.



(b) Distribution of the mean AttractRank value per group.

Fig. 7. Some illustration of the distribution of AttractRank value. (a) shows the distribution of labels obtained with hierarchical clustering and (b) shows the distribution of the mean AttractRank value per hour of each district.

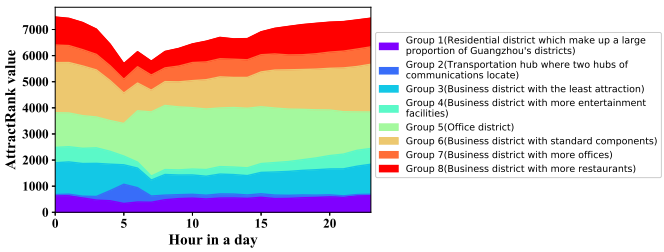


Fig. 8. Stacked area chart of group attraction: Its coloration indicates the proportion of the AttractRank value of each group. The AttractRank value of one group is defined as the summation of attraction over all districts in that group, where the attraction of districts is the mean AttractRank value per day.

South Station and Guangzhou Baiyun International Airport. This type of group can be considered as the transportation hub. According to the pattern of attraction changing over time, districts in group three, six and eight are of the same kind and group five has a converse pattern. It is a convincing explanation that they are, respectively, groups of office districts and business districts because they are more attractive to taxis alternately. Districts in group seven can be a group of some districts having functions of business district and office district at the same time. For the AttractRank value of four groups of business districts, they share a tendency to reach the valley at 6:00 and peak at 24:00. However, the attraction of group 6 changes more dramatically than that of group 3. Besides, the attraction of group 8 was higher than the one of group 6 at 18:00 while group 4 has additional peaks at 3:00 and 23:00. Based on the phenomenon mentioned above, we hypothesize the composition of each group and show it in the legend.

As explained above, the eight groups can be divided into 5 types, namely the residential district, the transportation hub, the business district, the office district, and the mixture of business district and office district. As mentioned before, the sum of AttractRank value per hour over every district is the number of operating records during that hour, which is corresponding to the line at the top. We can observe that the value of the line at the top varies from 5500 to 8000. The number of taxi records decreases gradually between 0 am and

5 am. Removing the attraction of group two, the sum will decline even more sharply. After that, an impulsion of going to work appears from 5 am to 6 am. The attraction of city districts in group five rises rapidly, and the attraction of those districts decreases gradually after 7 am. The proportion of attraction of office districts becomes smaller and smaller after 7 am. The proportion of the attraction of business districts gets larger and larger at the same time.

We can draw the conclusion that many people in Guangzhou city go to work during 6 am and 2 pm every day. The trend of summation of groups' attraction can be roughly described as a wave peaking at 0 am and reaching a trough at 5 am. That's supported by the knowledge that Guangzhou is famous as a never-sleeping city. It confirms that our algorithm can be used to explore the lifestyle of different city districts.

VIII. Conclusions

In this paper, we for the first time propose a novel method termed AttractRank for calculating the attraction of districts using taxi operating data. As distinguished from other district's indexes such as GDP and population of districts, the attraction index generated by AttractRank can represent the interaction between those districts. And then, an application is designed to explore the district's attraction of Guangzhou as follows.

- First, 134 districts are generated by applying the Constrained K-means clustering algorithm to Guangzhou taxi operating data.
- Based on the 134 districts, and taxi operating data, the district's AttractRank value for every period is calculated with AttractRank.
- Based on the time series of district's AttractRank values, hierarchical clustering is applied to explore the characters of these districts. By discussing the AttractRank values of different districts, we can reasonably predict the flow of people and further infer the function of each region, which is an important part of urban computing. The districts of Guangzhou can be roughly divided into five types, namely the residential district, the transportation hub, the business district, the office district, and the mixture of the business district and office district. Taxis run among them regularly every day.
- Four important time points are realized. They are the departure time of the train or plane, time to go to work by taxi, entertainment time after hours and time when people go home.

The extensive exploration suggests that our method has potential applications in urban planning and urban data mining.

References

[1] L. Zheng, D. Xia, X. Zhao, and W. Liu, "Mining trip attractive areas using large-scale taxi trajectory data," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Guangzhou, China, December 12-15, 2017*. IEEE, 2017, pp. 1217–1222.

[2] X. Mou, F. Cai, X. Zhang, J. Chen, and R. Zhu, "Urban function identification based on POI and taxi trajectory data," in *The 3rd International Conference on Big Data Research, ICBDR 2019, Cergy-Pontoise, France, November 20-22, 2019*. ACM, 2019, pp. 152–156.

[3] Y. Sun, H. Zhu, F. Zhuang, J. Gu, and Q. He, "Exploring the urban region-of-interest through the analysis of online map search queries," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Y. Guo and F. Farooq, Eds. ACM, 2018, pp. 2269–2278.

[4] M. Alhazzani, F. Alhasoun, Z. Alawwad, and M. C. González, "Urban attractors: Discovering patterns in regions of attraction in cities," *CoRR*, vol. abs/1701.08696, 2017.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[6] T.-H. Wen *et al.*, "Geographically modified pagerank algorithms: Identifying the spatial concentration of human movement in a geospatial network," *PloS one*, vol. 10, no. 10, p. e0139509, 2015.

[7] S. K. Mohanty, A. Dash, R. S. Mishra, and B. Dehury, "Economic development in the districts of india," in *The Demographic and Development Divide in India*. Springer, 2019, pp. 467–507.

[8] I. A. T. Hashem, V. Chang, N. B. Anuar, K. S. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *Int. J. Inf. Manag.*, vol. 36, no. 5, pp. 748–758, 2016.

[9] M. V. Moreno, F. Terroso-Sáenz, A. González-Vidal, M. Valdés-Vela, A. F. Skarmeta, M. A. Zamora, and V. Chang, "Applicability of big data techniques to smart cities deployments," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 800–809, 2016.

[10] J. Wang, J. Zhang, X. Yuan, Y. Tang, H. Hao, Y. Zuo, Z. Tan, M. Qiao, Y. H. Cao, L. Ai *et al.*, "Air quality data analysis and forecasting platform based on big data," in *2019 Chinese Automation Congress (CAC)*. IEEE, 2019, pp. 2042–2046.

[11] F. Xia, A. Rahim, X. Kong, M. Wang, Y. Cai, and J. Wang, "Modeling and analysis of large-scale urban mobility for green transportation," *IEEE Trans. Industrial Informatics*, vol. 14, no. 4, pp. 1469–1481, 2018.

[12] A. Galletta, L. Carnevale, A. Bramanti, and M. Fazio, "An innovative methodology for big data visualization for telemedicine," *IEEE Trans. Industrial Informatics*, vol. 15, no. 1, pp. 490–497, 2019.

[13] H. Han, Z. Liu, Y. Hou, and J. Qiao, "Data-driven multiobjective predictive control for wastewater treatment process," *IEEE Trans. Industrial Informatics*, vol. 16, no. 4, pp. 2767–2775, 2020.

[14] L. Li and S. X. Ding, "Performance supervised fault detection schemes for industrial feedback control systems and their data-driven implementation," *IEEE Trans. Industrial Informatics*, vol. 16, no. 4, pp. 2849–2858, 2020.

[15] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 2012, pp. 186–194.

[16] W. Xing and A. A. Ghorbani, "Weighted pagerank algorithm," in *2nd Annual Conference on Communication Networks and Services Research (CNSR 2004), 19-21 May 2004, Fredericton, N.B., Canada*, 2004, pp. 305–314.

[17] A. El-Geneidy and D. Levinson, "Place rank: valuing spatial interactions," *Networks and Spatial Economics*, vol. 11, no. 4, pp. 643–659, 2011.

[18] J. Xu, A. Li, D. Li, Y. Liu, Y. Du, T. Pei, T. Ma, and C. Zhou, "Difference of urban development in china from the perspective of passenger transport around spring festival," *Applied geography*, vol. 87, pp. 85–96, 2017.

[19] T. Anwar, C. Liu, H. L. Vu, and M. S. Islam, "Roadrank: Traffic diffusion and influence estimation in dynamic urban road networks," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1671–1674.

[20] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, 2001, pp. 577–584.

[21] G. Sun, V. Chang, G. Yang, and D. Liao, "The cost-efficient deployment of replica servers in virtual content distribution networks for data fusion," *Information Sciences*, vol. 432, pp. 495–515, 2018.

[22] V. Chang, "Towards data analysis for weather cloud computing," *Knowl. Based Syst.*, vol. 127, pp. 29–45, 2017.

[23] T. Soler and L. D. Hothem, "Coordinate systems used in geodesy: Basic definitions and concepts," *Journal of surveying engineering*, vol. 114, no. 2, pp. 84–97, 1988.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006.

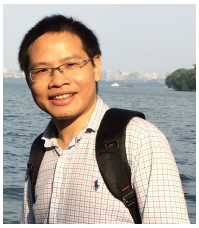
[25] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
[26] M. Bianchini, M. Gori, and F. Scarselli, "Inside pagerank," *ACM Transactions on Internet Technology (TOIT)*, vol. 5, no. 1, pp. 92–128, 2005.
[27] P. Berkhin, "A survey on pagerank computing," *Internet Mathematics*, vol. 2, no. 1, pp. 73–120, 2005.
[28] A. N. Langville and C. D. Meyer, "Deeper inside pagerank," *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.
[29] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.
[30] R. Serfozo, *Basics of applied stochastic processes*. Springer Science & Business Media, 2009.

Guangqiang Xie received the Ph.D. degree in Control Science and Engineering from Guangdong University of Technology, Guangzhou, China, in 2013. He is currently a Professor in Guangdong University of Technology and the Vice-Dean of School of computer. His research interests involve control of multi-agent systems and data mining.

Runpeng Zhang is an undergraduate in computer science, College of Computer, Guangdong University of Technology, China. He is currently with QG Technology Innovation Studio, Guangdong University of Technology, China. His research interests include data mining and machine learning.

Yang Li received the Ph.D. degree from Guangdong University of Technology, Guangzhou, China, in 2013. She is currently an Associate Professor with the School of Computer, Guangdong University of Technology, Guangzhou, China. Her research interests include privacy preserving and data mining.

Ling Huang received her undergraduate and master degree in 2009 and 2013 respectively from South China University of Technology. She is currently working toward the PhD degree at Sun Yat-sen University. She has published near 20 papers in international journals and conferences such as IEEE TKDE, IEEE TCYB, IEEE TNNLS, ACM TKDD, IEEE/ACM TCBB, Pattern Recognition, KDD, AAAI, IJCAI and ICDM. Her research interest is data mining.



Chang-Dong Wang received the Ph.D. degree in computer science in 2013 from Sun Yat-sen University, Guangzhou, China. He is a visiting student at University of Illinois at Chicago from Jan. 2012 to Nov. 2012. He joined Sun Yat-sen University in 2013, where he is currently an associate professor with School of Data and Computer Science . His current research interests include machine learning and data mining. He has published over 120 scientific papers in international journals and conferences such as IEEE TPAMI, IEEE TKDE, IEEE TCYB, IEEE TNNLS, ACM TKDD, IEEE TSMC-Systems, IEEE TSMC-C, KDD, AAAI, IJCAI, CVPR, ICDM, CIKM and SDM. His ICDM 2010 paper won the Honorable Mention for Best Research Paper Awards. He won 2012 Microsoft Research Fellowship Nomination Award. He was awarded 2015 Chinese Association for Artificial Intelligence (CAAI) Outstanding Dissertation. He is an Associate Editor in Journal of Artificial Intelligence Research (JAIR).

Hao Yang is an undergraduate in computer science, Guangdong University of Technology, China. He is currently with QG Technology Innovation Studio, Guangdong University of Technology, China. His research interests include data mining and reinforcement learning.

Jiahao Liang is an undergraduate in the Internet of things engineering, Academy of Automation, Guangdong University of Technology, China. He is currently with QG Studio, Guangdong University of Technology, China. His research interests include data mining and machine learning.