Highlights

City2vec:Urban Knowledge Discovery Based on Population Mobile Network

Yan Zhang,Xiang Zheng,Marco Helbich,Nengcheng Chen,Zeqiang Chen

- Capturing long-term population movements based on electronic map phone number data
- Raising the level of awareness of the city
- Compute embedding representations of nodes in city networks
- City vectors contain geographic location and culture attribute

City2vec:Urban Knowledge Discovery Based on Population Mobile Network

Yan Zhang[a], Xiang Zheng[c,e], Marco Helbich[d], Nengcheng Chen[a,b,*] and Zeqiang Chen[b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China
[c] School of Information Management, Wuhan University, Wuhan 430072, China
[b] National engineering research center of geographic information system, China University of Geosciences Wuhan 430074, China
[d] Department of Human Geography and Spatial Planning, Utrecht University, Heidelberglaan 2, 3584 CS, Utrecht, The Netherlands
[e] Collaborative Sensing and Knowledge Services Research Group, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Keywords:

City2vec

Population Migration

City Sensing

Complex Network

Population Flow

ABSTRACT

Due to the needs of social and economic development, population movements between cities often occur on a large scale. Spontaneous population movements between cities constitute a Mobile Network of enormous scale, and although a considerable amount of research has been conducted to parse this structure using algorithms related to complex networks, it is not possible to quantitatively describe the differences and similarities between different cities. We use graph embedding algorithm that extends the traditional complex network approach of one-dimensional cognition of cities to two-dimensional cognition. It could project city information into a high-dimensional mathematical space by learning the population flow relationships and computing a unique vector representation. We parsed the cell phone number data in about 80 million POIs in 334 cities across China provided by Tencent eMap, and constructed a city Mobile Network containing 2,662,596 directed edges that could effectively capture long-term, long-distance population migrations. The city embedding is actually a reduced dimensional representation of the mobile network, which can retain richer original information. Our research method can not only accomplish the tasks that traditional complex network analysis methods can do well, but also achieve higher dimensional cognition of cities (e.g., spatial relationships between cities and maturity of urban agglomerations), which is an effective complement to traditional graph theory methods.

1. Introduction

With China's rapid economic development and the further improvement of its transport infrastructure, the scale of migration is becoming increasingly large. According to the China 2021 census, 35% of China's population has a household registration that does not coincide with their actual location, and approximately 492.76 million people move freely between cities, including 124.84 million people who move across provinces [1]. Population movements of this magnitude are the result of people's spontaneous choice and voting with their feet (Singleton, 2015). Economist Charles Tiebout suggests that under the assumptions of no administrative restrictions on population movement, a large number of local governments with same tax systems and open and transparent information. Since the public services, employment opportunities and tax policy varies from city to city, residents will choose to settle in the city that offers the greatest benefits. The population is free to move out of unsatisfactory areas and into areas where they could meet their individual preferences.

During the first half of China's urbanisation, as cities in coastal areas took the lead in implementing the reform and opening-up policy, implementing market economy system, offering tax breaks and opening up markets, they attracted

*Corresponding author
✉ cnc@whu.edu.cn (N. Chen)
www.swe.whu.edu.cn (N. Chen)
ORCID(s): 0000-0002-3521-9972 (N. Chen)

[1] www.stats.gov.cn/tjsj/zxfb/202105/t20210510_1817176.html

large amounts of overseas capital and technology. The rapid economic development of the eastern coastal areas saw a sharp rise in demand for labour. Due to the attraction of higher income levels and better employment conditions, a large number of people migrated from the developing inland areas to the more developed coastal areas (Lu et al., 2013; You et al., 2018). In the second half of China's urbanisation, with the gradual deepening of economic system reforms, the movement of capital, industry and population accelerated, and the resources of inland urban agglomerations were enriched towards the inner central cities. As large cities could provide better healthcare services, have higher levels of education and more diverse job options; the migration of rural populations and populations from small and medium-sized cities to inland regional centres has also occurred. The above two approaches have been the main drivers of population movement in China over the last 30 years, implying a complex process of group decision-making and natural selection (Zhu and Lin, 2016).

Population migration is the most important manifestation of human movement across geographies (e.g., between two cities). It usually involves a short-term or long-term change of residence, is driven by factors related to, for example, the physical environment, socio-economics, and politics (Yang et al., 2019). Increasing amounts of highly granular data in space-time provide paramount sources to understand inter-city human movement. Extracting hidden patterns describing human mobility is among the long-standing challenges of urban research (Li et al., 2021a; Bi and Ye, 2021).

However, identifying long-term movements of people is more difficult in China than identifying short-term movements. People may travel to another city for a short period of time on business or to a transport hub city to change transport. This will generate much order data that such short-term movements could be easily captured, while a considerable number of commercial companies provide the appropriate data interfaces. The easiest way to identify long-term movements is to obtain data from driving licences or household registration authorities. Due to the household management system in China, economically developed cities set entry thresholds (Xu and Wu, 2022). This has resulted in a significant proportion of people working locally but not being recorded. In order to map the actual distribution of China's population, the government has conducted seven large-scale censuses in the last 70 years. Such censuses are large, costly and time-consuming, with the most recent one forming 679,000 census agencies and hiring over 7 million census workers[2].

Electronic maps are a twin mapping of the real world, and the wealth of data opens up new opportunities for researching the long-term mobility of the population. For example, the owner of a restaurant-type POI will provide a phone number for customers to contact. The fourth to seventh digits of the phone number represent the city in which the user's mobile phone number is registered. If the city where the POI's mobile phone is registered is not the same as the city where the POI is located, then this implicitly creates an origin-destination (OD) relationship from city A to city B. While people change where they live, mobile phone numbers do not generally change. Among the 80 million Tencent eMap POI data across the country, 2,662,596 similar interactive OD messages exist. Using this information, we could construct a national-scale "mobile network" to reveal the long-term mobility of the population.

In order to deal with such mobile networks of long-term population movements, the most common approach is to use social network analysis. Cities are considered as network nodes and inter-city population movements of varying intensity are considered as network edges. This approach allows the calculation of different indicators of urban nodes, such as Degree centrality, Closeness centrality, Harmonic centrality, Betweenness centrality, etc. These metrics provide a clear picture of how critical a city is in the network from different perspectives. In addition, as the migration of the population (the edge of the network) is closely related to the socio-economic elements of the city (Zheng et al., 2020; Hong et al., 2019), this also brings the possibility of uncovering the community structure of the city nodes community structure (city cluster) and hierarchy structure (city scale). However, traditional methods of network analysis are inadequate in the context of the population mobility network in this paper. We know that there is a spatial distance relationship between cities, and given that two cities are similar in attractiveness, most people will prefer the city that is closer to their hometown or has a more similar climate, but this spatial relationship is not reflected in the complex network analysis results described above. Furthermore, as urban agglomerations become more populated and economically developed, the boundaries between cities within an urban agglomeration become increasingly blurred. Traditional network analysis methods can measure the strength or weakness of cities, but cannot assess the maturity of urban cluster development or explain the tendencies in city pair selection. Here, we set the following two characteristics of urban agglomerations with highly mature development. Firstly, the movement of people within urban agglomerations is much more frequent than outside; In addition, cities with similar levels of development within urban

[2] stats.gov.cn/tjsj/zxfb/202105/t20210510_1817176.html

agglomerations have similar attractiveness to other cities people, and there is not much of a geographical proximity effect. Satisfying these two characteristics means that the cities of the cluster are already highly integrated.

The main contributions of this paper are twofold.

Firstly, we address the problem of identifying long-term population mobility by using a new data source to construct a population mobility network. Secondly, we propose a new method to compute the embedding representation of urban nodes, extending the level of human knowledge of population mobility.

Our study is organized as follows: the second section is the related research of this paper, which introduces the current research status of spatial interaction, the third section details our algorithm, the fourth section shows our experimental and validation process, and the last section concludes this study.

2. Related works

The simulation of intercity interactions has been an important area of urban analytics (Li et al., 2021a). Exploring the relationship between urban industrial functions and spatial interaction patterns of transportation has been of interest (Hu et al., 2021), whereas the quantification of the flow of people, logistics, capital, information and technology stands out. With the introduction of the concept of social sensing and the increasing availability of smartphones, a spatial data stream represented by online social relationship data has emerged (Liu et al., 2015; Li et al., 2021b; Zhang et al., 2021a).

This spatial data flow is actually recording and understanding the interactions of geographic entities, which are influenced by constant human movement. The human movement further quantifies the strength of connections between geographic entities, and in turn generates a spatial association graph of these interactions. This spatial association graph has been used to analyze the spatial structure and changing trends of cities (based on traffic flows generated from network trajectory data) (Zhang et al., 2021b),to analyze the similarity of locations (based on web review data)McKenzie and Adams (2018),to assess the polycentric structure of urban agglomerations (based on cell phone signaling data)(Hui et al., 2018),to measure the semantic relevance (based on web news records)(Hu et al., 2017), to analyze global trade patterns (based on AIS records of ships)(Kaluza et al., 2010), to assess the geographic choice of skilled workers for employment (based on resume data from online job sites)(Zhang et al., 2020a) and to inferr geographic attributes of cities (word co-occurrence relationships based on social media data) (Chen et al., 2021a).

The growing popularity of smart devices provides a new data stream where spatial social interaction measures are easily assessable and harmonized. This provides a huge amount of data from multiple sources for spatio-temporal analysis in geographic information science (Rocklage et al., 2021; Zhang et al., 2021c; Xu et al., 2021). Facing this huge network structure formed by spatial mapping of geographical entities, the most commonly used treatment scheme is the complex network analysis method. For such spontaneously generated networks with small-world and scale-free features, there are four main types of fundamental tasks, including Node classification, link prediction, Community Detection and Graph Classification (Zhou et al., 2020). Many scholars have applied and innovated around these fundamental tasks, and these studies mainly focus on complex network community discovery based on modularity and spectral clustering (Newman, 2003; Lei et al., 2019), link prediction and centrality analysis (Trouillon et al., 2016), influence calculation and heterogeneous network alignment (Chauhan et al., 2021; Hari et al., 2021). It also includes some work at the algorithmic level of graph space optimization, using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to extract and merge traffic flow networks to form better complex network graph layouts (Von Landesberger et al., 2015; Wang et al., 2022a) and combining K-shortest paths & hierarchical clustering to parse urban transportation network structure (Yue et al., 2019; Wang et al., 2022b).

There are also interesting studies on cities that use the spatial differences between where social media users are registered and where they live, i.e., the flow of information in social media networks, to describe the relationships of urban interactions (Cvetojevic and Hochmair, 2021).The urban relationships contained in social media are used to measure the nature of spatial connectivity at different scales (Li et al., 2021c) and thus to analyze spatial inequalities in regional development (Ouyang et al., 2020).

Spatial data on various types of information, material, and energy flows can also be used to resolve urban structures, analyze spatially and functionally connected urban agglomerations, and discover industrial synergies among them (Hui et al., 2018; Shao et al., 2020; Wang et al., 2022c). The network analysis method and bus networks data can also be used to study the geospatial heterogeneity, directional imbalance and spatial agglomeration features (Wang et al., 2020; Zeng et al., 2019). Mobile signalling is also a good source of spatio-temporal interaction data. Based on these data, a study constructs a large complex network containing 60 cities. They use community discovery algorithms to find the sub-networks in the network and reveal the hierarchical structure of the cities (Zhang et al., 2020b).

The above studies share the common drawback that they can only describe one property of nodes, such as: the degree of nodes, the centrality of nodes, etc. (Yao et al., 2021; Charyyev and Gunes, 2019). In the mathematical description of a city, we often get a kind of 'one-hot' encoding vector based on different evaluation indicators, which is of variable length and difficult to quantify under different indicator systems. The Word2vec method, as the originator of various later 2vec methods, solves the problem of mapping sparse multidimensional vectors to dense vectors using neural networks, and is an excellent feature extraction method (Mikolov et al., 2013). On this basis, according to different application scenarios, doc2vec (Le and Mikolov, 2014), which is applied to document vectorization, tweet2vec (Dhingra et al., 2016), which is applied to social media feature extraction, and item2vec (Dhingra et al., 2016), which is applied to recommendation systems, have emerged one after another. Node2vec and Graph2vec methods first introduced vectorization methods into the graph structure represented by knowledge graphs, and they use clustering methods to pre-label nodes as the recommendation system "cold start "knowledge, which efficiently and rapidly establishes the connection of nodes (Grover and Leskovec, 2016; Narayanan et al., 2017; Palumbo et al., 2018).

These methods mentioned above have also been used in GIS in some applications. Considering point-of-interest data with geotags as documents and using Word2vec method to vectorize land patches, different urban functional areas can be identified (Yao et al., 2017); The Place2vec method treats the adjacencies of geographic entities as spatial contexts and obtains better spatial embedding results (Yan et al., 2017); using Word2vec to extract the semantic information to match unstructured addresses (Lin et al., 2020). These studies solve the problem of semantic alignment of data with stronger geographical representation capability compared to traditional methods (Chen et al., 2021b).

Graph structures are widely available in natural sessions, and graph algorithms represented by graph neural networks have been used extensively in many fields and have become a popular research topic in geographic studies (Maduako and Wachowicz, 2019; Zhu et al., 2018).We propose a City2vec method based on Mobile Network and Node2vec algorithm, which uses a depth-first strategy (DFS) and breadth-first strategy (BFS) to traverse the city association graph. This algorithm could obtain the relationship sequence of cities and thus calculate the embedding expression. It not only preserves the complex network properties in the Mobile Network, but also obtains the city embedding properties, which brings the possibility of downstream tasks.

3. Methods

The research methodology is divided into three parts as shown in Figure 1: the first is to build a population Mobile Network with the interaction information implied by the social sensing data spontaneously generated by humans (POI data). The second uses neural networks to sample the graph knowledge, gets an embedded expression of the city. The last part is based on the calculation results of the second part for downstream tasks.

Feature engineering is an important aspect in machine learning and deep learning. In network-structured data, we convert the feature extraction task into an unsupervised feature learning task. The underlying structure of the network can be highly nonlinear, and learning network features requires maintaining the original network structure, so there are difficulties in designing a suitable model to capture the structural information of it.

Algorithms such as Skip-Gram and CBOW have emerged in natural language processing to learn continuous feature representations of words. It scans words in a document and computes the embedding features of the words to predict nearby words (Crivellari and Beinat, 2019). Such an approach is based on the distributional assumption, namely that words tend to have similar meanings in similar contexts, and optimize a custom graph-based objective function using stochastic gradient descent (SGD) motivated. The Word2vec model based on the CBOW algorithm considers the contextual relationships of words (x_i), sets a sampling window of size c, and performs sliding sampling of the corpus (Rong, 2014) as shown in Figure 2B. The model is trained based on the sampled data and predicts the probability of occurrence of x_t based on the contextual content. The mathematical formula is as follows:

$$P(x_t|x_{t-c}^{t+c}) = P\left(x_t|\tau(x_{t-c}, x_{t-c+1}, \cdot\cdot\cdot, x_{t+c-1}, x_{t+c})\right) \tag{1}$$

τ refers to the word vector addition operation in Word2vec that adds the vectors of all adjacent words in the window. The input corpus was transformed into the vector of a specified dimension using the word2vec tool, in which words with similar meaning have closer spatial distance.

The Word2vec model in natural language processing transforms sparse knowledge into dense real domain vector

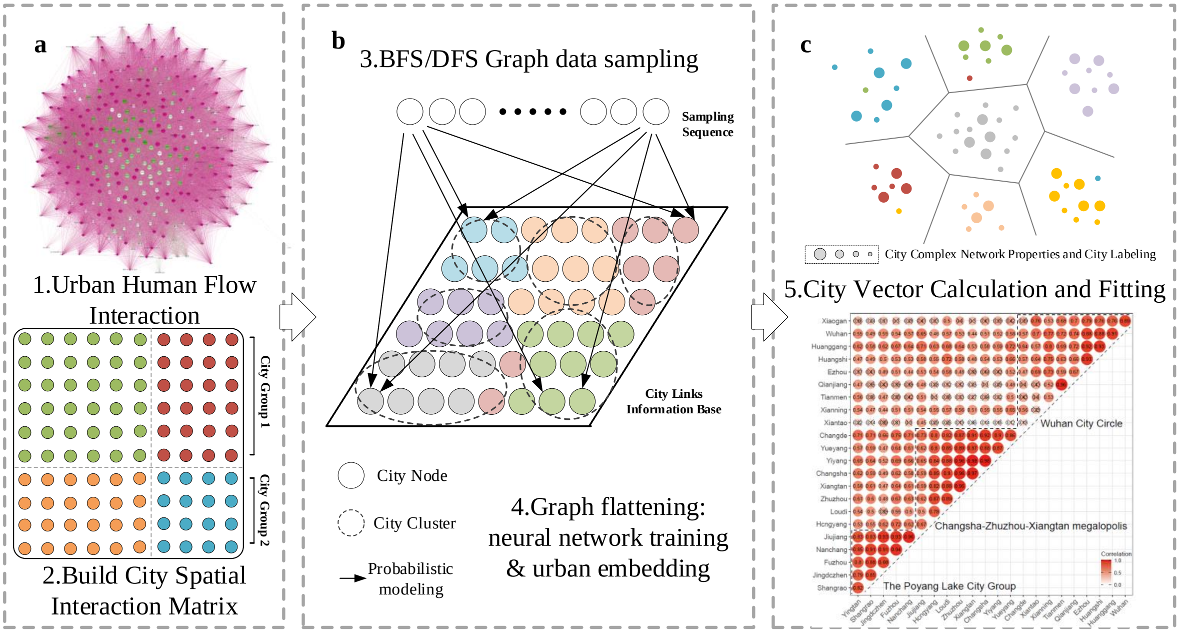
a
1.Urban Human Flow Interaction
City Group 1
City Group 2
2.Build City Spatial Interaction Matrix
b
3.BFS/DFS Graph data sampling
Sampling Sequence
City Links Information Base
City Node
City Cluster
Probabilistic modeling
4.Graph flattening: neural network training & urban embedding
c
City Complex Network Properties and City Labeling
5.City Vector Calculation and Fitting
Xiaogan
Wuhan
Huanggang
Huangshi
Ezhou
Qianjiang
Tianmen
Xianning
Xiantao
Changde
Yueyang
Yiyang
Changsha
Xiangtan
Zhuzhou
Loudi
Hengyang
Jiujiang
Nanchang
Fuzhou
Jingdezhen
Shangrao
Wuhan City Circle
Changsha-Zhuzhou-Xiangtan megalopolis
The Poyang Lake City Group
Correlation

Figure 1: The technical workflow: a) construction of the national population interaction matrix based on the graph information implied by POI cell phone numbers; b) sampling the Mobile Network using BFS/DFS algorithm and calculating the embedding vector of cities; and c) using the coalescent subgroup algorithm to identify urban agglomerations.

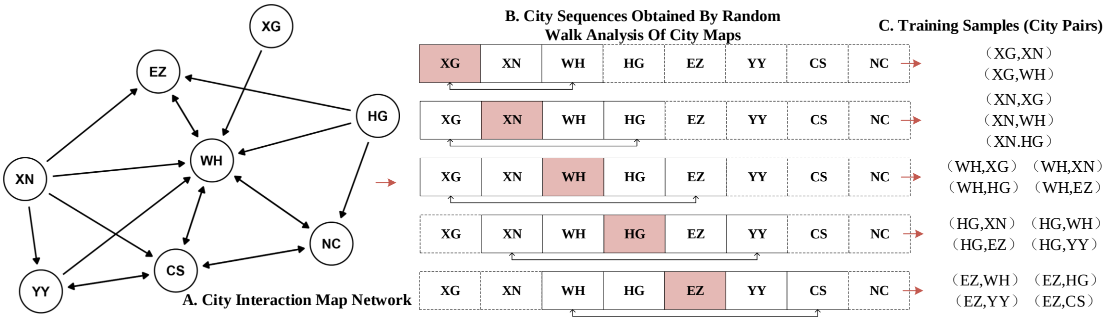

Figure 2: Schematic diagram of the acquisition of training samples (sampling window c of size 2, wandering sequence length l of 8, node name abbreviation of city name).

knowledge. This knowledge embedding algorithm is semantically driven, i.e., knowledge embedding is computed based on word co-occurrence probabilities (Olson et al., 2021). Instead, we study spontaneous population mobility, which is socio-economically driven, i.e., learning characteristics from population Mobile Network.

We can think of the network as a "document" and we transform the feature learning of the network into a maximum likelihood estimation problem. An ordered sequence of nodes is generated by sampling the nodes of the network (Palumbo et al., 2018). We describe the migratory network by $G = (V, E)$, where V represents city nodes and E represents the migratory flow between cities. Let $f : V \rightarrow \mathbb{R}^d$ the mapping function from nodes to feature representations we aim to learn for a downstream prediction task. As shown in Figure 3, the number N of neural units in the hidden layer is the dimension of our output city vector d. Here d is also the dimensionality of our feature representation. Equivalently, f is a weighted matrix of size $|V| \times d$ parameters. For every source node $u \in V$, we define $N_S(u) \subset V$ as a network neighborhood of node u generated through a neighborhood sampling strategy S. We attempt to optimise the following objective function Equation 2 as the loss function of the model,which maximizes the log-probability of observing a network neighborhood $N_S(u)$ for a node u conditioned on its feature representation, given by f:

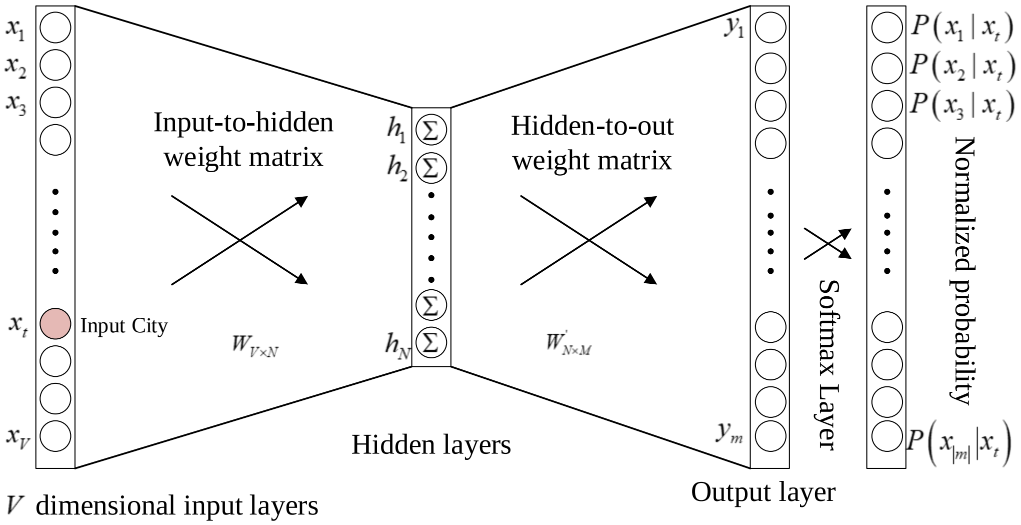

Figure 3: The training process of embedding vectors. We use the sequences generated by graph sampling as corpus,a simple CBOW model with only one word in the content is shown.

$$\max_f \sum_{u\in V} logPr(N_S(u)|f(u)) \tag{2}$$

The model continuously adjusts the weight matrices $W'_{N\times M}$ and $W_{V\times N}$ during the training process (Figure 3). It could predict the conditional probability P of the output word x_t given the contextual information and its weight matrix W. In the iteration of the model, Hoffman tree coding with negative sampling algorithm is often applied to speed up the training.

Based on the above graph sampling algorithm and population Mobile Network, we propose a City2vec algorithm for learning the features of city nodes and mapping cities to vector space to maximize the retention of graph information. We use the algorithm to perform biased Deepwalk sampling of the Mobile Network structure (Grover and Leskovec, 2016). We "flatten" the graph structure into r bars and a wandering sequence of length n. Where r represents the number of random walks generated by each node in the graph, and n represents the number of nodes per random walk (i.e., the length of the retrieved sequence).

The random walk generates a sequence of each city node uniformly. To avoid this, we introduced the biased wandering method from the node2vec algorithm (Grover and Leskovec, 2016). As shown in Figure 4, the sequence embedding expression of city nodes is generated by introducing two parameters p and q and introducing BFS with DFS in the random walking process.The BFS approach focuses on neighbouring nodes, such as the paths marked in red in Figure 4B. It covers a shorter distance but a larger area and can be a good description of the local features of a graph. The DFS approach focuses more on further nodes, such as the paths marked in blue in Figure 4B. It covers a smaller area, but over longer distances and gives a good description of the global features of the graph. The parameter p represents the probability of sampling a neighbouring node of the previous node, and q represents the probability of sampling a non-neighbouring node of the previous node. The probability that the current urban node u samples a new node t is as Equation 3, where S_{prev} is the previous node.

$$\alpha = \begin{cases} \frac{1}{p} & if\ d_{S_{prev},t} = 1 \\ 1 & if\ d_{S_{prev},t} = 0 \\ \frac{1}{q} & if\ d_{S_{prev},t} = 2 \end{cases} \tag{3}$$

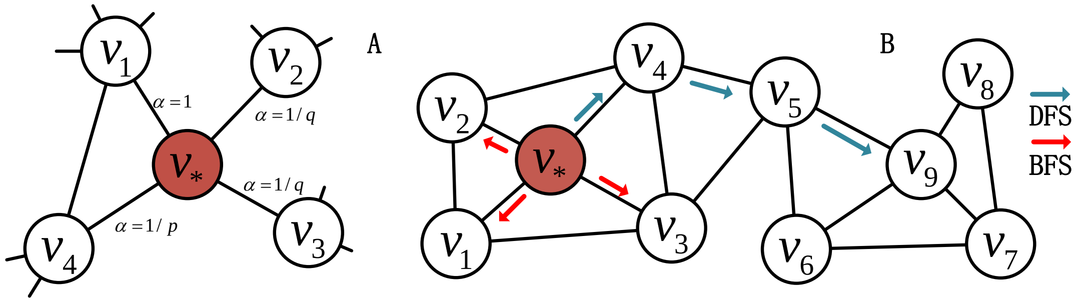

Figure 4: Node sequence sampling schematic. Panel A shows the probability schematic of the next wandering of the node v_*. PanelB shows the generation of city sequence according to the method of BFS and DFS.

Table 1
Various attributes of Tencent eMap POI

ID	POI Name	Address	Lat	Lng	County	City	POI Type	Phone Number
1	Baiyun Mountain	Dahe Village, Pulpit Town	37.196	113.971	Xingtai County	Xingtai	Entertainment travel	None
2	Baiyunshan Forestry	Baiyun Mountain Forestry	37.117	114.003	Xingtai County	Xingtai	Corporate enterprise	None
3	CNOOC Lincheng Baiyun Road Gas Station	100m north of Jinjiang Garden	37.446	114.533	Lincheng County	Xingtai	Automobile gas station	$Xingtai^A$
4	CNOOC Lincheng Baiyunshan Road Convenience Store	East of Renmin Street	37.446	114.533	Lincheng County	Xingtai	Shopping convenience store	$Wuhan^B$
5	Baby House Kindergarten	Intersection of Xinjiang Street &Baiyunshan Road	37.069	115.678	Qinghe County	Xingtai	Schools kindergardens	None

Note: Although the phone number is public, but for privacy reasons, here we hide the specific number.
TypeA: The POI owner's registered city information can be parsed, but it is the same as the city where the POI is located, and no interaction relationship can be extracted ($Xingtai \rightarrow Xingtai$).
TypeB: The POI owner's registered city information can be parsed, and it is not the same as the city where the POI is located, and the interaction relationship can be extracted ($Wuhan \rightarrow Xingtai$).

After biased sampling of the graph structure, the problem of computing city graph vectors is effectively transformed into the problem of computing city word vectors. The acquired city sequence is considered as a corpus in a natural language processing task, and it is sampled using the algorithm shown in Figure 3. The algorithm is built on the basic assumption that cities (words) in close proximity in the sequence have a higher probability of co-occurrence than cities (words) that are farther apart in the sequence, and we perform training by feeding city pairs to the neural network. As shown in the Figure 2, we demonstrate the process of generating city training samples when the sampling window c is 2.

Our study area is 334 prefecture-level cities in China. To prepare the interaction data necessary for the City2vec algorithm, we collected a total of about 80 million points of interest (POI) data from Tencent Maps[3] across the country. As shown in Table 1, the POI data contain POI name, address description, latitude and longitude, county name, city name, POI type, and voluntary public cell phone number information. The cell phone number in China consists of 11 digits, and the first three digits represent the network identification code, which is used to distinguish different operators. The fourth to seventh digits represent the area code of the cell phone number's attribution, i.e. the area where the cell phone number is handled. This part of the information is very useful, and we build region-code matching libraries to parse this information. The last four bits of data are the user identification code, which is not used in this paper.

Due to cultural differences [4], both email and phone numbers are important in the US, but mobile phone numbers are far more important than email in China. A person hardly ever changes its mobile phone number, and if the registered location of the phone does not match the area where the user is currently employed, this implies spatial mobility from city A (registered location) to city B (place of employment). This flow is rarely noticed compared to migration data calculated based on transport big data (trains, flights), which is more detailed and records long periods of time

[3] https://map.qq.com
[4] http://www.cac.gov.cn/2017-08/25/c_1121541921.htm

(implying migration of the population) and long distances (a long trip in the transport network may be divided into several small segments, such as changing to different modes of transport, transferring and transiting at transport hubs), and our article is also based on the similarities and differences between these two types of spatial flows at the temporal and spatial scales (traffic flow and mobile flow). Among all the POI data, there are about 22.66 million data containing specific phone numbers (phone number is not empty), and 2,662,596 directed links (interaction information generated by the inconsistency of two addresses) are extracted, covering 334 major prefecture-level cities in China. As shown in Figure 5, this interaction information portray the migration and mobility of the population between cities. We call this spatial interaction network as Mobile Network.

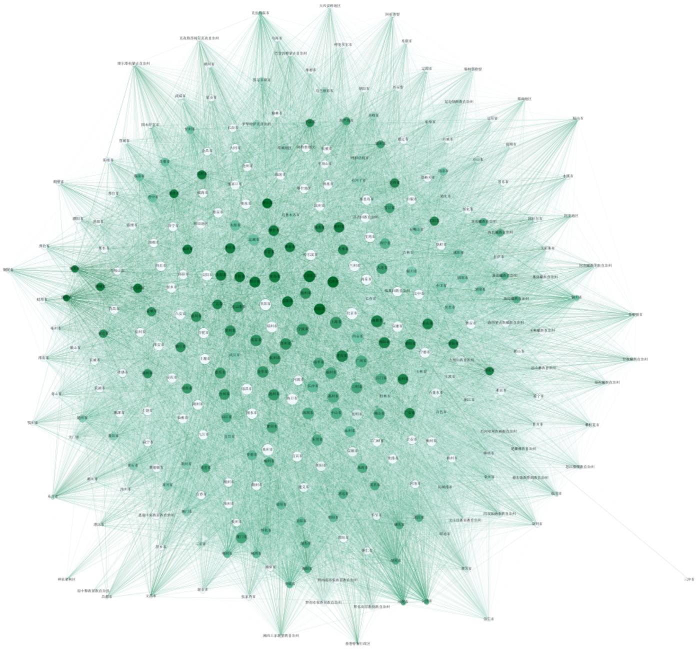

Figure 5: Schematic diagram of the constructed China Mobile Network (only those connected with edge weights greater than 500 are extracted for display, nodes represent different cities, the colors of points and edges represent different weights, the complete network has 334 nodes and 2,662,596 edges.The interaction refers to the OD relationship between the registered city and the current city of residence recorded by the POI, the more records means the stronger the human interaction between the two cities).

In addition,in order to verify whether the City vector contains information on economic properties, we also counted a variety of socio-economic indicators for each city. Those indicators were collected from the China Statistical Yearbook, Suomi National Polar-Orbiting Partnership's Visible Infrared Imaging Radiometer Suite (NPP/VIIRS) satellite night light remote sensing data [5], land use type data, MODIS NDVI vegetation index data, traffic road network data, demographic data and administrative boundary data. Among them, the spatial resolution of NPP/VIIRS nighttime light data reaches 15 second (about 450 m), the adopted wide-angle radiosonde eliminates the light oversaturation phenomenon and enhances the detection sensitivity, and the in-orbit check procedure further improves the image clarity. We use the total number of lights in the area (total intensity) to reflect the lighting characteristics of the area, and construct the total number of lights in the area (TNLI), whose calculation formula is:

$$TNLI = \sum_{i=1}^{n} DN_i \tag{4}$$

DN_i is the image element radiation value of the raster in the region i, and n is the total number of image elements in the region. The night-light remote sensing data can measure the level of regional economic development and observe human activities to some extent (Levin et al., 2020). Based on the above data we plotted the distribution of four

[5] ladsweb.nascom.nasa.gov

indicators as shown in Appendix C Figure 14.

4. Case study

In this section, we first trained the embedding representation of cities and performed a clustering analysis to identify the major urban agglomerations. Secondly, we compare the differences between mobile network and traffic network and find that mobile network is better at capturing long-term and multi-scale population migration states. Finally, based on this feature, we exploit its socio-economic attributes and calculate the spectrum map.

4.1. Training city vectors and identifying city clusters

Human language has thousands of words, and when word vector training is performed, the training dimension is often set to 100 or even 300 dimensions to capture more word features. However, Mobile Network has only hundreds of nodes, and the network information is less informative compared to the semantics of human language. we set the city vector dimension to be about 1/15 of the number of network nodes (20 dimensions); considering that the number of cities in a city cluster is generally not more than 10, we set walk_length and the scan window size to 10, which can capture long-distance interaction information; in order to make our city vector fully trained, we set the number of walks per node to 200; to accelerate the model training, we open 4 threads for synchronous calculation; the wandering parameters p,q refer to the literature (Meng and Masuda, 2020) for the default settings.

The Mobile Network is sampled and computed by the neural network to obtain the embedding representation of cities shown in Figure 6.We used the Kmeans method and calculated the Silhouette Coefficient to determine the optimal number of clusters (best clusters=27, we calculated the Silhouette Coefficient for the number of clusters from 2 to 50). To verify the effectiveness of the City2vec method and reveal the knowledge implicit in the Mobile Network, we conducted a controlled experiment. Group 1, based on city vectors, uses K-Means algorithm and HDBSCN algorithm to identify city clusters, and Group 2, based on traditional methods to parse the structure of Mobile Network, community structure detection algorithm (Louvain algorithm) and spectral clustering algorithm to identify city clusters, respectively. It was found that K-Means+City2vec achieved the best results in the identification and detection of the three major urban agglomerations in the middle reaches of Yangtze River (Table 2).

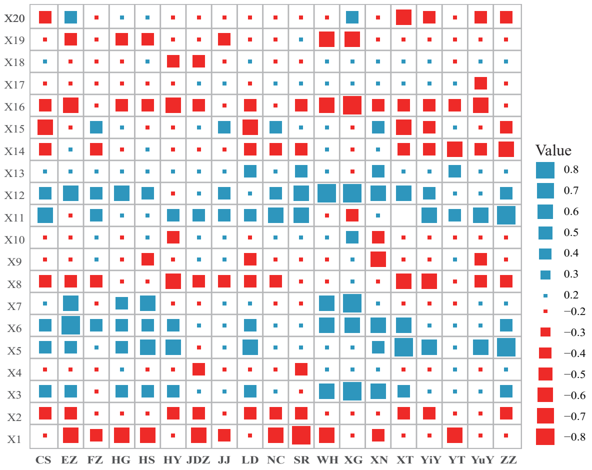

Figure 6: City embedding calculation results of twenty dimensions (The horizontal axis denotes the city dimension and the vertical axis denotes the observation dimension, and we obtained vectors for all 334 cities, and here only the vectors for the major cities in the central reaches of the Yangtze River are shown. There are three main urban agglomerations, 1. Changsha,CS, Hcngyang,HY, Loudi,LD, Yiyang,YiY, Yueyang,YuY, Zhuzhou,ZZ; 2. Ezhou,EZ, Huanggang,HG, Huangshi,HS, Wuhan,WH, Xiaogan,XG, Xianning,XN, Xiangtan,XT; and 3. Fuzhou,FZ, Jingdczhen,JDZ, Jiujiang,JJ, Nanchang,NC, Shangrao,SR, Yingtan,YT, where cities within urban agglomerations have approximate values in some dimensions).

Figure 7 shows the clustering schematic of the city vector (using K-means clustering), and we also downscaled the city vector using the method of principal component analysis (PCA) (Peng et al., 2021). Similar cities are classified into one cluster, and cities in different clusters in the two-dimensional space are still well distinguished. Then we use the K-means algorithm to label each city, and cities with the same label are classified into a city cluster with a high degree of similarity within the city cluster. Compared with the cross-province movement of population, the intra-province movement of population is more frequent. The results of our clustering have a considerable degree of consistency with real government urban agglomerations plans. Most of the urban agglomerations are broadly similar to administrative boundaries (labels=1,3,4,5,8,9,10,12,14,15,17,19,23,26), some of them share similar cultural practices (labels=7,13,16,20,24,25), and the rest of them belong to smaller but geographically connected subcultural circles (Outliers).

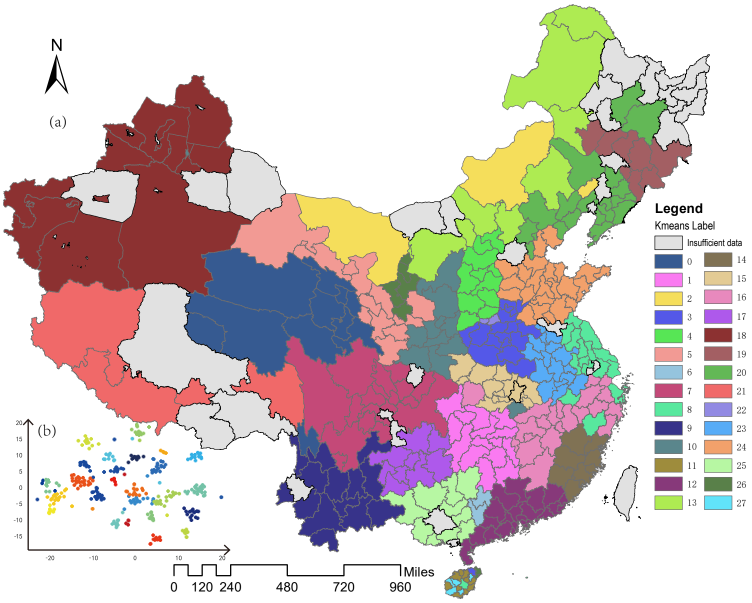

Figure 7: Results of using K-means clustering to identify urban agglomerations (subplot b represent the posterior distribution of city vectors downscaled from 20 dimensions to two dimensions, and the main plot a represents the spatial distribution of different city clusters).

In addition, we carried out a comparative analysis of the cities in the three major urban agglomerations identified in the middle reaches of the Yangtze River (labels=1,15,16). Given that the Closeness Centrality indicator shown in Figure 11 allows for a better coupling of the various social indicators, we ranked the major cities in the central reaches of the Yangtze River according to Closeness Centrality. Wuhan (0.97) and Changsha (0.949) are 1 and 2, which are in fact the provincial capitals of their provinces and the only cities in the cluster with a GDP of over RMB 1 trillion (Table2). Surprisingly, Yingtan (0.922) is more important in the mobile network than the provincial capital Changsha (0.919). As an important copper industrial base and railway hub in China, it attracts a large floating population.

We used the K-means algorithm, the Dbscan algorithm, and the complex network community discovery algorithm to cluster the vector of cities in the central reaches of the Yangtze River, and the clustering results were compared with real city cluster classification criteria. It can be seen from Table 2 that the K-means algorithm works the best, misclassifying only Xianning city. The complex network community discovery algorithm based on the Louvain algorithm incorrectly classified two cities (Meng et al., 2020), and the Dbscan algorithm the worst performer. The good clustering results in Table 2 also imply that the trained city vectors adequately learn information about city interactions as well as socio-economic characteristics. People tend to choose cities with high similarity to the city they currently live in as their destination, which also offers the possibility of city recommendation.

4.2. Structure Comparison of Mobile Network and Traffic Network

To verify the reliability of our network, we consider Tencent traffic migration data as a comparison set. Tencent's traffic data platform counts the total passenger traffic between cities including trains, automobiles, and airplanes in a year [6]. We call the city network formed by Tencent's urban traffic migration data a Traffic Network.

[6] https://heat.qq.com/qianxi.php

Table 2
Comparison of indicators of key cities in the central reaches of Yangtze River urban agglomeration.

City	Pop	Light	Water Supply	Annual Electricity	Kmeans Label	Dbscan Label	Modular Class	Degree Centrality	Closeness Centrality	Harmonic Centrality	Betweeness Centrality	Clustering
Wuhan	884	155.798	115.958	580.337	🟢	🟢	🟢	48.38	0.97	0.984	235.239	0.688
Xiaogan	518	17.403	4.087	132.063	🟢	⚪	🟢	8.066	0.829	0.897	237.755	0.734
Huangshi	273	15.809	7.058	137.406	🟢	🟢	🟢	4.821	0.761	0.843	47.756	0.761
Huanggang	741	21.195	3.306	122.596	🟢	🟢	🟢	8.692	0.804	0.878	72.541	0.75
Ezhou	111	9.62	6.225	67.15	🟢	⚪	🟢	2.664	0.695	0.781	13.571	0.814
Xianning	305	12.442	3.507	88.726	🔴	⚪	⚪	6.792	0.802	0.876	70.357	0.751
Nanchang	532	50.521	35.733	230.001	🟡	🟡	🟡	24.423	0.919	0.956	204.721	0.703
Jiujiang	523	27.229	7.943	197.61	🟡	🟡	🟡	11.585	0.86	0.918	116.665	0.719
Jingdczhen	170	8.839	4.092	57.703	🟡	🟡	🟡	5.697	0.761	0.843	50.7	0.763
Shangrao	789	19.399	6.42	157.923	🟡	⚪	🟡	12.653	0.886	0.936	77.178	0.734
Fuzhou	432	14.897	5.401	85.034	🟡	🟡	🟡	8.755	0.86	0.918	50.654	0.741
Yingtan	129	6.122	2.304	47.931	🟡	⚪	🟡	6.238	0.922	0.958	46.839	0.726
Changsha	729	99.3	65.054	363.693	🔴	🔴	🔴	40.218	0.949	0.973	231.182	0.687
Zhuzhou	403	22.537	15.233	119.124	🔴	🔴	🔴	9.681	0.824	0.893	42.536	0.755
Xiangtan	289	23.821	7.772	120.182	🔴	🔴	🔴	7.378	0.756	0.839	25.216	0.78
Yueyang	569	23.369	6.294	149.008	🔴	⚪	🔴	10.55	0.853	0.914	62.895	0.736
Hcngyang	801	17.486	11.584	150.689	🔴	⚪	⚪	9.493	0.833	0.9	57.434	0.743
Yiyang	478	9.976	5.259	81.456	🔴	🔴	🔴	7.138	0.824	0.893	70.189	0.751
Loudi	455	12.586	4.775	142.596	🔴	🔴	🔴	5.023	0.814	0.886	27.246	0.768

Note1: Population / Ten Thousand People, Water Supply / Ten Million Cubic Meters, Electricity Consumption / Billion kWůh, Light and Weighted DegreeCentrality / Thousand.

Note2: Wuhan City Cycle 🟢; Poyang Lake City Group 🟡; Changsha-Zhuzhou-Xiangtan Megalopolis 🔴; Misclassified ⚪

It can be seen that cities in the Traffic network have strong scale-free characteristics, and although the overall degree of the network is low (Mean Degree=122), there are a few key cities that play a dominant role in the traffic network and hold most of the traffic (Figure 8 d,e). In contrast, the overall degree of the Mobile Network is high (Mean Degree=420), and there are no significant outliers, and the difference between key cities with high degrees is not particularly significant compared to the average cities. This is due to the fact that Traffic Network records the short-term movement of the population, and transportation hub cities take on the function of traffic "distribution", which greatly strengthens the core position of such cities. The Mobile Network is driven by socio-economic factors, mainly economic, policy, natural environment and employment attraction (e.g., the rapid economic development of a certain city attracts residents of other cities to go to work; the warm and pleasant environment of a certain city attracts residents of other cities to go to do business.). The population in a Mobile Network moves in both directions, i.e., from large cities to small cities and from small cities to large cities (as shown in Figure 8 f).

In order to compare the differences between Mobile network and Traffic network, we downscaled the city vector using PCA method, and we can see (Figure 9) that the city nodes of Traffic network (red) are more discrete than Mobile network (blue). This is also consistent with the results in Figure 8(b,c), where the average distance from cities in Mobile Network to the network prime vector is 0.631, which is higher than that of Traffic network (0.586), showing a more compact distribution. The cities in the middle reaches of the Yangtze River (core cities are Changsha, CS, Wuhan, Nanchang, NC) are more concentrated in Figure 9 (confidence ellipse 1, 2), showing a strong correlation. Mobile network is more "equal" than Traffic network (Figure 8 e,f). The traffic data will frequently record the interaction between two cities, reflecting the short-term population migration, which actually raises the weight of the traffic hub cities. Our Mobile network is built on the basis of relatively stable POI, which is closer to the long-term population migration (permanent migration).

We calculated Pearson correlation coefficients between the city vectors of key cities in the central reaches of the Yangtze River urban agglomeration. As shown in Figure 10, it is clear that the similarities between cities within urban agglomerations are much higher than the similarities between cities in different urban agglomerations, forming three distinctly different sub-urban agglomerations.

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{5}$$

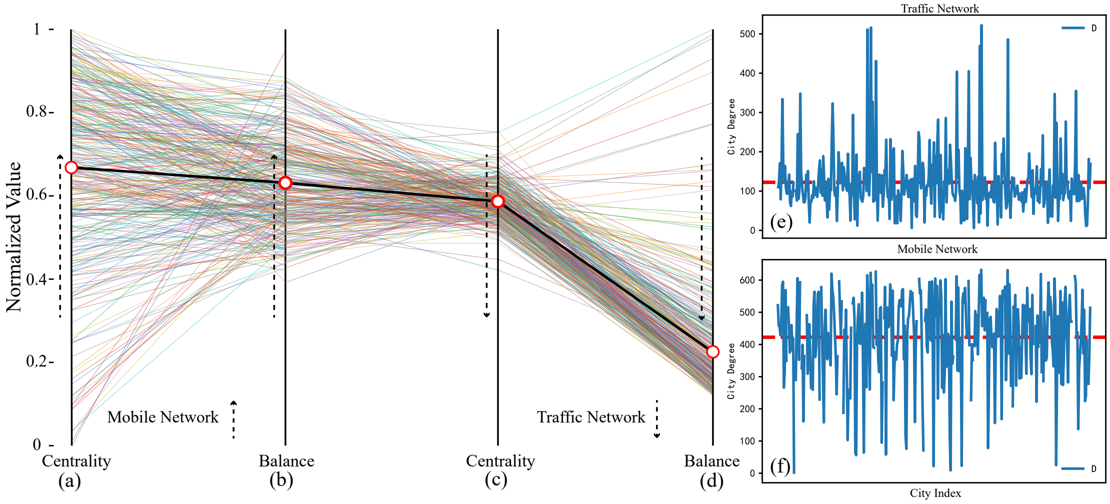

Figure 8: Properties of Cities in Mobile Network vs. Traffic Network (a,c denote the normalized centrality of city; b,d denote the balance degree of city, which is the normalized distance to the network center of mass vector. e,f are the distributions of degree centrality in different networks, respectively).

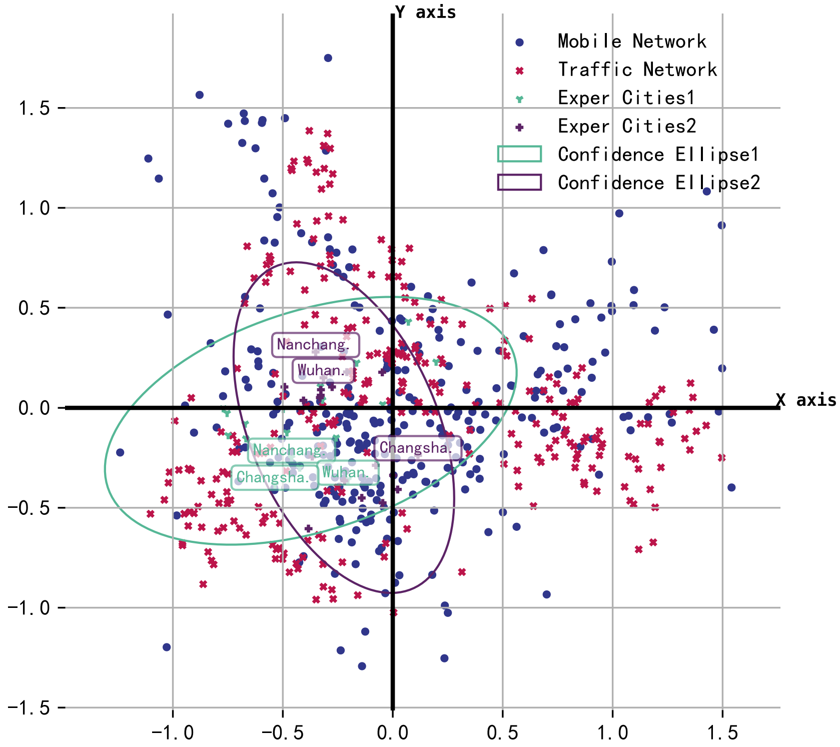

Figure 9: The results of city node dimensionality reduction visualization of Mobile Network and Traffic network (we also plot the confidence ellipse 1, Mobile Network, confidence ellipse 2, Traffic network of key cities in the middle reaches of Yangtze River urban agglomeration).

It can be seen (Figure 10) that the intra-group similarity of Traffic network (b) is more significant. And the correlation between groups is very weak.This means that Traffic network has a distance attenuation effect and focuses more on spatial interactions at close distances. While Mobile network (a) is good at capturing long-distance spatial

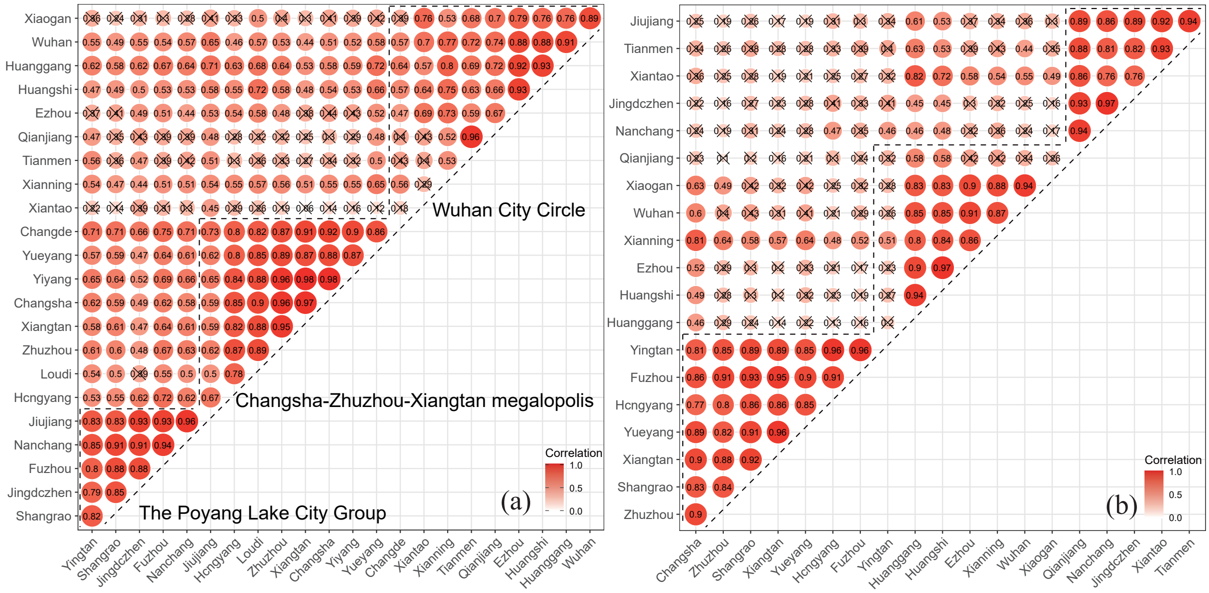

Figure 10: City vector similarity results of Mobile Network (a) vs. Traffic network (b) (divided into three sub-city clusters based on similarity between city vectors, Pearson correlation coefficients calculated for city vectors, x represents failure of Pearson's correlation test at a confidence level of $p = 0.05$).

associations in addition to short-distance interactions. We can summarize the main features of these two networks from Figure 8,9,10. Traffic network can be used to capture the short-term, short-distance migration status of the population, which can be applied to the risk assessment of public health events as well as natural disaster events (Lee et al., 2021); while the Mobile Network (a) proposed in this paper can be used to capture the long-term, multi-scale migratory state of a population, could be used to analyze the complex coupling effects of networks and economies, which is what we analyze in the next section (4.3).

4.3. Socio-economic, geographic and complex network attributes of coupled city vectors

Considering that the main driver of population migration is economic, we fit the socioeconomic indicators to the complex network attributes (Lao et al., 2016). We start with an analysis of a part of cities, about hundreds of cities in 10 urban agglomerations, which are, respectively, Changsha-Zhuzhou-Xiangtan Megalopolis (label=1), Central Plains Urban Agglomeration (label=3), Taiyuan Urban Agglomerations (label=4), Chengdu-Chongqing City Group (label=7), Yangtze River Delta Urban Agglomerations (label=8), Guanzhong Plain City Group (label=10), the Pearl River Delta City Group (label=12), Urban Agglomeration on the West Side of the Straits (label=14), Wuhan City Cycle (label=15), The Poyang Lake City Group (label=16). The linear fit results of Figure 11 (a-d) illustrate that the degree values of city nodes have an extremely strong positive correlation with the economic attributes at the $p = 0.01$ confidence level.

Figure 11 (e-h) shows box plots of each urban indicator for different urban agglomerations. Figure 11 (e) shows that Central Plains Urban Agglomeration (label=3) has the highest average urban population, Taiyuan Urban Agglomerations (label=4) has the lowest average urban population, and Yangtze River Delta Urban Agglomerations (label=8) has the highest urban population limit. Figure 11 (f) illustrates that the Yangtze River Delta Urban Agglomerations (label=8) and the Pearl River Delta City Group (label=12) have the highest water consumption, and these two urban agglomerations are also the economic and industrial centers of China, and require large amounts of industrial & agricultural water. Figure 11 (g) shows that the Yangtze River Delta Urban Agglomerations (label=8) are still the most energy intensive urban group in China. The Poyang Lake City Group (label=16) also has a high energy consumption due to the large amount of rare earth and metal minerals industry. Figure 11 (h) shows the total night light remote sensing of different city groups. The central reaches of Yangtze River city group is still the "brightest" region in China, reflecting the active socio-economic activities, while the resource-based Poyang Lake City Group (label=16) has a large difference in brightness between cities.

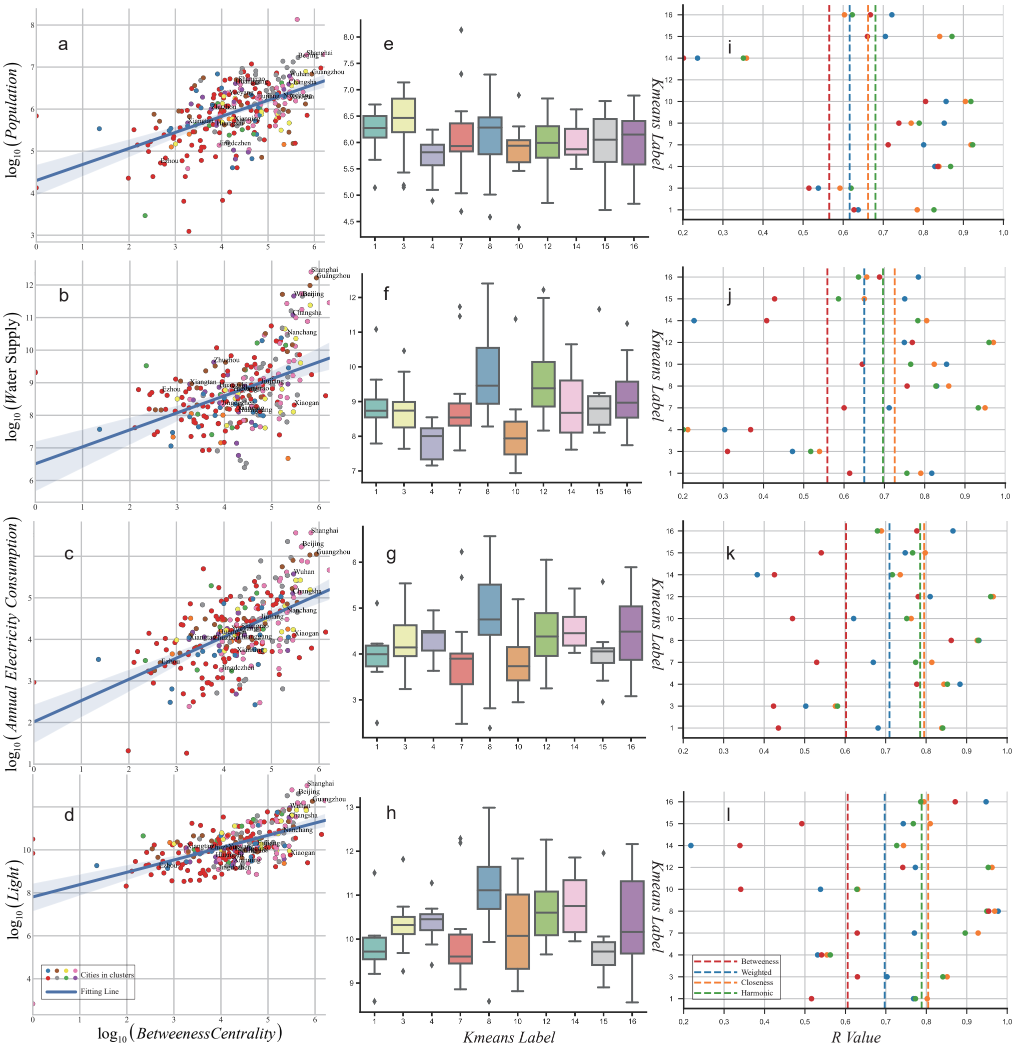

Figure 11: Interaction of socio-economic indicators with complex network attributes at the scale of urban agglomerations. (Fitted curves of network attributes and city metrics: a-d is the fit of Betweeness centrality with the logarithm of population, water supply, annual electricity consumption, and light total for 334 cities in China; Box line plots of urban indicators for different urban clusters: e-h is the box plot of the four indicators for the selected 10 urban agglomerations;Fitting degree of different urban indicators for different urban clusters: i-l, the horizontal axis represents the goodness of fit of the linear regression, the vertical axis represents different city groups, and the dotted data represents the fitted R-value between a certain type of centrality of a city group and the indicator, the blue dots represent Weighted Degree Centrality, orange dots represent Closeness Centrality, green dots represent Harmonic Closeness Centrality, red dots represent Betweeness Centrality, and the line represents the mean value of different centrality in all urban agglomerations, the data were logarithmically processed before fitting.)

As Figure 11 (i-l) shows the results of fitting urban network attributes to socio-economic indicators at the scale of urban agglomerations, Closeness Centrality and Harmonic Closeness Centrality (yellow dashed line and green dashed line) fit best with night remote sensing and various socio-economic indicators from the Statistical Yearbook. The Closeness Centrality of the entire graph has a weighted average fitted R-value of approximately 0.8 for all indicators, proving that the urban Mobile Network is informative. The importance of cities in the Mobile Network is highly correlated with total urban light and urban energy consumption, and less correlated with total population and water supply, implying that developed cities create more total economic output with fewer resources and population.

In addition, experiments were carried out to verify whether the city vector contained geographical features. We use

Pearson correlation coefficients to calculate the similarity between cities two by two, and generate city spectra based on cities with different extremes. For example, the city vector $City_N$ of the most northern city and the city vector $City_S$ of the most southern city are used as the two ends of the spectrum, and the spectral value of each city $City_i$ is $r_{City_N City_i} - r_{City_S City_i}$,$r$ denotes the calculation of the Pearson correlation coefficient), so that the generated vector can be verified to contain the geographic location attribute.

As shown in Figure 12, we calculated seven sets of spectra, where each vertical line in the spectrum represents a city, and the black lines represent the spectral mean, indicating the propensity of the mobile population to the city (the degree of recognition of the city). We can divide Figure 12(a,b) into one category, representing the two largest urban agglomerations in China in terms of geographic space, namely the Pearl River Delta urban agglomeration (core cities are Guangzhou,GZ and Shenzhen,SZ) and the Bohai Rim urban agglomeration (core cities are Beijing,BJ and Tianjin,TJ). To measure the propensity to migrate to the core cities under the same urban agglomeration. Figure 12(c,d) shows the comparison between coastal (Shanghai,SH) and inland (Wuhan,WH), and between eastern (Shanghai,SH) and western (Chengdu,CD), respectively. As can be seen, the core cities of the urban agglomerations (Figure 12 e,f), have little difference in their attractiveness to the population, with a more neutral urban tendency (black median lines). Pearl River Delta urban agglomeration is clearly more developed than Bohai Rim urban agglomeration and has a more similar source of population.

In contrast, the urban tendency (black median line) in the comparison group (Figure 12 c,d) has shifted significantly, indicating that the coastal cities are more attractive to the population than the inland cities and western cities, showing a clear geospatial difference.

Figure 12 (e,f) represents the urban tendency in the middle reaches of Yangtze River urban agglomeration. Although Wuhan (WH) has a higher GDP than Changsha (CS), Changsha (CS) has a stronger ability to attract population than Wuhan (WH). According to the data of the 7th National Census in 2021, Changsha added more than 3 million people in the past 10 years (2010-2020), while Wuhan added 2.542 million people in 10 years, which is significantly lower than Changsha, and our study strongly supports this conclusion.

In addition, as shown in (Figure 12, g), we mark northern cities (cities north of 35 degrees north latitude, 35 degrees being the natural dividing line in China and the 800 mm annual precipitation dividing line) in red and southern cities (cities south of 35 degrees north latitude) in blue. Beijing and Guangzhou are the most representative large cities in the south (GZ) and north (BJ) of China, respectively, and we choose these two cities as the extremes. It can be seen that the spectrum of the northern city is more biased towards Beijing, and the spectrum of the southern city is clearly biased towards Guangzhou, while the mean value of the spectrum lies around 0. This implies that the vectors we generate contain location information, which can be clearly corroborated in the spectral maps.

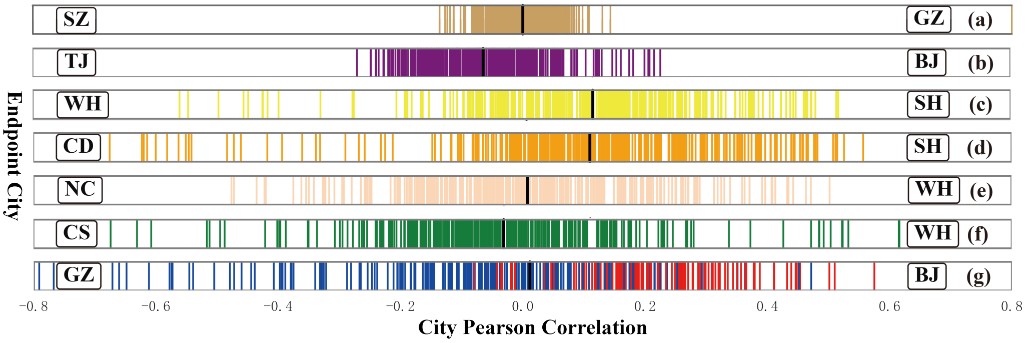

Figure 12: City spectrum based on Pearson's correlation coefficient (a,b indicate competition between central cities within the same urban area; c,d indicate competition between coastal cities and inland cities; e,f indicate competition between inland centres, and g indicate competition between cities in the north and south. A more concentrated spectrum indicates a more mature urban agglomeration and a blurring of the boundaries between cities, with the black median line indicating the relative competitiveness between city pairs).

The enhancements of our method compared to traditional network analysis methods are as follows:

1. As shown in the Figure 12. We could calculate the similarity between cities and the results of the "vote" of more than 300 cities across the country for that city pair, to fully explain the level of competitiveness of the city and the origin of its population. This is far more intuitive and accurate than the boring comparison of the various types

of centrality of cities in Table 2. 2. we can assess the development of urban agglomerations (compactness of the network). A highly developed urban agglomeration gradually "smoothes out" the boundaries between cities, and the "voting" results of other cities outside the agglomeration become more concentrated. As shown in the Figure 12 (a, d), city cluster a (with core cities SZ, GZ) is significantly more integrated than city cluster d (with core cities TJ, BJ). 3. The city vectors we calculated also contain geographical and climate attributes. As shown in spectral plot g, there is a clear difference between southern and northern cities, indicating that the distance between cities & home and climatic conditions will be taken into account when humans make migration decisions. In addition, the results also contain the impact of distance. As Wuhan and Changsha are each the central cities of their provinces. Figure 12 f also reveals that the economic level and population attractiveness of the two cities are not very different, but due to distance differences, residents moving out of smaller cities prefer the closer central cities.

5. Conclusion and outlook

This research constructs a spatial interaction network (Mobile Network) based on a large number of urban co-occurrence relations provided by the floating population and trains urban vectors embedded with various kinds of knowledge. We extend the one-dimensional sensing of cities (i.e., describing cities by computing various attributes or urban economic indicators through traditional complex network methods) to two-dimensional sensing (i.e., one can vectorise network nodes with multiple dimensions to compare different cities). Our study confirms that the city embedding algorithm achieves better classification results than the traditional network community delineation algorithm in the city agglomerations identification task. Trained city vectors are good at capturing long term and long distance migration information and are enriched with socio-economic attributes, geographic location attributes and complex network attributes. Our research focus more on the economic aspects, due to the fact that data are derived from electronic maps, which highly relevant to business. And more urban information (e.g., urban spatial location, urban commentary data, urban streetscape data) can be introduced in the future to expand our understanding of population migration in a different perspective.

A. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

B. Acknowledgments

This research was supported by the National Key R&D Program (no.2018YFB2100500), National Nature Science Foundation of China (nos. 41971351, 41771422, 41890822), and China Scholarship Council (202106270048). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

C. Supplementary materials

In order to measure the dominance of the urban agglomeration core city over the neighbouring cities, we have proposed the indicator $influence$. It calculates the flow between the core city and the surrounding cities as a proportion of the total flow in the surrounding cities, with the following formula:

$$Influence_{i,j} = \frac{Flow_{i,j}}{\sum\limits_{k=1}^{k=cities} Flow_{k,j}\,(j \neq k)}\,(i \neq j) \tag{6}$$

$Flow_{i,j}$ denotes the migration scale between city i and city j, and $\sum\limits_{k=1}^{k=cities} Flow_{k,j}\,(j \neq k)$ denotes the total population migration size of city j. The core cities of the urban agglomeration belong to the hub node in the network and have higher dominance over the ordinary cities. As can be seen in Figure 13, the population migrates further to the central cities. Tianmen, Qianjiang and Xiantao as three county-level cities are smaller in size and extremely closely connected to each other, and can be considered as a whole city for analysis.

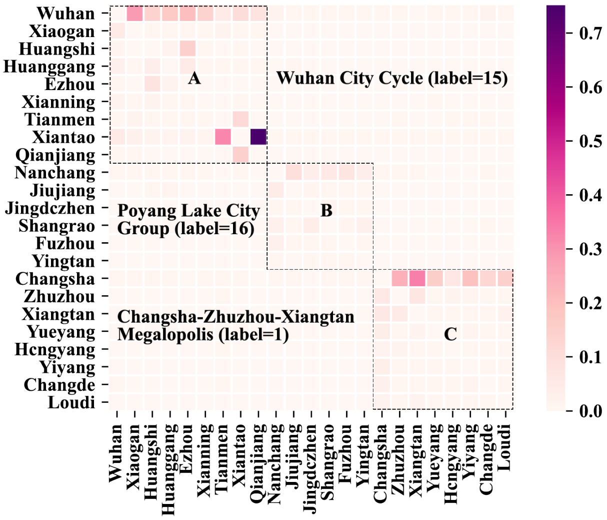

Figure 13: Spatial equality and dominance of urban nodes in urban agglomerations.

D. Computation of complex network properties

Complex networks are topological abstractions of a large number of real complex systems, between regular networks and random networks, and similarly, population migration between cities is a typical complex network system (Li et al., 2017; Dong et al., 2020). We introduce several types of centrality indicators of complex networks to assess them. Cities with higher centrality tend to have more economic activity and can provide a large number of jobs. These cities have a strong attraction to incoming populations as well as a high spatial influence. At the same time, these cities are also the most vulnerable nodes, and disruptions to these cities can disrupt the interactive links between cities across the network (Zhang et al., 2020c).

Indicator 1: Degree centrality reverses the number of edges in the entire complex network. Relatively speaking, the more edges connected to a node, the more important that node is. A node with more edges (i.e. the more "important" cities) usually looks like a "transport hub". The Degree centrality of node x_i is calculated as Equation 7.

$$C_D\left(x_i\right)=\frac{\deg\left(x_i\right)}{n-1} \tag{7}$$

Indicator 2: Closeness centrality indicates how close a node is to all other nodes in the complex network. This metric measures how easy it is for a node to reach other nodes, by capturing the inverse of the average of the distances to all nodes in a complex network. The Closeness centrality of node x_i is calculated as Equation 8.

$$C_C\left(x_i\right)=\frac{n-1}{\sum\limits_{j\neq i} dist\left(x_i,x_j\right)} \tag{8}$$

Indicator 3: The Harmonic centrality reverses the sum and reciprocal operations in the definition of closeness centrality. The Harmonic centrality of node x_i is calculated as Equation 9.

$$C_H\left(x_i\right)=\frac{1}{n-1}\sum_{j\neq i}\frac{1}{dist\left(x_i,x_j\right)} \tag{9}$$

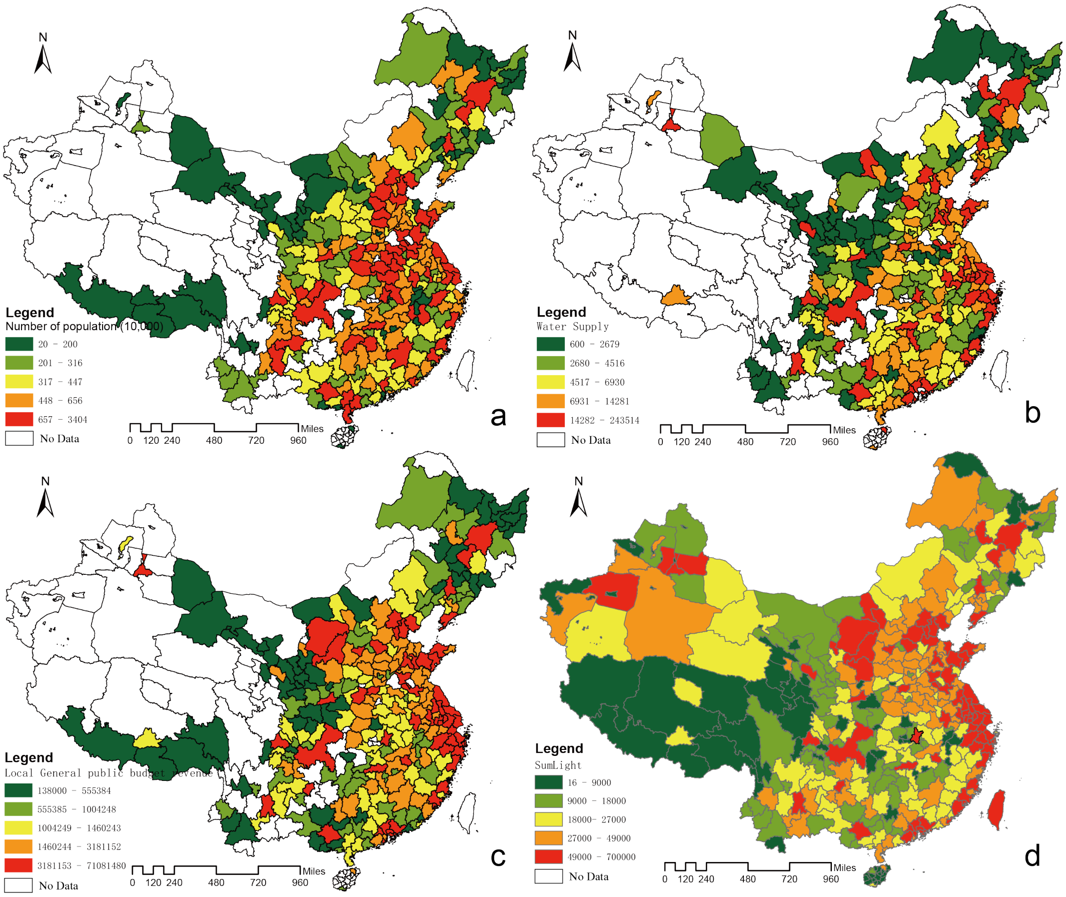

Figure 14: Distribution of indicator values for each city in China (a, number of people in the city b, water supply c, local financial budget d, total amount of lights in terms of image brightness).

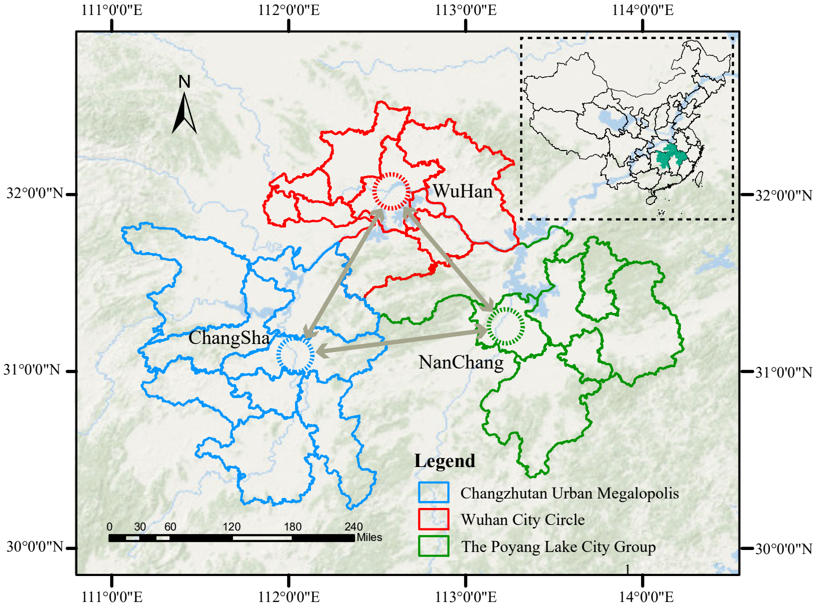

Figure 15: Overview of the central reaches of the Yangtze River urban agglomeration.

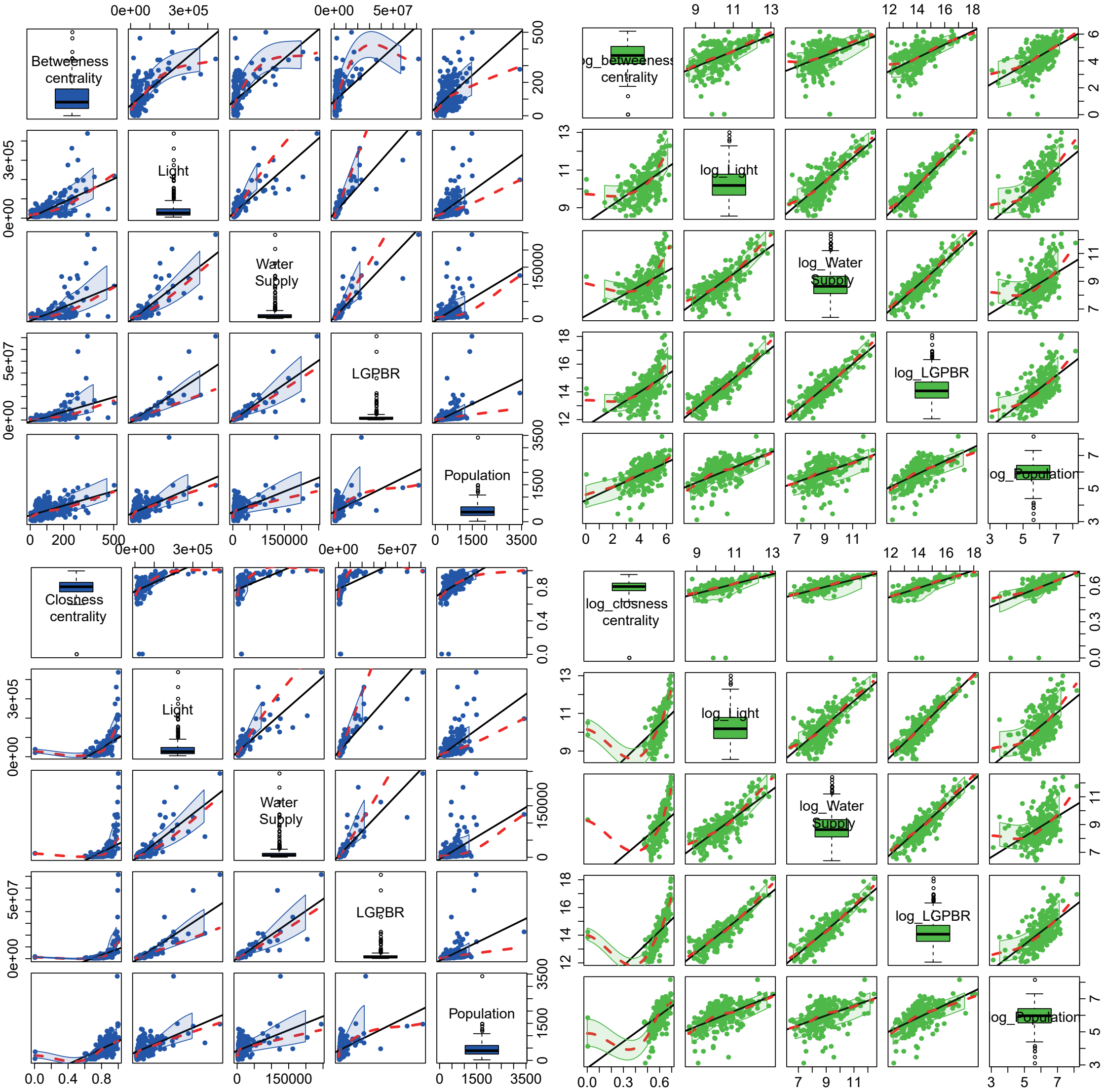

Figure 16: Linear fit & curve fitting results of complex network indicators to socio-economic indicators.

Among them, $x_i \in V$, $n = |V|$ and $dist\left(x_i, x_j\right)$ means the distance from node x_i to node x_j.

Indicator 4: Betweenness centrality of node x_i indicates the number of shortest paths through the node between all pairs of nodes in a complex network. The more nodes the complex network has with high betweenness centrality, the more efficient it is. The Betweenness centrality of node x_i is calculated as Equation 10.

$$C_B\left(x_i\right) = \frac{2\sum\sum\frac{g_{jk}(i)}{g_{jk}}}{(n-1)(n-2)} \tag{10}$$

With $g_{jk}(i)$ is the number of geodesics (Shortest Route) from node x_j to node x_k containing x_i. Degree centrality describes the "sociability" of a city node and does not consider the dominance over other cities. Betweenness centrality measures the dominance of a node and is the intermediary node for communication between two cities. Closeness centrality is more about the independence of the node and calculates the ability of the node not to be controlled by other nodes. We perform the calculation of various metrics of the network on the Gephi platform.

References

Singleton, J.D.. Sorting charles tiebout. History of political economy 2015;47(suppl_1):199–226.

Lu, C., Wu, Y., Shen, Q., Wang, H.. Driving force of urban growth and regional planning: A case study of china's guangdong province. Habitat international 2013;40:35–41.

You, Z., Yang, H., Fu, M.. Settlement intention characteristics and determinants in floating populations in chinese border cities. Sustainable cities and society 2018;39:476–486.

Zhu, Y., Lin, L.. Studies on the temporal processes of migration and their spatial effects in china: progress and prospect.. Scientia Geographica Sinica 2016;36(6):820–828.

Yang, X., Fang, Z., Xu, Y., Yin, L., Li, J., Lu, S.. Spatial heterogeneity in spatial interaction of human movementsinsights from large-scale mobile positioning data. Journal of Transport Geography 2019;78:29–40.

Li, L., Derudder, B., Kong, X.. A machine learning approach to the simulation of intercity corporate networks in mainland china. Computers, Environment and Urban Systems 2021a;87:101598.

Bi, H., Ye, Z.. Exploring ridesourcing trip patterns by fusing multi-source data: A big data approach. Sustainable Cities and Society 2021;64:102499.

Xu, D., Wu, X.. Separate and unequal: hukou, school segregation, and educational inequality in urban china. Chinese Sociological Review 2022;:1–25.

Zheng, Y., Zhang, X., Dai, Q., Zhang, X.. To float or not to float? internal migration of skilled laborers in china. International Journal of Environmental Research and Public Health 2020;17(23):9075.

Hong, J., Tang, M., Wu, Z., Miao, Z., Shen, G.Q.. The evolution of patterns within embodied energy flows in the chinese economy: A multi-regional-based complex network approach. Sustainable Cities and Society 2019;47:101500.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., et al. Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach. Computers, Environment and Urban Systems 2021;87:101619.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al. Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers 2015;105(3):512–530.

Li, Z., Huang, X., Ye, X., Jiang, Y., Martin, Y., Ning, H., et al. Measuring global multi-scale place connectivity using geotagged social media data. Scientific reports 2021b;11(1):1–19.

Zhang, Y., Chen, Z., Zheng, X., Chen, N., Wang, Y.. Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. Journal of Hydrology 2021a;603:127053.

Zhang, Y., Zheng, X., Chen, M., Li, Y., Yan, Y., Wang, P.. Urban fine-grained spatial structure detection based on a new traffic flow interaction analysis framework. ISPRS International Journal of Geo-Information 2021b;10(4). URL: https://www.mdpi.com/2220-9964/10/4/227. doi:10.3390/ijgi10040227.

McKenzie, G., Adams, B.. A data-driven approach to exploring similarities of tourist attractions through online reviews. Journal of Location Based Services 2018;12(2):94–118.

Hui, E.C., Li, X., Chen, T., Lang, W.. Deciphering the spatial structure of china's megacity region: A new bay areathe guangdong-hong kong-macao greater bay area in the making. Cities 2018;.

Hu, Y., Ye, X., Shaw, S.L.. Extracting and analyzing semantic relatedness between cities using news articles. International Journal of Geographical Information Science 2017;31(12):2427–2451.

Kaluza, P., Kölzsch, A., Gastner, M.T., Blasius, B.. The complex network of global cargo ship movements. Journal of the Royal Society Interface 2010;7(48):1093–1103.

Zhang, X., Zheng, Y., Zhao, Z., Ye, X., Zhang, P., Wang, Y., et al. The education-chasing labor rush in china identified by a heterogeneous migration-network game. Scientific reports 2020a;10(1):1–17.

Chen, N., Zhang, Y., Du, W., Li, Y., Chen, M., Zheng, X.. Ke-cnn: A new social sensing method for extracting geographical attributes from text semantic features and its application in wuhan, china. Computers, Environment and Urban Systems 2021a;88:101629.

Rocklage, M.D., Rucker, D.D., Nordgren, L.F.. Mass-scale emotionality reveals human behaviour and marketplace success. Nature Human Behaviour 2021;:1–7.

Zhang, Y., Chen, N., Du, W., Li, Y., Zheng, X.. Multi-source sensor based urban habitat and resident health sensing: A case study of wuhan, china. Building and Environment 2021c;:107883.

Xu, L., Chen, N., Chen, Z., Zhang, C., Yu, H.. Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. Earth-Science Reviews 2021;:103828.

Zhou, J., Shen, J., Yu, S., Chen, G., Xuan, Q.. M-evolve: Structural-mapping-based data augmentation for graph classification. IEEE Transactions on Network Science and Engineering 2020;.

Newman, M.E.. The structure and function of complex networks. SIAM review 2003;45(2):167–256.

Lei, Y., Zhou, Y., Shi, J.. Overlapping communities detection of social network based on hybrid c-means clustering algorithm. Sustainable Cities and Society 2019;47:101436.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.. Complex embeddings for simple link prediction. In: International Conference on Machine Learning. PMLR; 2016, p. 2071–2080.

Chauhan, L., Ram, U., Hari, K., Jolly, M.K.. Topological signatures in regulatory network enable phenotypic heterogeneity in small cell lung cancer. Elife 2021;10:e64522.

Hari, K., Ram, U., Jolly, M.K.. Identifying more equal than others edges in diverse biochemical networks. Proceedings of the National Academy of Sciences 2021;118(16).

Von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G., Kerren, A.. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. IEEE transactions on visualization and computer graphics 2015;22(1):11–20.

Wang, P., Zhang, T., Zheng, Y., Hu, T.. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. International

Journal of Geographical Information Science 2022a;:1–27doi:10.1080/13658816.2022.2032081.

Yue, H., Guan, Q., Pan, Y., Chen, L., Lv, J., Yao, Y.. Detecting clusters over intercity transportation networks using k-shortest paths and hierarchical clustering: a case study of mainland china. International Journal of Geographical Information Science 2019;33(5):1082–1105.

Wang, P., Hu, T., Gao, F., Wu, R., Guo, W., Zhu, X.. A hybrid data-driven framework for spatiotemporal traffic flow data imputation. IEEE Internet of Things Journal 2022b;doi:10.1109/JIOT.2022.3151238.

Cvetojevic, S., Hochmair, H.H.. Modeling interurban mentioning relationships in the us twitter network using geo-hashtags. Computers, Environment and Urban Systems 2021;87:101621.

Li, Z., Huang, X., Ye, X., Jiang, Y., Yago, M., Ning, H., et al. Measuring global multi-scale place connectivity using geotagged social media data. arXiv preprint arXiv:210203991 2021c;.

Ouyang, J., Fan, H., Wang, L., Yang, M., Ma, Y.. Site selection improvement of retailers based on spatial competition strategy and a double-channel convolutional neural network. ISPRS International Journal of Geo-Information 2020;9(6):357.

Shao, Q., Liu, X., Zhao, W.. An alternative method for analyzing dimensional interactions of urban carrying capacity: case study of guangdong-hong kong-macao greater bay area. Journal of Environmental Management 2020;273:111064.

Wang, S., Zhang, X., Chen, N., Wang, W.. Classifying diurnal changes of cyanobacterial blooms in lake taihu to identify hot patterns, seasons and hotspots based on hourly goci observations. Journal of Environmental Management 2022c;310:114782.

Wang, Y., Deng, Y., Ren, F., Zhu, R., Wang, P., Du, T., et al. Analysing the spatial configuration of urban bus networks based on the geospatial network analysis method. Cities 2020;96:102406.

Zeng, C., Song, Y., Cai, D., Hu, P., Cui, H., Yang, J., et al. Exploration on the spatial spillover effect of infrastructure network on urbanization: A case study in wuhan urban agglomeration. Sustainable Cities and Society 2019;47:101476.

Zhang, W., Fang, C., Zhou, L., Zhu, J.. Measuring megaregional structure in the pearl river delta by mobile phone signaling data: A complex network approach. Cities 2020b;104:102809.

Yao, Y., Wang, J., Hong, Y., Qian, C., Guan, Q., Liang, X., et al. Discovering the homogeneous geographic domain of human perceptions from street view images. Landscape and Urban Planning 2021;212:104125.

Charyyev, B., Gunes, M.H.. Complex network of united states migration. Computational Social Networks 2019;6(1):1–28.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013, p. 3111–3119.

Le, Q., Mikolov, T.. Distributed representations of sentences and documents. In: International conference on machine learning. PMLR; 2014, p. 1188–1196.

Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.. Tweet2vec: Character-based distributed representations for social media. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany: Association for Computational Linguistics; 2016, p. 269–274. URL: http://anthology.aclweb.org/P16-2044.

Grover, A., Leskovec, J.. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016, p. 855–864.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., Jaiswal, S.. graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:170705005 2017;.

Palumbo, E., Rizzo, G., Troncy, R., Baralis, E., Osella, M., Ferro, E.. Knowledge graph embeddings with node2vec for item recommendation. In: European Semantic Web Conference. Springer; 2018, p. 117–120.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., et al. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. International Journal of Geographical Information Science 2017;31(4):825–848.

Yan, B., Janowicz, K., Mai, G., Gao, S.. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems. 2017, p. 1–10.

Lin, Y., Kang, M., Wu, Y., Du, Q., Liu, T.. A deep learning architecture for semantic address matching. International Journal of Geographical Information Science 2020;34(3):559–576.

Chen, N., Zhang, Y., Du, W., Li, Y., Chen, M., Zheng, X.. Ke-cnn: A new social sensing method for extracting geographical attributes from text semantic features and its application in wuhan, china. Computers, Environment and Urban Systems 2021b;88:101629. URL: https://www.sciencedirect.com/science/article/pii/S0198971521000363. doi:https://doi.org/10.1016/j.compenvurbsys.2021.101629.

Maduako, I., Wachowicz, M.. A space-time varying graph for modelling places and events in a network. International Journal of Geographical Information Science 2019;33(10):1915–1935.

Zhu, D., Huang, Z., Shi, L., Wu, L., Liu, Y.. Inferring spatial interaction patterns from sequential snapshots of spatial distributions. International Journal of Geographical Information Science 2018;32(4):783–805.

Crivellari, A., Beinat, E.. From motion activity to geo-embeddings: Generating and exploring vector representations of locations, traces and visitors through large-scale mobility data. ISPRS International Journal of Geo-Information 2019;8(3):134.

Rong, X.. word2vec parameter learning explained. arXiv preprint arXiv:14112738 2014;.

Olson, A.W., Calderon-Figueroa, F., Bidian, O., Silver, D., Sanner, S.. Reading the city through its neighbourhoods: Deep text embeddings of yelp reviews as a basis for determining similarity and change. Cities 2021;110:103045.

Levin, N., Kyba, C.C., Zhang, Q., de Miguel, A.S., Román, M.O., Li, X., et al. Remote sensing of night lights: A review and an outlook for the future. Remote Sensing of Environment 2020;237:111443.

Meng, L., Masuda, N.. Analysis of node2vec random walks on networks. Proceedings of the Royal Society A 2020;476(2243):20200447.

Peng, H., Ke, Q., Budak, C., Romero, D.M., Ahn, Y.Y.. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. Science Advances 2021;7(17):eabb9004.

Meng, W., et al. Analysis on the reliability of coastal city transportation from the perspective of resilience: A case study of yantai city, shandong province, china. In: IOP Conference Series: Earth and Environmental Science; vol. 580. IOP Publishing; 2020, p. 012019.

Lee, J.N., Mahmud, M., Morduch, J., Ravindran, S., Shonchoy, A.S.. Migration, externalities, and the diffusion of covid-19 in south asia. Journal of Public Economics 2021;193:104312.

Lao, X., Zhang, X., Shen, T., Skitmore, M.. Comparing china's city transportation and economic networks. Cities 2016;53:43–50.

Li, R., Dong, L., Zhang, J., Wang, X., Wang, W.X., Di, Z., et al. Simple spatial scaling rules behind complex cities. Nature communications 2017;8(1):1–7.

Dong, L., Huang, Z., Zhang, J., Liu, Y.. Understanding the mesoscopic scaling patterns within cities. Scientific reports 2020;10(1):1–11.

Zhang, Y., Chen, N., Du, W., Yao, S., Zheng, X.. A new geo-propagation model of event evolution chain based on public opinion and epidemic coupling. International Journal of Environmental Research and Public Health 2020c;17(24):9235.