An Empirical Analysis of Public Transit Networks Using Smart Card Data in Beijing, China

Most existing studies on public transit network (PTN) rely on either small-scale passenger flow data or small PTN, and only traditional network parameters are used to calculate the correlation coefficient. In this work, the real smart card data (SCD) (when passenger tap in and tap out a station) of over eight million users is used as a proxy of passenger flow to dynamically explore and evaluate the structure of large-scale PTNs with tens of thousands of stations in Beijing, China. Three types of large-scale PTNs are generated, and the overall network structure of PTNs are examined and found to follow heavy-tailed distributions (mostly power law). Further, three traditional centrality measures (i.e., degree, betweenness and closeness) are adopted and modified to dynamically explore the relationship between PTNs and passenger flow. Our findings show that, the modified centrality measures outperform the traditional centrality measures in estimating passenger flow.

Keywords: Smart Card Data (SCD), Public transport systems, Network Centrality, Correlation analysis, Passenger Flow.

1 Introduction

With urbanisation accelerates, there are increasing challenges for the public transportation system (PTS), such as traffic congestion, air pollution and energy consumption. A high percentage of people in the city heavily relies on PTS, of which the efficiency plays an important role in people's travel behaviour, especially in the developing areas such as China that associated with huge population migrated to cities (Si et al. 2016; Dimitrov & Ceder 2016; Shanmukhappa et al. 2018; T. Zhang et al. 2018; Yan et al. 2018). Public Transit Networks (PTNs), i.e., stations and lines of subway and bus, are one of the most important infrastructures in PTS, and a better understanding of the structure and efficiency of PTNs, in particular, its relationship with passenger flow, will potentially benefit decision-makers, transportation planners and related urban studies (Gao et al. 2013; Tang et al. 2013; Zhao et al. 2017). There are many existing studies dedicated to the analysis of PTNs from different perspectives, and one of the basic assumptions underlying most current work is the passenger flow can reveal and reconstruct the urban structure. Besides, the evolution of PTNs is strongly linked to the process of urbanisation, which can be delineated by the passenger flow on PTNs (Von Ferber et al. 2009; Xu et al. 2016; Zhang et al. 2016; H. Zhang et al. 2018). The complex network theory has been widely applied in different fields, such as social science work, and transportation is one of the most important applications (Scott 1988; Kim & Hastak 2018).

Most existing studies on PTNs and its relationship between traffic flow can be generally characterized into two groups. In the first group, complex network theory is used to model PTN and examine the structure from the perspective of network analysis. For example, (Von Ferber et al. 2005) found that the degree distributions of the three PTNs of Berlin, Düsseldorf, and Paris demonstrated a power law distribution (scale-free network). (Sienkiewicz & Hołyst 2005) carried out fundamental research to analyse the statistics of PTNs and investigated the

topology of the network structure based on space L and space P (refer to (Dimitrov & Ceder 2016) for more detail). They analysed the clustering, assortative characteristics and betweenness of PTNs of 22 cities, each of which comprises bus stations from minimum 152 to maximum 2,881. The assessment findings for the network topology indicate that the degree distribution in space L was followed a power law, while degree the distribution for space P was presented as an exponential function. Meanwhile, (Feng et al. 2016) examined directed and weighted bus transit networks from a view of complex networks. The empirical properties of the bus transit of Harbin were reported that the cumulative distributions of weighted degree, degree, number of routes that connect to each station, and node weight (peak-hour trips at a station) uniformly follow the exponential law. (Li et al. 2018) have evaluated the vulnerability of PTSs from the perspective of topological properties of the PTNs and attack tolerance. The abovementioned studies either suffers from small scale (e.g., 152 stations) or do not use real traffic flow to examine the relationship between the structure of PTNs.

In the second group, the relationship between PTNs and passenger flow is examined. For instance, (Luo et al. 2019) conducted a regression models to reveal the correlative relationship between passenger flow distribution and the conventional network properties for the train system in Hague and Amsterdam cities. In Luo's study, the conventional network centrality measures were computed based on two topologies of the network, namely, space L and space P. The conventional network centrality is based on the topology of the network only. However, one of the limitations is the passenger flow is collected based on survey data, which suffers from being small scale and not representative in terms of real passenger flow, this is all implicit in the low-frequency city. The high- and low-frequency are introduced in (Batty 2018).

With the advent of GPS-enabled floating cars such as taxis and buses, the road traffic data become more wide coverage, which provides opportunities to evaluate and analyse the traffic flow (Kerner et al. 2005; Tang et al. 2015; Liu et al. 2018). Regarding the relationship between traffic flow and traffic network (mainly road network), there are many studies as well. There are many studies dedicated to examining the structure of traffic networks using different models, e.g., (Jiang et al. 2011; Mukherjee 2012; Tian et al. 2016; Wang et al. 2017). Further, the relationship between traffic flow and urban traffic network is explored (Kazerani & Winter 2009a; Kazerani & Winter 2009b; Ye et al. 2016). Among these studies, traffic flow is collected by several methods, including Annual Average Daily Traffic (AADT), street survey, video images, roadside detectors, and GPS trajectory data. Although the abovementioned studies focused on the road network instead of PTNs with fixed routes, the methods can be applied to PTNs in this study. For example, a regression model proposed by (Pun et al. 2019) by combining five centrality measures. Meanwhile, with the rapid development of technology and the emergence and use of smart cards in the payment of transportation fees, smart card data offers an unprecedented opportunity to analyse and evaluate passenger flow and to understand travel behaviour with high accuracy, as compared to data collected by surveying. Besides, the smart card data can be used to detect the movement of the passengers within the high-frequency city (Batty 2018). In the high-frequency city, estimating and predicting the passenger flow is one of the most critical issues due to its crucial role in transportation planning and management.

We argue that, 1) most existing studies rely on either small-scale passenger flow data or small PTN; 2) only traditional network parameters (e.g., degree, betweenness, closeness, etc.) are used to calculate the correlation. In this study, the passenger flow is derived from smart card data (SCD) in Beijing, China. There are three types of PTNs are modelled in the study area: subway network, bus network, and the integration between subway and bus networks based on transfer relations that are called the PT network. Besides, the PTNs are modelled in two different representations, namely, station-based and line-based representations. It is worth to be noted that we used a large-scale network in comparison with previous studies (see Table1). The structure of the PTNs in the study area is examined from the perspective of a complex network. Then the correlation between passenger flow and the PTNs as mentioned earlier, are examined on an hourly basis using conventional centralities (degree, closeness, and betweenness) in the network analysis based on the topological properties. To better understand the high-frequency city, we proposed the modified centrality measurements considering both the topological and geometrical properties of PTNs to investigate further the relation between PTNs and passenger flow, of which the correlation analysis demonstrates acceptable performance.

Table 1: An illustration of the element of the generated Networks

Network Representation Type	Network Element	Subway System	Bus System	PT (Bus + Subway) System
Station-based network	Nodes	278	35674	35952
	Edges	700	903813	935551
Line-based network	Nodes	36	1574	1610
	Edges	428	281874	306422

This paper is organized as follows. The overview of the methodological framework, network modelling and network analysis are presented in section 2. Section 3 describes the study area, experimental data, results and discussion, followed by the conclusion in the last section.

2 Methodology

In this study, three traditional and modified network centralities (degree, closeness and betweenness) for analysing the relationship between structural features of PTN and passenger flow are proposed as shown in Figure 1.

The framework consists of three main stages, as follows:

• **Modelling Public Transit Network**: The primary stage is to create networks from the PTN individually and collectively, namely bus system, subway system and the two systems together. Each PTN was modelled as station- and line-based network.

• **Analysis of network parameters**: In this stage, both traditional and modified centrality measures of each network are computed for station- and line-based network of bus, subway and merged network.

• **Analysing passenger flow**: In this stage, the passenger flow was computed based on the smart card data for each network and the relationship between PTN is analysed.

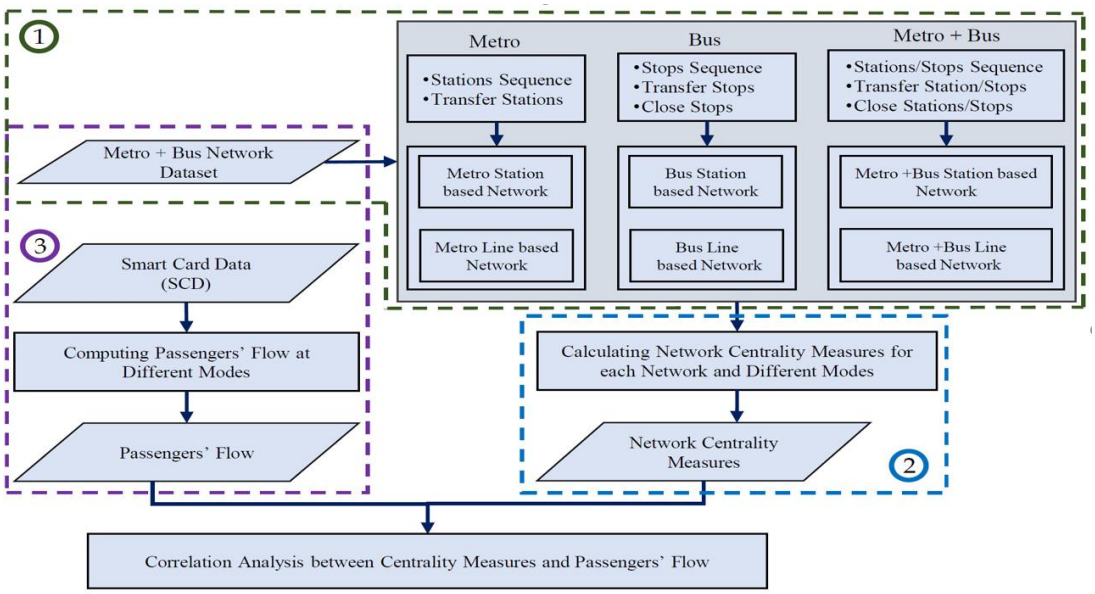

Figure 1:The adopted framework for network analysis

2.1 Modelling Public Transit Networks

In this paper, each of the PTNs is translated into a directed graph $G = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges (links) as demonstrated in Figure 2. In the current work, considering the spatial and transfer attributes of the nodes, a graph G is represented by $G = (V(X, Y, T), E)$ where V and E are described as:

$$V = \{v_i(x_i, y_i, t_i): i = 1, 2, \ldots, p;\ x_i = longitude, y_i = latitude, t_i = number\ of\ transfer\ edges\} \quad (1)$$

$$E = \{e_{ik} \rightarrow (v_i(x_i, y_i, t_i), v_k(x_k, y_k, t_k)) \forall (v_i(x_i, y_i, t_i), v_k(x_k, y_k, t_k)) \in V: i = k = 1, 2, \ldots, p\} \quad (2)$$

Where p is the number of nodes in the network. In the next term, $v_i(x_i, y_i, t_i)$ is represented as a particular node and determined by its longitude, latitude and number of transfer edges at the node. e_{ik} is a specified edge that connected two nodes v_i and v_k.

The transforming PTSs into the two representations of the graph described below in more detail.

2.1.1 Station-based network's structure

In this network, the nodes represent the subway station, bus stops or both of them together in the subway network, bus network, subway with bus (PT) network respectively as shown in Figures 2B, 2C, 2D successively. This representation is only the L-space representation in the literature, but the name is different.

Subway stations-based network's structure

As mentioned early, the nodes of the subway stations-based network are the subway stations. Further, the edges are located between every two successive nodes if they are on the same

line, called connected edges, and additional edges are linking each pair of different intersection lines of the transfer station, called within-station transfer edges. Figure 2C presents the corresponding subway stations-based network representations. So, this directed graph consisted of 278 nodes and 700 edges.

Bus stations-based network's structure

This Network comprised of 35674 nodes and 903813 edges, where the nodes are the bus stops. The edges of this graph consisted of connected edges and foot transfer edges. Where the connected edge is a link exists between two nodes if they are on the same line. The foot transfer edge exists between two stops if the distance based on the walkable street network less than a bus threshold distance. The bus threshold distance is the maximum allowable walking distance for transferring between bus stops and equivalent to $10\ minutes$ walking time with assuming the walking speed $70\ m/min$ for transferring between bus stops. An example of this graph is shown in Figure 2B.

PT stations-based network's structure

All of the subway stations and the bus stops are one of the basic units for the PT network, which are represented the nodes. The edges of this network comprised of connected edges, within-station transfer edges and foot transfer edges. All components of the edges of this graph are described above, but there is a little difference related to foot transfer edges. Since the density of subway station is considerably lower than bus stops, the threshold for transferring between subway stops and bus stops is equivalent to $15\ minutes$ walking time with assuming the walking speed $70\ m/min$ for transferring between bus stops. This directed graph consists of 35952 nodes and 935551 edges. A sample of this graph is given in Figure 2D.

2.1.2 Line-based network's structure

The representation of line-based network represents lines in each system as nodes and a relationship between lines as edges. This way, the transfer edge between lines establish links between the nodes. The corresponding line-based network representations are also shown in Figures 2E, 2F, 2G.

Subway line-based network's structure

In the subway line-based network, the subway lines are considered the nodes. The edges are located between the lines that have a shared transfer station. This network comprises 36 nodes and 428 edges. This network is shown in Figure 2F.

Bus line-based network's structure

This Network comprised of 1,574 nodes and 281,874 edges, where the nodes are the bus lines. The edges of this graph are links exist between two nodes if they have foot transfer edges. An example of this graph is illustrated in Figure 2E.

PT line-based network's structure

All of the subway lines and the bus lines are represented the nodes. The edges of this network are links that exist between two nodes if they have foot transfer edges or common transfer

station. There are 1610 nodes and 306422 edges in this network. Figure 2G shows a sample of this network.

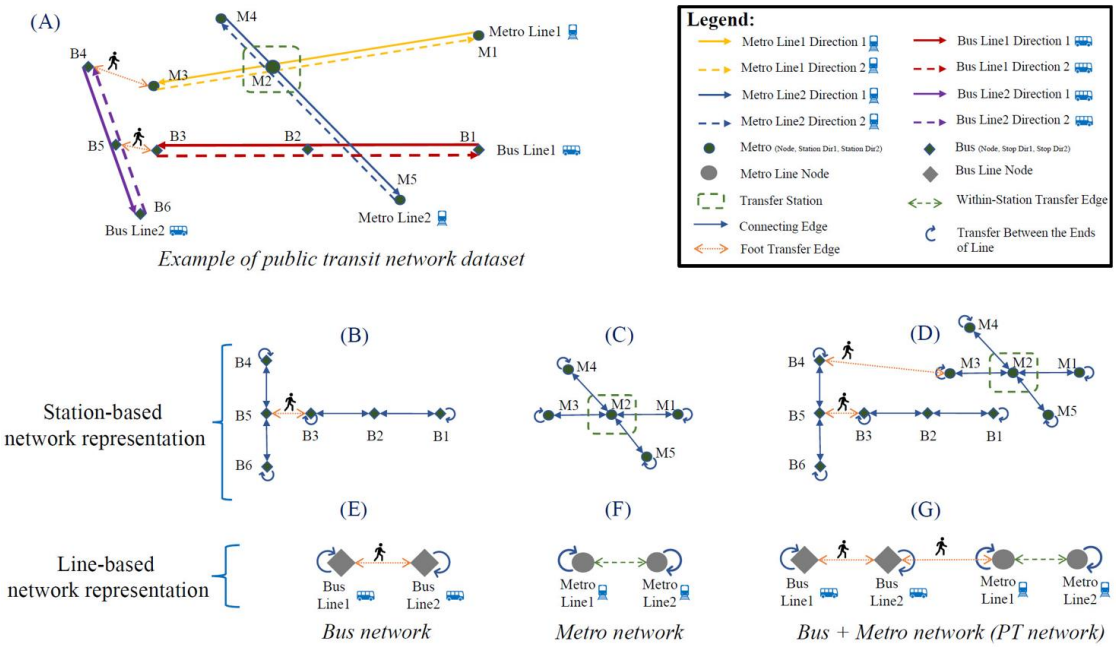

Figure 2: Modelling the graphs based on different representations

2.2 Analysis of the structure of PTNs

In this section, the directed graph for each system is analysed with station-based network and line-based network representations for Beijing. An extensive discussion of both traditional and modified techniques for three of network parameters, namely degree, closeness and betweenness are given in this section. These centrality measures contribute to illustrate the fundamental properties of networks by characterising the relative importance of a node within the graph. These three centrality measures were selected because each of these measures gave different values that can be distinguished for all nodes compared to the other centrality measures (eigenvector, page rank, and clustering), which have a very slight difference in their values. Furthermore, the degree centrality represents the number of adjacent nodes, so it suitable to analyse the correlation between the degree centrality and the passenger flows. It is likely that the higher number of nearest stations the greater passenger flow. Moreover, the betweenness and closeness centrality measures are to express the proximity of a node to their counterparts so that it represents the shortest path. Thus, it is best suited to represent the transfer stations, which are also likely to attract more passengers.

2.2.1 Traditional network centrality measures

Centrality is essential to understand the structural properties of the public transit network. There are three widespread centrality measures, namely degree, betweenness, and closeness.

Degree

Degree centrality is the most basic yet crucial parameter in network analysis. Degree centrality is a local measure and is also called connectivity in space syntax (Huang et al. 2016). As

above mentioned, the network is represented as a connectivity graph $G(V, E)$. Let v_i and v_k are any two nodes of G. If v_i and v_k are connected by an edge directly, they considered adjacent. The number of edges adjacent to the node is indicated the degree centrality of the node. The nodes with higher degree are more central. The degree centrality of v_k is defined by:

$$c_D(v_k) = \sum_{i=1}^{p} a(v_i, v_k) \qquad \text{where } a(v_i, v_k) = \begin{cases} 1, & if\ v_i, v_k\ are\ adajcent \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The maximum possible size of $c_D(v_k)$ is $p - 1$ (Hage & Harary 1996). Therefore, a node's degree centrality $c_D'(v_k)$ is the ratio

$$c_D'(v_k) = \frac{\sum_{i=1}^{p} a(v_i, v_k)}{p-1} \tag{4}$$

Closeness

Closeness centrality indicates the nearness of a node to all other nodes. The closeness is the sum of the lengths of the shortest paths between one node and all other nodes of a connected graph. The closeness centrality of v_k is determined as follows:

$$c_C(v_k)^{-1} = \sum_{i=1}^{p} d(v_i, v_k) \tag{5}$$

Where $d(v_i, v_k)$ is the shortest-path distance between v_i and v_k. As (Freeman 1978) indicates, the closeness measure is effectively a measure of inverse centrality, since it grows as nodes become more distant. (Beauchamp 1965) has already solved this problem; closeness is normalised by the sum of minimum possible distances $p - 1$. Hence the closeness centrality of the node is defined by:

$$c_C'(v_k) = \frac{p-1}{\sum_{i=1}^{p} d(v_i, v_k)} \tag{6}$$

So, by implementing the Equation (6), higher values of closeness indicate higher centrality.

Betweenness

Betweenness centrality is one of the essential centrality measures for analysing network structure and defines as the sum of the fraction of all-pairs shortest paths that pass through v_k:

$$c_B(v_k) = \sum_{s,t \in V} \frac{\sigma(s, t|v_k)}{\sigma(s, t)} \tag{7}$$

Where s is the source node, t is the target node, $\sigma(s, t)$ is the number of shortest (s, t)-paths, and $\sigma(s, t|v)$ is the number of those paths passing through some node v_k other than s, t. As all network is directed graph, the maximum possible betweenness value of a node, $c_B(v_k)$, is $(p - 1)(p - 2) = p^2 - 3p + 1$. The betweenness $c_B'(v_k)$ is the ratio of $c_B(v_k)$ to this maximum expression, so that:

$$c_B'(v_k) = \frac{c_B(v_k)}{p^2 - 3p + 1} \tag{8}$$

Equations 4, 6, and 8 are used as a conventional technique for centrality measures. The conventional technique is used to capture the topological properties of PTNs only.

Furthermore, the topological and geometric attributes of the PTNs are considered in the developed modified technique for centrality measures.

2.2.2 modified technique for centrality measures

Modified centrality measures for station-based representation is developed and achieved by combining the conventional centrality measures and the transfer attribute related to a station (node). As for line-based representation, modified centrality measures are attained by integrating the conventional centrality measures and the transfer attribute attached to each line (node). Where the transfer attribute for both representations is the transfer edges at a node, refer to Equation 1. Because of the passenger flow is affected by the number of transfer edges that linked to the station (More transfer edges in the station, more people commute through this station), a modified centrality measure for v_k, $c_{v_k}^{mod}$, can be calculated using the following formula:

$$c_{v_k}^{mod} = c_{v_k}^{conv} * (t_v v_k)^{\alpha} \quad (9)$$

Where $c_{v_k}^{conv}$ is the conventional centrality measures, t_{v_k} is the number of transfer edges at a node v_k, and α is an influencing factor. To obtain ideal value for α, the optimisation method proposed in (Zhao et al. 2017) is implemented. The objective function, which used in the optimisation method, is used to maximise the sum of correlation coefficients between centrality measures and the passenger flow at each time slot, where the day is divided to a 24-time slot.

3 Study area and data processing

3.1 Study area

The Beijing metropolitan region is a municipality situated in the north of China, as shown in Figure 3A. It is one of the most important capitals of countries in the world, one of the oldest countries in the world and one of the most important centres in the world in various disciplines at present. Beijing Municipality consists of 16 districts with an area of approximately 16,500 (Figure 3B). The Population at the end of 2016, according to the Beijing Municipal Bureau of Statistics and NBS Survey Office in Beijing, is 21.72 million residents with a growth rate of 0.1% (Beijing Municipal Bureau of Statistics and NBS Survey Office in Beijing 2017). Thus, the density of the population exceeds 1,300 residents per square kilometre. Beijing is China's largest hub of transportation. Its metropolitan and micropolitan regions have a well-designed system of transport. A total of 8.25 billion commuters using Urban Public Transit in 2016, a slight decrease of 2% compared with last year. Where 3.69 billion passengers were transported by Buses transit; Subway transit completed 3.66 billion passenger traffic, an increase of 10.2% on the same year (Beijing Transportation Research Centre 2017). In 2016, there were 18 operating lines for the subway network, 334 operating stations, 53 transfer stations, 554 kilometres of operating mileage, and 5,024 vehicles. Figure 3C shows the spatial distribution of subway system where the station points are coloured in dark green in a circular shape, and the subway lines are coloured in different colours, in which the label of each line are included in the figure. As for Bus Network, there were 876 operating lines in the city, which was the same as last year. The number of vehicles operated was 22,688 bus (Beijing

Transportation Research Centre 2017). (Figure 3D) describes the spatial distribution of stops and lines for public bus service respectively, in dark red in square shape and blue.

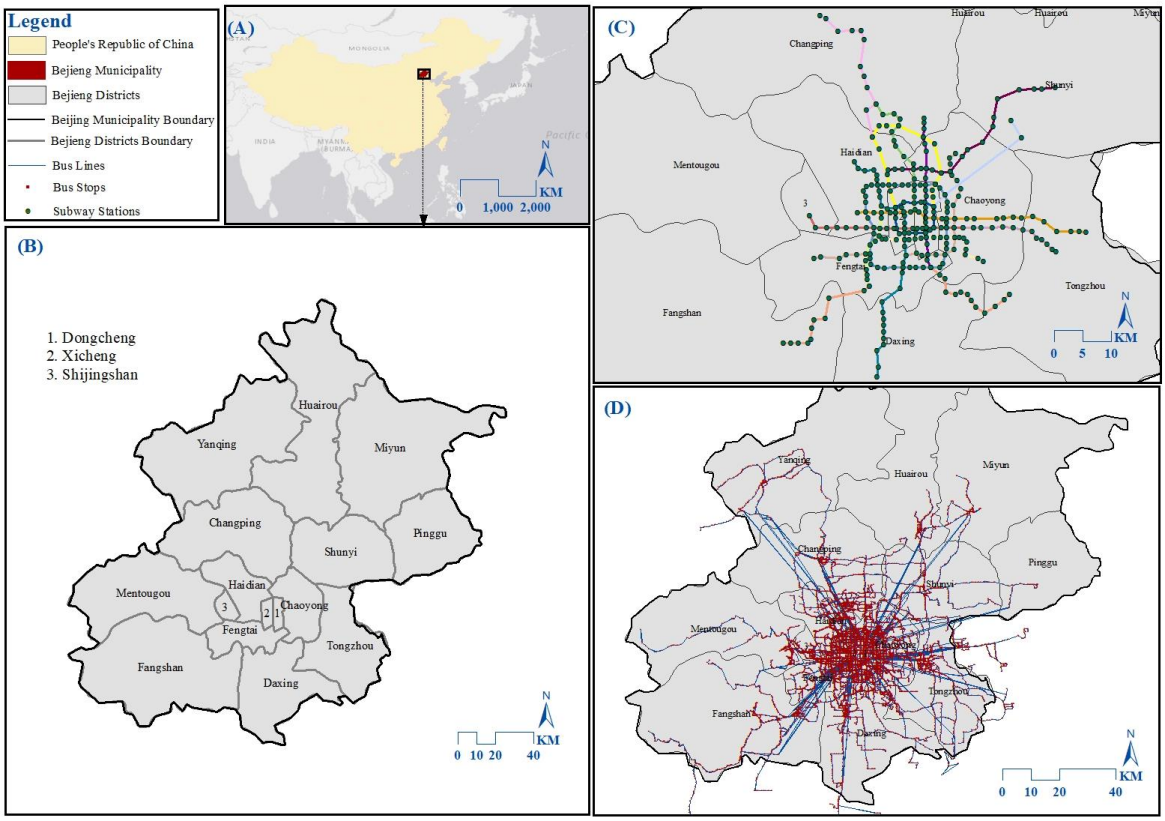

Figure 3: Location of the study area (Beijing) within China (A), Beijing Metropolitan Area (B), Spatial Distribution of subway Network in Beijing (C), and Spatial Distribution of Bus Network in Beijing (D)

3.2 Data processing

In this work, a network dataset of the public transit system in Beijing (Subway and Bus) and Smart card Data (SCD) are used to identify the extent to which the passenger flows are affected by the centrality measures of the network. The Subway network is developing appreciably every year with the opening of several new stations when completed in recent years. In 2016, The subway system consisted of 18 lines, 278 operating stations. It should be noted that the stations at the intersection of multiple lines, stations are referred to as several nodes. For instance, station (Jiangoumen) is a transfer station between line 1 and line 2, so this station considered as two nodes in the PTNs. Thus, the number of subway stations become 334 rather than 278 stations where there are 53 transfer stations. Beijing's Bus network is one of the city's largest, most used and cost-effective forms of urban and suburban transportation. In the used network dataset, the public bus service comprised of 803 and 1574 lines with considering and without considering the direction of travel respectively. Additionally, the bus system includes 42024 bus stations, considering that the transfer station represents the number of intersections, where the number of distinct stations is 35674. A sample of the available original and modified PTNs dataset is illustrated in Table 2 and Table 3 successively. The dataset includes the following main fields:

- DATASOURCE: The transit modes that passengers take where the AFC and DoubleIC indicate Subway and bus system respectively.
- LINE: The unique number of Public transit lines.
- DIR: The direction of travel (DOT) for Public transit line. In the original network dataset, there are three distinct value 0, 1, and 2, in which both 1 and 2 represent the going and returning direction in sequence for the bus network. However, at the same time, the DOT for the subway system is not defined; it was referred to as 0 as presented in Table 2. Therefore, the Dot of the subway was modified as existed in the bus network. The loop lines were taken into consideration in modifying the DOT, as shown in Table 3.
- STATION NAME: The station's name.
- STATIONNUM: The sequence number of stations per the direction of travel.
- LAT: The latitude coordinates of stations.
- LON: The longitude coordinates of stations.

Table 2: A sample of original Beijing's subway and bus network dataset

DATASOURCE	LINE	DIR	STATIONNAME	STATIONNUM	LAT	LON
AFC	Line 1	0	Dongdan	17	116.4125	39.9070
AFC	Line 1	0	Jianguomen	18	116.4288	39.9072
AFC	Line 2	0	Jianguomen	9	116.4288	39.9072
...	...	...	...	...	...	...
DoubleIC	18	1	Anyuan Dongli	17	116.4046	39.9817
DoubleIC	18	2	Anyuan Dongli	5	116.4052	39.9818
DoubleIC	409	1	Anyuan Dongli	19	116.4046	39.9817
DoubleIC	409	2	Anyuan Dongli	21	116.4052	39.9818

Table 3: A sample of modified Beijing's subway and bus network dataset

DATASOURCE	LINE	DIR	STATIONNAME	STATIONNUM	LAT	LON
AFC	Line 1	1	Jianguomen	18	116.4288	39.9072
AFC	Line 1	2	Jianguomen	6	116.4288	39.9072
AFC	Line 2	1	Jianguomen	9	116.4288	39.9072
AFC	Line 2	2	Jianguomen	10	116.4288	39.9072
...	...	...	...	...	...	...
DoubleIC	18	1	Anyuan Dongli	17	116.4046	39.9817
DoubleIC	18	2	Anyuan Dongli	5	116.4052	39.9818

Beijing Transportation Smart Card, known as Yikatong Card, was first issued on May 2006. It is a contactless smart card that can be applicable to almost all means of transport in the city of Beijing. Because of the smart cardholders could receive a high discount rate (i.e., 75% fare reductions for students and 50% fare reduction for regular passengers) and save time queuing at either ticket machines or ticket window, more than 90% of the ridership paid their trips by the smart card (Ma et al. 2017).

The Public transit smart card records provide for one week from Monday, April 11 to Sunday, November 17, in 2016. In this week there are four weekdays and two days on the weekend. This week is the week after the Qingming Festival. Beijing transportation system has two types of AFC system, the flat fares and distance-based fare systems. Transit riders pay a fixed rate for flat fare by tapping their smart cards on the card reader when entering; only check-in scans are necessary. For the distance-based fare system, the commuters should

tap their smart card when boarding and alighting (Ma et al. 2013). The flat fare system is generally used in the bus system, while most of the subway lines adopt the distance-based fare system (except for the Airport Line with a single fare of 25 RMB) (Zou et al. 2018). Table 3 shows some examples of the provided smart card data. The smart card data involves the following main fields:

- DATASOURCE: The transit modes that passengers take where the AFC and DoubleIC indicate Subway and bus system respectively.
- LINE: The unique number of Public transit lines.
- DIR: The direction of travel (DOT) for Public transit line, in which both 1 and 2 represent the going and returning direction.
- ON_TIME and OFF_TIME: The earliest and latest time of time slot.
- ON_LON and ON_LAT: The longitude and latitude of the station that commuters swipe their smart cards when they are boarding (origin), respectively.
- OFF_LON and OFF_LAT: The longitude and latitude of the station that commuters swipe their smart cards when they are alighting (Destination) respectively.
- DUR: The average of the duration of the trip between the same OD stations in the same time slot.
- NUM: The number of passengers who travelled from and to same OD stations in the same time slot (flow).

4 Results and discussion

4.1 Distribution of centrality measures

Whereas normalised conventional centrality measures are determined in the station-based representation using Equation 4, 6, and 8, The modified centrality measures are measured based on Equation 9.

Figure 4 shows subplots for degree distribution in both representations. It can be seen that most of the distributions of modified degree fitted by a power law, in which most nodes have only a few links but, by contrast, there exist some nodes which are extremely linked. The bus network in line-based representation described by a Gaussian function, in which the parameters read as: $\mu_1 = 0.19, \sigma_1 = 0.19$ and $\mu_2 = 0.08, \sigma_2 = 0.22$, respectively, for conventional and modified degree centrality. Furthermore, the distribution of conventional degree follows a Gaussian function for the subway network in both representation, and PT in the line-based representation. The parameters of each distribution exist in the legend of each plot.

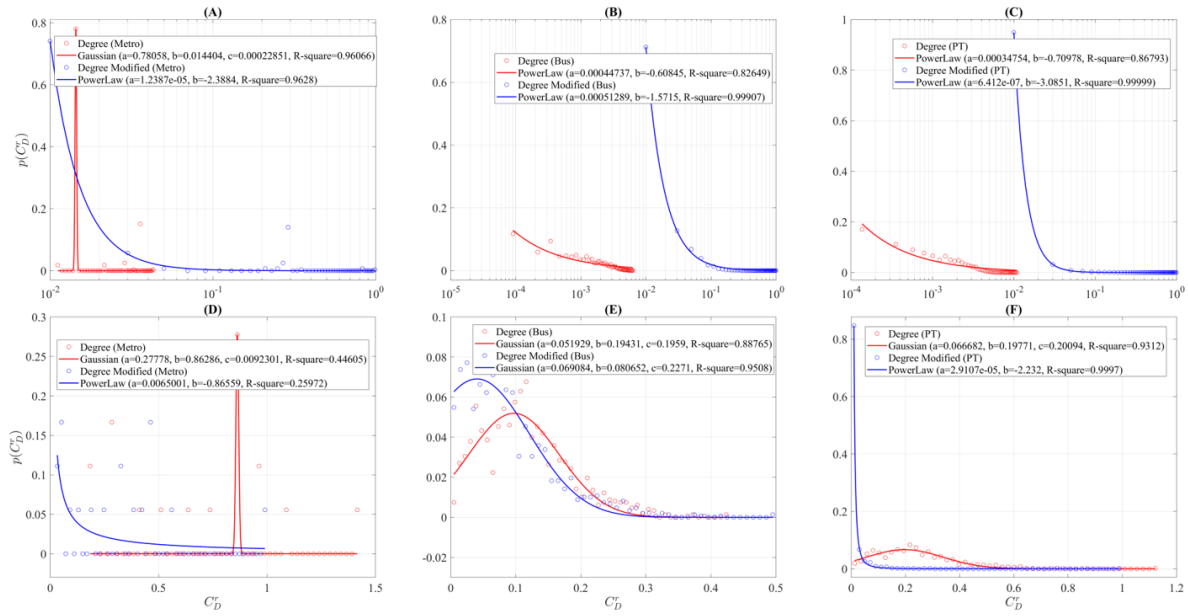

Figure 4: Degree distribution for (A) and (D) subway network, (B) and (E) bus network, and (C) and (F) PT network. Plots (A), (B), and (C) show the distributions in station-based representation in semi-log scale while plots (D), (E), and (F) in line-based representation in the linear scale

The closeness distributions are depicted in Figure 5. It can be found that all the conventional closeness is presented as a Gaussian function, while the most of modified centrality can be described by a power law except for bus network in the line-based representation, with parameters $\mu = 0.13, \sigma = 0.20$.

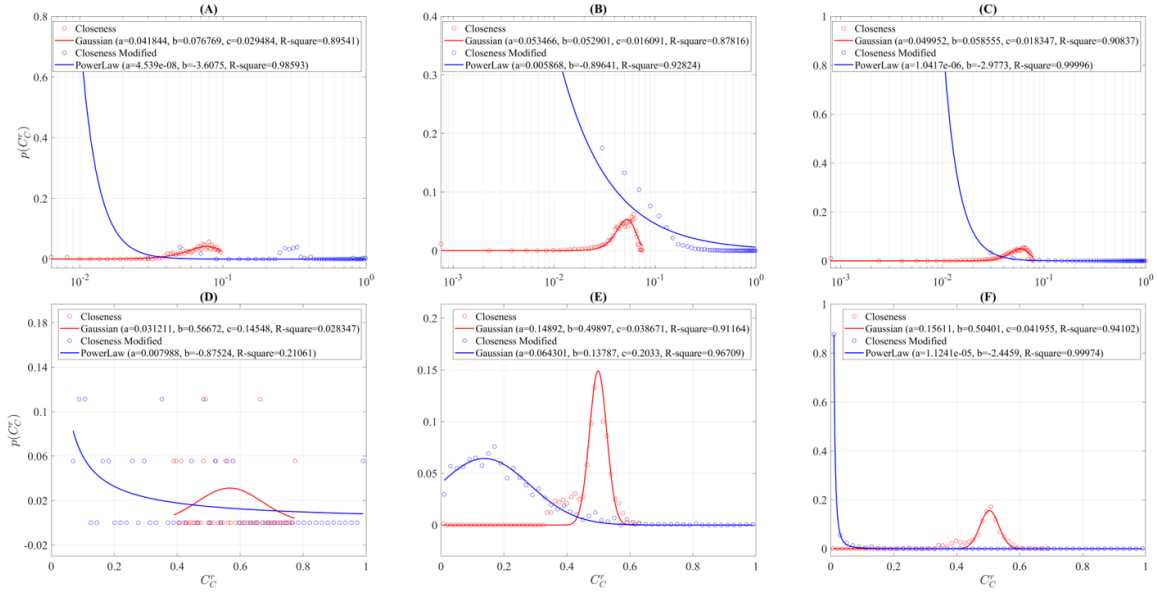

Figure 5: Closeness distribution for (A) and (D) subway network, (B) and (E) bus network, and (C) and (F) PT network. Plots (A), (B), and (C) show the distributions in station-based representation in semi-log scale while plots (D), (E), and (F) in line-based representation in the linear scale

Figure 6 illustrates with the exception of the conventional betweenness in the station-based representation, and the power law fit for all of both conventional and modified betweenness for the three networks in both representations. Values of the scaling parameter of power law are between 1.06 to 6.3.

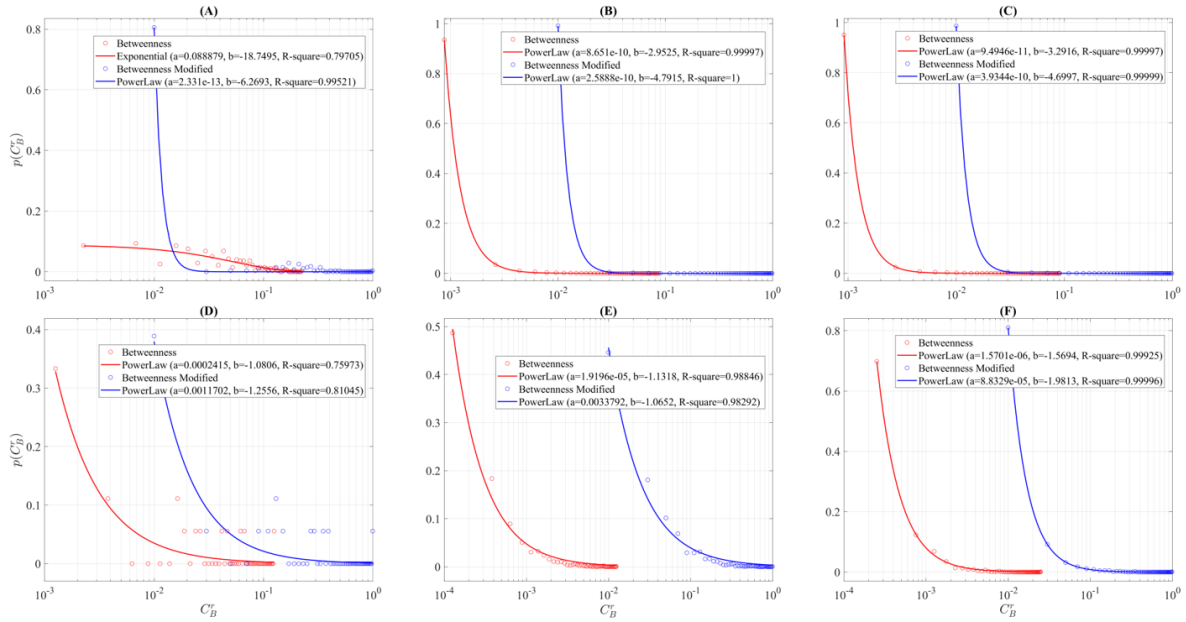

Figure 6: Betweenness distribution for (A) and (D) subway network, (B) and (E) bus network, and (C) and (F) PT network. Plots (A), (B), and (C) show the distributions in station-based representation in semi-log scale while plots (D), (E), and (F) in line-based representation in the linear scale

4.2 Passenger flow extraction

In this study, the smart card data is used to analysis the passenger flow. As stated before, the day is separated into 24-time slots. The passenger flow is computed per each time slot at each node for all network at both representations. The total ingoing flow is calculated through a station by aggregating numbers of passenger who swipe their smart card when they are boarding, where the outgoing flow is attained in the same way. It is expected the incoming flow to be equal to the outgoing flow. The correlation between ingoing and outgoing flow is demonstrated in Figure 7. In Figure 7A, it is quite visible that the majority of subway stations are located along the fitted line, which its slope parameter $a = 1.01$. The fitted line for the subway network is compatible with the stated expectation.

In contrast, the incoming flow and outgoing flow for bus stops are inconsistent with the expectation, as presented in Figure 7B. This inconsistent is explained by the fact that the density of bus stations is much larger than in the subway. Thus, it is normal for most passengers to use the same subway stations to travel a round-trip to their homes or work while the behaviour in the bus network may be different, where commuters can use different stations when returning from their work or home and vice versa.

The total passenger flows at each time slot are computed for the three PTSs. It should be noted that the total flow at both representations is equal. Hence the total flow for station-based representation is demonstrated in Figure 8. It can be seen that the passenger flow is a regular and predictable periodic mobility routine in the level of 7 days, which reflects the regular human activities, especially on the first four days, weekdays. It can be observed that sharp peaks occur in two periods between 7 am to 10 am and 5 pm to 8 pm, respectively, at morning and evening rush hours. On weekends, from time slot 121 to 168, the temporal variations show similar behaviour, but the total passenger flow on the weekends is lower than on weekdays.

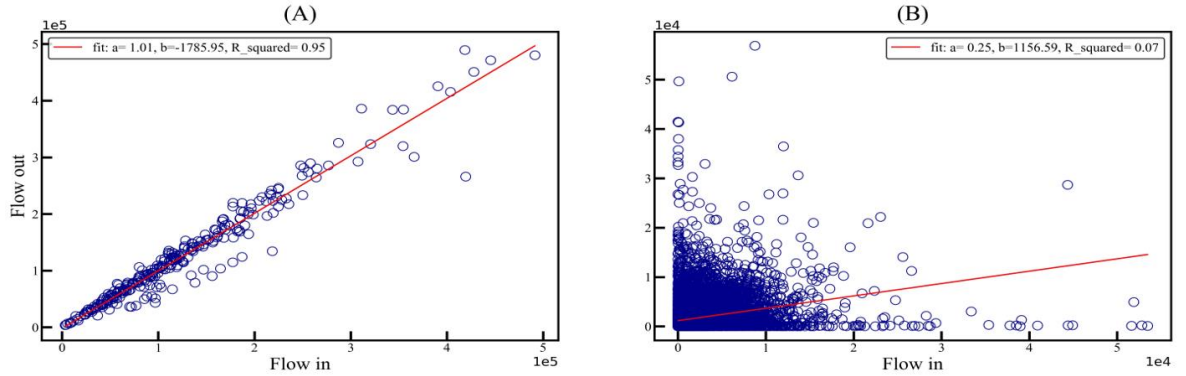

Figure 7: Ingoing versus outgoing flow at each station for (A) subway network and (B) bus network

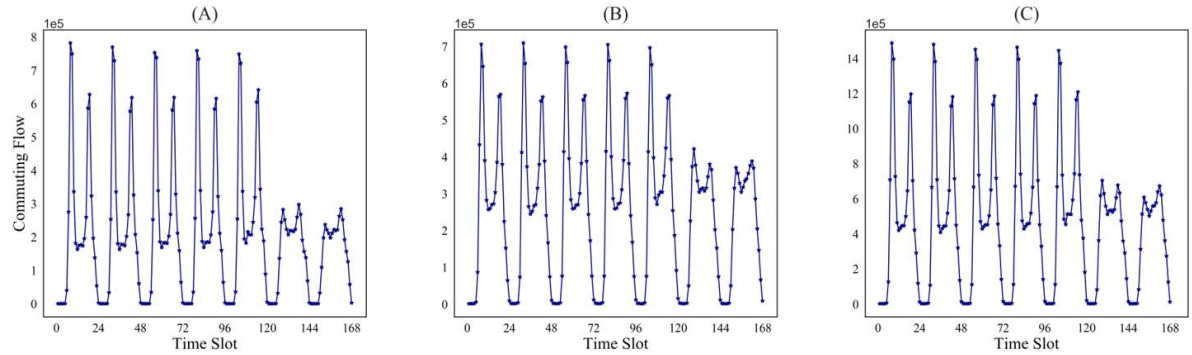

Figure 8: Total flow at station-based representation for (A) subway network, (B) bus network, and (C) PT network

The passenger flow at each node (station/line) for the 168-time slot are aggregated. The heat-map charts in Figure 9 represent the passenger flow at each node, where each pixel in the chart reflects the passenger flow at a specified node in a specific time slot. In each graph, the time slots, 168-time slot, are represented in the x −axis, and the y −axis refers to the ID of the nodes. It can be noticed that the passenger flows for each stop/line ID in all graphs reflect the regularity and periodicity in the macro-level. So, this is evidence that the flow can be predicted at the macro-level. It can be observed that the passenger flow for subway network at both representations is much higher than the bus network. The passenger flow for PT network is only a combination of the bus network and subway.

Moreover, the spatial distribution of the aggregated passenger flow of 7 days for the six representations is computed and visualised in Figure 10 It can be seen that the high passenger flow existed in the downtown of Beijing. The passenger flow related to the subway system is higher than the bus network, as shown in Figure 10. Line 10 and Line 4 in the subway system have a higher incoming flow, while the airport express subway line has a minimum passenger flow, as illustrated in Figure 10D.

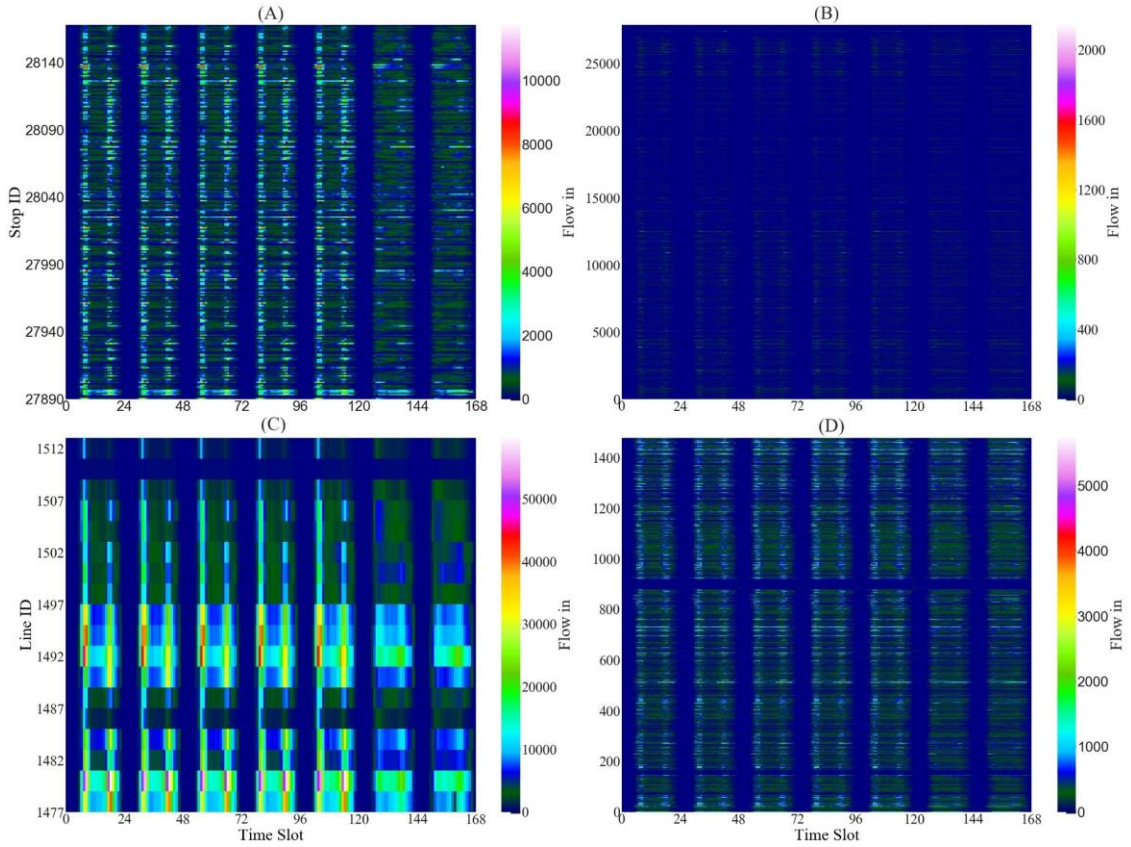

Figure 9: Passenger flow at different representation: (A) subway station-based representation, (B) bus station-based representation, (C) subway lines-based representation, and (D) bus lines-based representation

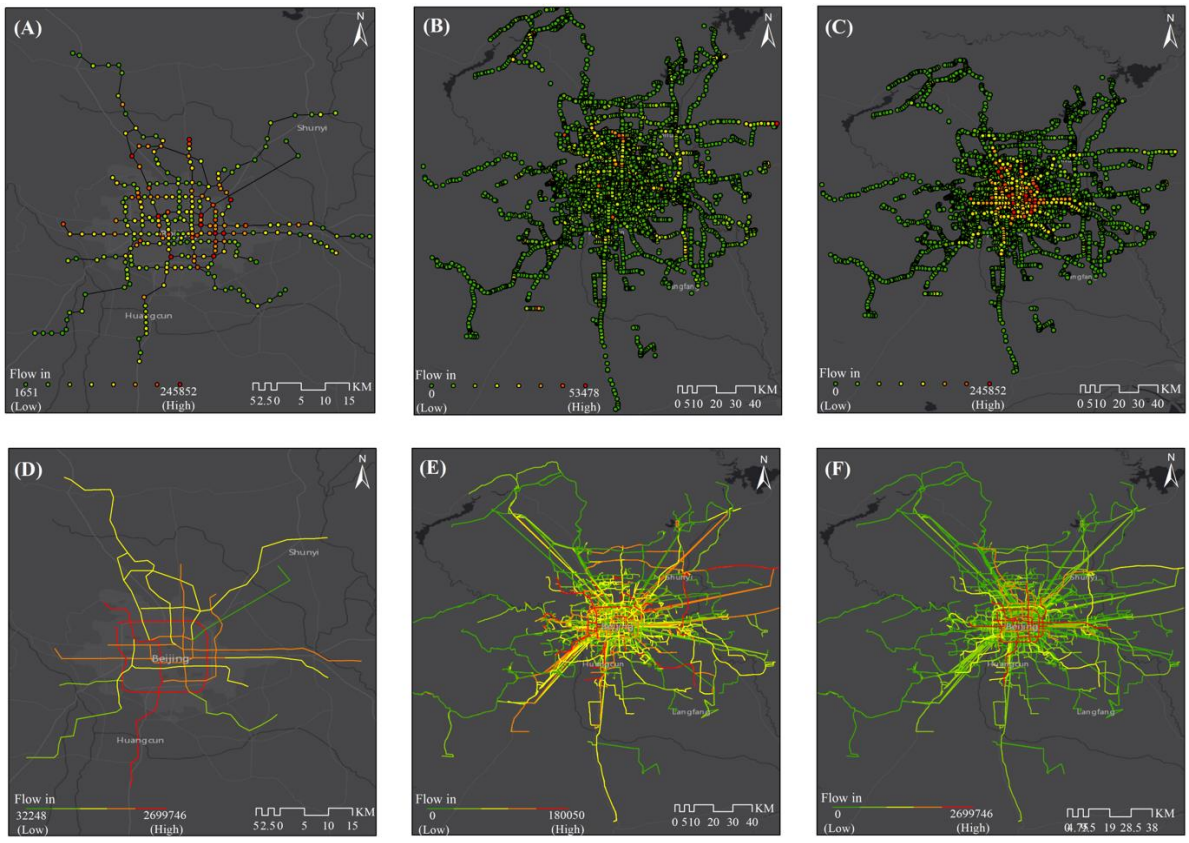

Figure 10: Spatial distribution of passenger flow over seven days for (A) subway station-based representation, (B) bus station-based representation, (C) PT station-based representation, (D) subway lines-based representation, (E) bus lines-based representation, and (F) PT lines-based representation

4.3 Correlation between passenger flow and network centrality

4.3.1 Correlation analysis for station-based representation

Figure 11 compares the spatial distribution of the degree centrality that measured using the conventional and modified techniques at station-based representation for the three networks. Where the green and red points reflect the nodes with lower and higher centrality, in sequence. The most exciting aspect of this graph is the subway stations have the highest value of the conventional degree centrality, considering that the subway station has multiple transfer edges than bus station beside the connecting edges for PT network graph, as shown in 11C. Furthermore, the subway stations have the highest value of the modified centrality measure, both because of the previous reason and because the subway station has a higher passenger flow. It is also intuitive in this Figure is the nodes with higher degree centrality are located in the central city because the density of the station is considerably high.

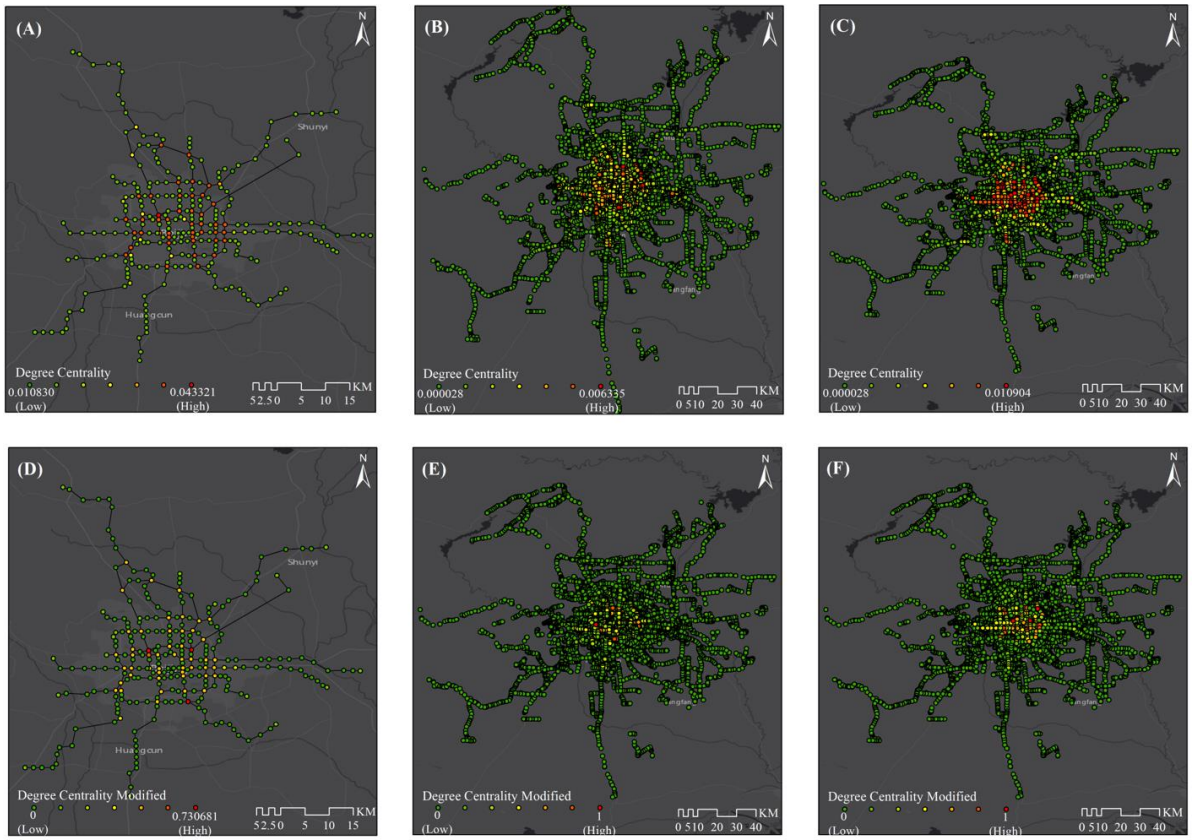

Figure 11: Spatial distribution of conventional degree centrality for the (A) subway network, (B) bus network, and (C) PT network. Spatial distribution of modified degree centrality for the (D) subway network, (E) bus network, and (F) PT network-all at station-based representation

Figure 12 displays the spatial distribution of the conventional and modified closeness centrality at station-based representation for the three networks, in which the lowest values for the closeness centrality are shown in the green, and the highest values are shown in the red. The nodes with higher closeness centrality are situated in downtown. Because of considering the transfer attribute of the networks, there is a medium and high value of the conventional closeness centrality have been converted into low values of the modified closeness centrality. The spatial distribution of the betweenness centrality at station-based representation follow similar spatial patterns. Because of considering the transfer attribute of the networks, there is a medium and high value of the conventional closeness centrality have

been converted into low values of the modified closeness centrality. Due to the limited space, they are not put in the paper

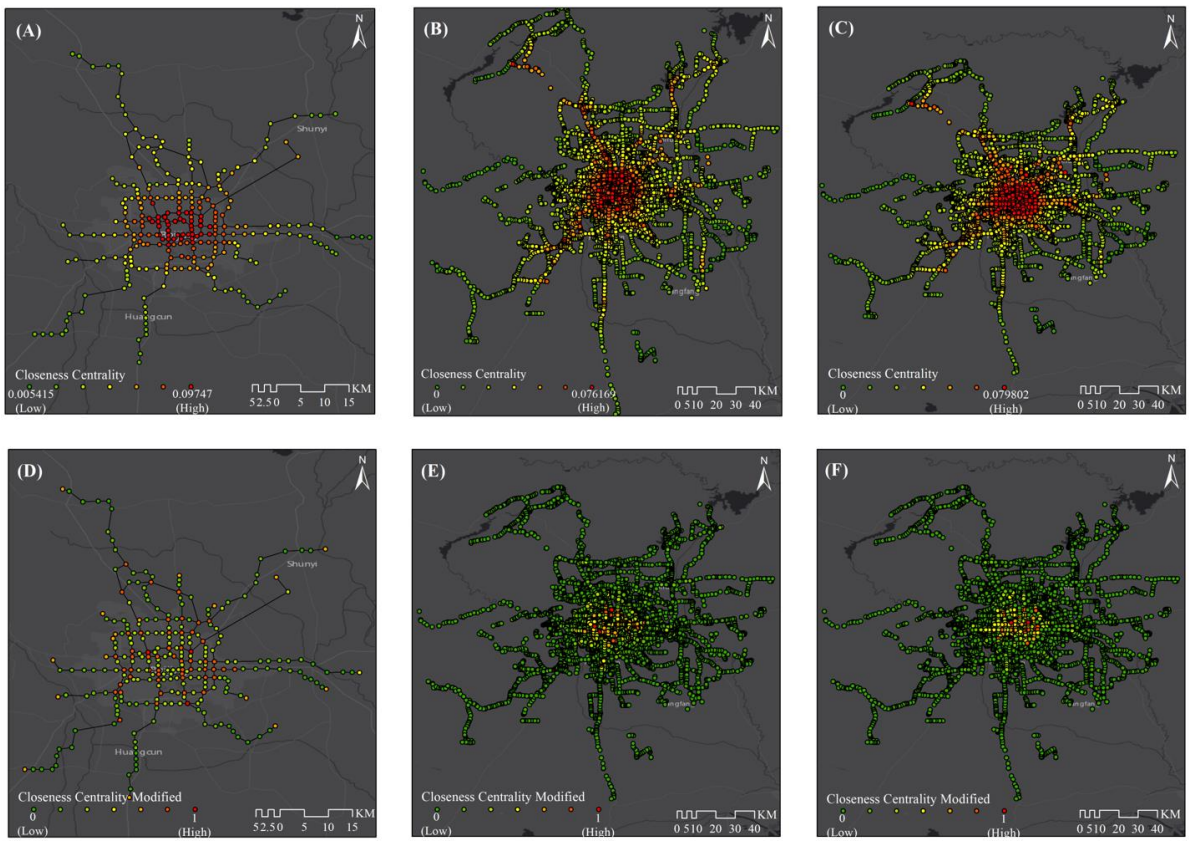

Figure 12: Spatial distribution of conventional closeness centrality for the (A) subway network, (B) bus network, and (C) PT network. Spatial distribution of modified closeness centrality for the (D) subway network, (E) bus network, and (F) PT network- all at station-based representation

The correlation coefficients between passenger flow and centrality measures are computed at station-based representation and taken in combination with station-based passenger Flow. Table 4 displays the mean correlation coefficients between passenger flow and the degree centrality. It is apparent from the Table 4 that the conventional degree centrality and passenger flow, at station-based representation, reflect a low correlation. Hence, it is evidence that there is a shortcoming of using the conventional degree for predicting the passenger flow. Although there is a slight improvement in the correlation between the modified degree centrality and passenger flow, it is not reliable in predicting the flow.

Furthermore, Table 4 shows the mean correlation coefficients between passenger flow and the closeness centrality. As shown in Table 4, the conventional closeness centrality and station-based passenger flow exhibit low correlation, especially for bus and PT networks. Where the mean value of the correlation coefficients between conventional closeness centrality and passenger flow in all time slots are 0.31, 0.11, and 0.08 for the subway network, bus network, PT network, respectively, this shows that the use of conventional closeness to predict and analyse the passenger flow is weak. There is an increase of 9% in the modified closeness centrality for subway network compared with the conventional one, an increase of 6% in the modified closeness centrality for bus network compared its conventional counterpart and, a considerable increase of 287% in modified closeness centrality For PT network, As presented in Table 4. Despite these increases, the correlation coefficients values

are small, indicating the modified closeness centrality inefficiency in predicting flow at station-based representation.

Moreover, Table 4 shows the correlation coefficients between passenger flow and the betweenness centrality at station-based representation for the three networks. The mean value of the correlation coefficients between conventional betweenness centrality and passenger flow in all time slots are 0.31, 0.04, and 0.38 for the subway network, bus network, PT network, respectively. Hence, the conventional betweenness at station-based representation is not appropriate for analysing and predicting the passenger flow, particularly in the bus network. The most likely causes of inefficiency of using the conventional centrality for bus network only is the high density of the bus stops. However, the observed difference between the correlation values in PT network and the corresponding in bus network is significant, this because of the combination of the subway network that attracts more passenger, which improved the performance of the bus network.

Table 4: The mean correlation coefficient between passenger's flow and selected centrality measures in station based-representation

Centrality Measures	Technique Type	Subway System	Bus System	PT System
Degree	Conventional	0.30	0.11	0.14
	Modified	0.33	0.14	0.35
Closeness	Conventional	0.31	0.11	0.08
	Modified	0.34	0.12	0.31
Betweenness	Conventional	0.311	0.04	0.38
	Modified	0.31	0.09	0.45

4.3.2 Correlation analysis for line-based representation

Moving on now to consider the centrality measures and the passenger flow at line-based representation. Firstly, the conventional centrality measures are computed based on the Equation 4, 6, and 8 at line-based representation and the modified centrality measures are calculated according to Equation 9 after obtaining the optimised parameter α for a degree, closeness, and betweenness for three networks.

The spatial distribution of the conventional degree centrality and the modified degree centrality are set out in Figure 13, where the lower and higher values of the degree centrality are indicated in green and red colours, respectively. What is striking about the results in this Figure is the subway lines have the highest value of the conventional and modified degree centrality. This is because most of the subway network is connected to a large number of different subway lines and bus lines as well, compared with bus network, as can be seen from Figure 13 C and F. It can be observed that Line 10 in subway network have the highest value of degree centrality, due to its geometric properties. Where it has 16 distinct transfer station, 57 kilometres of operating mileage (highest operating mileage) and this line is around the line. Because of considering the geometric properties of the networks, there is a medium and high value of the conventional centrality measures have been changed into low values of the modified centrality measures.

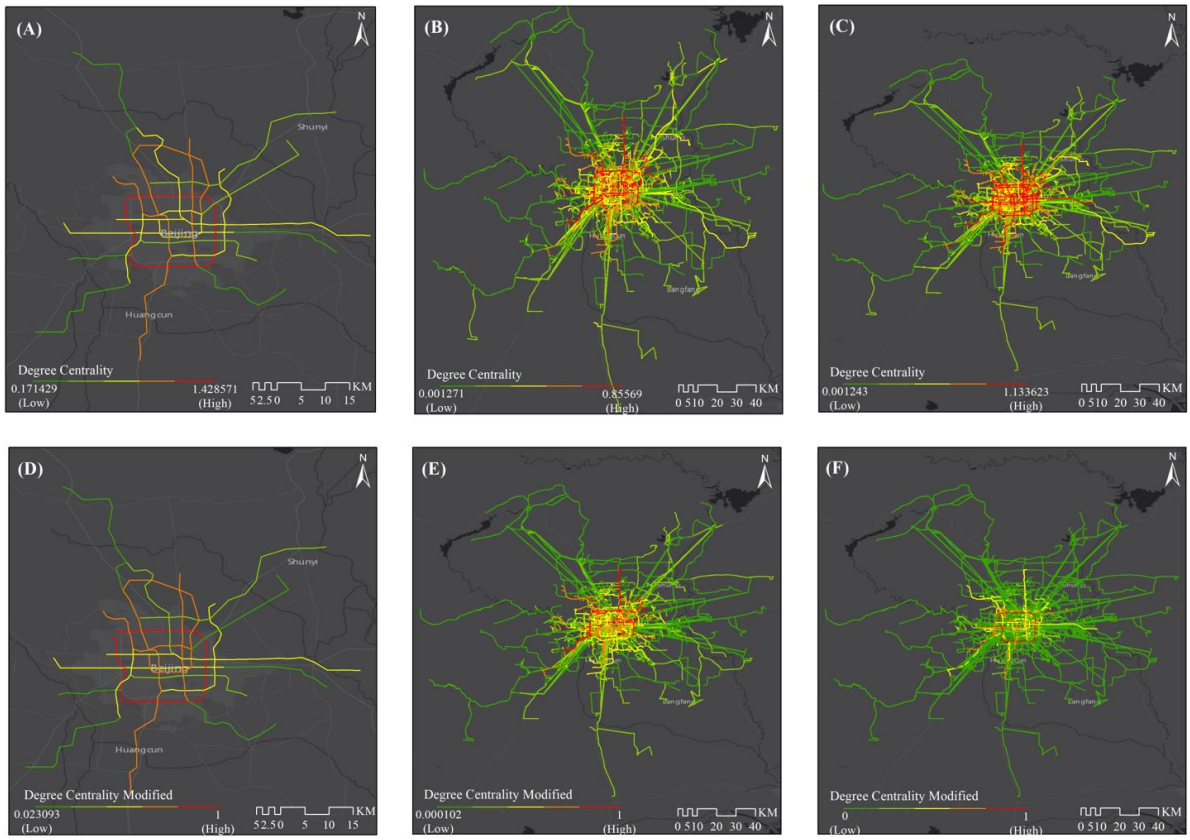

Figure 13: Spatial distribution of conventional degree centrality for the (A) subway network, (B) bus network, and (C) PT network. Spatial distribution of modified degree centrality for the (D) subway network, (E) bus network, and (F) PT network-all at line-based representation

Figure 14 provides the spatial distribution of the conventional closeness centrality and the modified closeness centrality. Red colour reflects the highest closeness centrality value, while the green colour stands out the lowest value of the closeness centrality. Similarly, Line 10 in the subway network have the highest value of closeness centrality. As for the spatial distribution of the conventional closeness centrality and the modified betweenness centrality for line-based representation is compared and follow the same spatial patterns of the other centrality measures. The subway lines have the highest value of the conventional and modified betweenness centrality, especially Line 10 in the subway network. Due to the limited space, they are not put in the paper.

The correlation coefficients between the passenger flows and the centrality measures are calculated at line-based representation. The correlation coefficients between the selected centrality measures and the line-based passenger flow for each period are provided in Table 5. As can be seen in Table 5, the conventional degree and the line-based passenger flow exhibit a high correlation, in particular for the subway network, compared with station-based representation. Contrary to the station-based representation, the conventional degree centrality at line-based representation is appropriate for predicting and analysing the passenger flow. The mean values of the correlation coefficients between the modified degree and passenger flow at line-based representation for the subway network, bus network, and PT network at all time slot are 0.81, 0.30, and 0.47, consecutively. Thus, it is indicated that the modified degree centrality for the subway network is more convenient for analysing and predicting passenger flow than that for bus network and PT network.

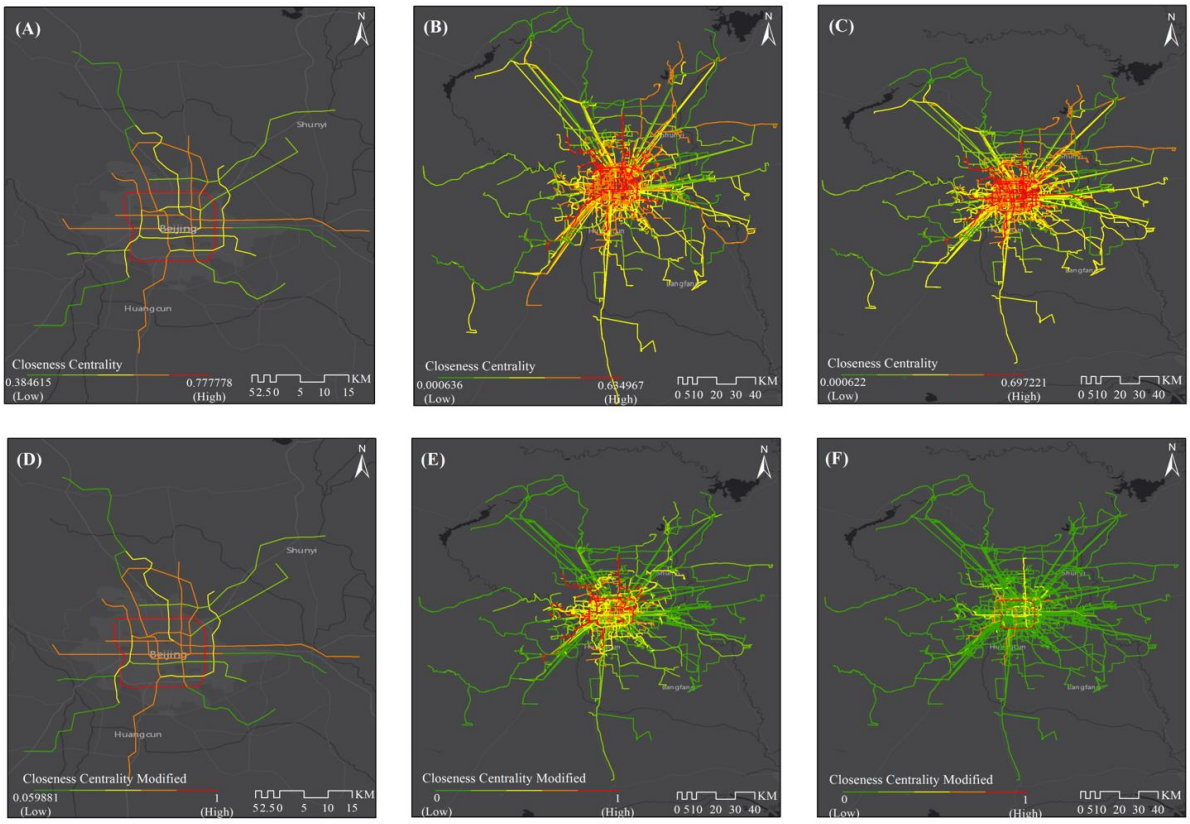

Figure 14: Spatial distribution of conventional closeness centrality for the (A) subway network, (B) bus network, and (C) PT network. Spatial distribution of modified closeness centrality for the (D) subway network, (E) bus network, and (F) PT network- all at line-based representation

Table 5: The mean correlation coefficient between passenger's flow and selected centrality measures in line based-representation

Centrality Measures	Technique Type	Subway System	Bus System	PT System
Degree	Conventional	0.77	0..29	0.31
	Modified	0.81	0.30	0.47
Closeness	Conventional	0.75	0.24	0.21
	Modified	0.82	0.28	0.45
Betweenness	Conventional	0.67	0.31	0.53
	Modified	0.68	0.31	0.54

Table 5 shows the correlation coefficients between the closeness centrality and the line-based passenger flows for each time slot at line-based representation for three networks. In Table 5, the conventional closeness and the line-based passenger flow demonstrates a high correlation, in particular for the subway network when setting against station-base representation. There is a significant difference between station-based representation and the line-based representation, in which the conventional closeness centrality at line-based representation is convenient to predict and analyse the passenger flow. The mean values of the correlation coefficients between the modified closeness and passenger flow at line-based representation for the subway network, bus network, and PT network at all time slot are 0.82, 0.28, and 0.45, consecutively. Hence, relying on the modified closeness centrality for the subway network to analyse and predict the passenger flow is better than depending on both the bus network and the PT network.

Table 5 demonstrates a high correlation between the conventional line-based passenger flow, particularly for the subway network and PT network in comparison to station-based representation. The conventional betweenness centrality at the line-based representation differs from the corresponding at the station-based representation in that the line-based centrality is suitable for predicting and analysing the passenger flow. It can be noticed that there is a slight improvement in the correlation between the modified betweenness centrality and passenger flow in comparison to the conventional betweenness centrality. Consequently, the line-based passenger flow can be predicted by using the betweenness centrality at line-based representation.

5 Conclusion

In the high-frequency city, estimating and predicting the passenger flow is one of the most critical issues due to its crucial role in the planning and management of the PTSs. However, limited research has been carried out on examining the correlation between the PTNs properties and the passenger flow. Through correlation coefficients value, we can identify the extent to which network properties can be used to predict the passenger flow. This study was undertaken to model PTNs from the PTSs (Bus, Subway, and Bus with subway systems) at two representation, namely station-based representation and line-based representation.

Moreover, SCD was used to compute the passenger flow based on corresponding each PTN and its representation. The network properties were analysed based on the conventional and modified technique where the conventional network centrality measures were computed based on the topological properties of the network only. By contrary, the modified network centrality measures were calculated based on both topological and geometrical properties of the network at different representations. The degree, closeness, and betweenness centralities were chosen for network analysis. The results of this study indicate that the conventional and modified network centralities for all PTNs at the station-based representation are not appropriate for predicting the passenger flow. In contrast, the conventional and modified network centralities at the line- based representation for the subway network and the modified network centrality for PT network are convenient to predict the passenger flow. So, it can be exploited to aid the policy makers and transportation agencies to take the appropriate decision for cancelling or adding lines with determining its route and the transfer stations. To describe the urban structure, we need to explore the interrelationship between the land use and the PTN centralities. In the best of our knowledge, however, there are limited research that concentrate on the impact of PTN centrality indicators on various types of urban land use. Therefore, in our next work we will explore the interrelationship between the PTN centrality indicators and Land use pattern.

Acknowledgement

The authors thank Dr LAI, Jianhui for processing the public transit big data used in this study, and the anonymous reviewers, who helped to improve this work. Author 2 acknowledges funding support from a start-up project (1-ZE6P) funded by The Hong Kong Polytechnic University.

References

Batty M. 2018. Visualizing aggregate movement in cities. Philos Trans R Soc B Biol Sci. 373(1753).

Beauchamp MA. 1965. An improved index of centrality. Behav Sci. 10(2):161–163.

Beijing Municipal Bureau of Statistics and NBS Survey Office in Beijing. 2017. Population and employment annual report in 2017 [Internet]. [accessed 2019 May 28]. http://tjj.beijing.gov.cn/nj/main/2018-tjnj/zk/indexeh.htm

Beijing Transportation Research Centre. 2017. Beijing Traffic Development Annual Report in 2017 (in Chinese) [Internet]. [accessed 2019 Jul 24]. http://www.bjtrc.org.cn/List/index/cid/7.html

Dimitrov SD, Ceder A. 2016. A method of examining the structure and topological properties of public-transport networks. Phys A Stat Mech its Appl. 451(2016):373–387.

Feng S, Hu B, Nie C, Shen X. 2016. Empirical study on a directed and weighted bus transport network in China. Phys A Stat Mech its Appl. 441:85–92.

Von Ferber C, Holovatch T, Holovatch Y, Palchykov V. 2009. Public transport networks: Empirical analysis and modeling. Eur Phys J B. 68(2):261–275.

Von Ferber C, Holovatch Y, Palchykov V. 2005. Scaling in public transport networks. Condens Matter Phys. 8(141):225–234.

Freeman LC. 1978. Centrality in social networks conceptual clarification. Soc Networks. 1(3):215–239.

Gao S, Wang Y, Gao Y, Liu Y. 2013. Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality. Environ Plan B Plan Des. 40(1):135–153.

Hage P, Harary F. 1996. Centrality. In: Isl Networks Commun Kinship, Classif Struct Ocean. Cambridge: Cambridge University Press; p. 165–203.

Huang L, Zhu X, Ye X, Guo W, Wang J. 2016. Characterizing street hierarchies through network analysis and large-scale taxi traffic flow: a case study of Wuhan, China. Environ Plan B Plan Des. 43(2):276–296.

Jiang J, Metz F, Beck C, Lefevre S, Chen J, Wang QA, Pezeril M. 2011. Double power-law degree distribution and informational entropy in urban road networks. Int J Mod Phys C. 22(1):13–20.

Kazerani A, Winter S. 2009a. Modified Betweenness Centrality for Predicting Traffic Flow. In: 10th Int Conf GeoComputation, Sydney, Aust. Sydney, Australia.

Kazerani A, Winter S. 2009b. Can betweenness centrality explain traffic flow. In: 12th Agil Int Conf Geogr Inf Sci. Hanover, Germany; p. 1–9.

Kerner BS, Demir C, Herrtwich RG, Klenov SL, Rehborn H, Aleksić M, Haug A. 2005. Traffic state detection with floating car data in road networks. In: IEEE Conf Intell Transp Syst Proceedings, ITSC. Vol. 2005. Vienna, Austria; p. 700–705.

Kim J, Hastak M. 2018. Social network analysis: Characteristics of online social networks after a disaster. Int J Inf Manage. 38(1):86–96.

Li X, Guo J, Gao C, Su Z, Bao D, Zhang Z. 2018. Network-based transportation system analysis: A case study in a mountain city. Chaos, Solitons and Fractals. 107:256–265.

Liu X, Chow JYJ, Li S. 2018. Online monitoring of local taxi travel momentum and congestion effects using projections of taxi GPS-based vector fields. J Geogr Syst. 20(3):253–274.

Luo D, Cats O, van Lint H. 2019. Can passenger flow distribution be estimated solely based on network properties in public transport systems? Transportation (Amst).:1–20.

Ma X, Liu C, Wen H, Wang Y, Wu YJ. 2017. Understanding commuting patterns using transit smart card data. J Transp Geogr. 58:135–145.

Ma X, Wu YJ, Wang Y, Chen F, Liu J. 2013. Mining smart card data for transit riders' travel patterns. Transp Res Part C Emerg Technol. 36:1–12.
Mukherjee S. 2012. Statistical analysis of the road network of India. Pramana - J Phys. 79(3):483–491.
Pun L, Zhao P, Liu X. 2019. A Multiple Regression Approach for Traffic Flow Estimation. IEEE Access. 7:35998–36009.
Scott J. 1988. Social Network Analysis. Sociology. 22(1):109–127.
Shanmukhappa T, Ho IWH, Tse CK. 2018. Spatial analysis of bus transport networks using network theory. Phys A Stat Mech its Appl. 502:295–314.
Si B, Fu L, Liu J, Shiravi S, Gao Z. 2016. A multi-class transit assignment model for estimating transit passenger flows-a case study of Beijing subway network. J Adv Transp. 50(1):50–68.
Sienkiewicz J, Hołyst JA. 2005. Statistical analysis of 22 public transport networks in Poland. Phys Rev E - Stat Nonlinear, Soft Matter Phys. 72(4):1–11.
Tang J, Liu F, Wang Y, Wang H. 2015. Uncovering urban human mobility from large scale taxi GPS data. Phys A Stat Mech its Appl. 438:140–153.
Tang J, Wang Y, Liu F. 2013. Characterizing traffic time series based on complex network theory. Phys A Stat Mech its Appl. 392(18):4192–4201.
Tian Z, Jia L, Dong H, Su F, Zhang Z. 2016. Analysis of Urban Road Traffic Network Based on Complex Network. Procedia Eng. 137:537–546.
Wang S, Zheng L, Yu D. 2017. The improved degree of urban road traffic network: A case study of Xiamen, China. Phys A Stat Mech its Appl. 469:256–264.
Xu Q, Mao BH, Bai Y. 2016. Network structure of subway passenger flows. J Stat Mech Theory Exp. 2016(3).
Yan C, Wei X, Liu X, Liu Z, Guo J, Li Z, Lu Y, He X. 2018. A new method for real-time evaluation of urban traffic congestion: a case study in Xi'an, China. Geocarto Int.:1–16.
Ye P, Wu B, Fan W. 2016. Modified Betweenness-Based Measure for Prediction of Traffic Flow on Urban Roads. Transp Res Rec J Transp Res Board. 2563(1):144–150.
Zhang H, Shi B, Yu X, Mou Z, Li M, Wang L, Zhuge C. 2018. Transfer stability of urban subway network with passenger flow: Evidence in Beijing. Int J Mod Phys B. 32(14).
Zhang T, Dong S, Zeng Z, Li J. 2018. Quantifying multi-modal public transit accessibility for large metropolitan areas: a time-dependent reliability modeling approach. Int J Geogr Inf Sci. 32(8):1649–1676.
Zhang X, Chen G, Han Y, Gao M. 2016. Modeling and analysis of bus weighted complex network in Qingdao city based on dynamic travel time. Multimed Tools Appl. 75(24):17553–17572.
Zhao S, Zhao P, Cui Y. 2017. A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China. Phys A Stat Mech its Appl. 478:143–157.
Zou Q, Yao X, Zhao P, Wei H, Ren H. 2018. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. Transportation (Amst). 45(3):919–944.