International Journal of Geographical Information Science, 2015
Vol. 29, No. 6, 889–911, http://dx.doi.org/10.1080/13658816.2014.999244

Taylor & Francis
Taylor & Francis Group

Finding community structure in spatially constrained complex networks

Yu Chen, Jun Xu* and Minzheng Xu

State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing, China

(*Received 26 November 2013; final version received 12 December 2014*)

One feature discovered in the study of complex networks is community structure, in which vertices are gathered into several groups where more edges exist within groups than between groups. Many approaches have been developed for identifying communities; these approaches essentially segment networks based on topological structure or the attribute similarity of vertices, while few approaches consider the spatial character of the networks. Many complex networks are spatially constrained such that the vertices and edges are embedded in space. In geographical space, nearer objects are more related than distant objects. Thus, the relations among vertices are defined not only by the links connecting them but also by the distance between them. In this article, we propose a geo-distance-based method of detecting communities in spatially constrained networks to identify communities that are both highly topologically connected and spatially clustered. The algorithm is based on the fast modularity maximisation (CNM) algorithm. First, we modify the modularity to geo-modularity Q^{geo} by introducing an edge weight that is the inverse of the geographic distance to the power of n. Then, we propose the concept of a spatial clustering coefficient as a measure of clustering of the network to determine the power value n of the distance. The algorithm is tested with China air transport network and BrightKite social network data-sets. The segmentation of the China air transport network is similar to the seven economic regions of China. The segmentation of the BrightKite social network shows the regionality of social groups and identifies the dynamic social groups that reflect users' location changes. The algorithm is useful in exploring the interaction and clustering properties of geographical phenomena and providing timely location-based services for a group of people.

Keywords: complex network; spatial constraint; community; modularity

1. Introduction

Our world is full of many types of networks, such as the World Wide Web, social networks, and neural networks. The discovery of small-world (Watts and Strogatz 1998) and scale-free (Adamic *et al.* 2000) properties in many natural and artificial networks has inspired research of complex networks in many disciplines (Michalis *et al.* 1999, Strogatz 2001, Girvan and Newman 2002, Chen 2003, 2006, Barabási and Oltvai 2004). Many geographical phenomena are interrelated and form different types of networks. Many networks of geographical phenomena have the properties of a complex network, such as road networks, flight networks, and migration flow networks. Sen *et al.* (2003) found

*Corresponding author. Email: xujun@lreis.ac.cn
Present address of Yu Chen and Minzheng Xu is University of Chinese Academy of Sciences, Beijing 10049, China.

small-world properties of the Indian railway network; Bagler (2008) studied the airport network of India, which is connected by air links, and found that it is a small-world network with a truncated power-law degree distribution. By analysing the topological patterns of urban street networks and traffic flows, Jiang (2007, 2009) detected the power law distributions of both street length and connectivity degree and found that the most heavily used 20% of streets accommodate 80% of the traffic flow. People have begun to use the principles of complex networks to solve geographical problems. Crucitti *et al.* (2006) studied four centrality indices in urban street networks of different world cities and explored the different patterns of centralities in planned and self-organised cities. Wang *et al.* (2011) examined three centrality indices of cities in the air transport network of China, and the results indicated that they are highly correlated with socioeconomic indicators of cities. However, many studies are based on the existing methods of complex networks, which concentrate on the topological structure of a network, while seldom consider the spatial arrangement of a network.

A feature that has been discovered in the study of complex network is community structure (Girvan and Newman 2002). This structure refers to vertices that are gathered into several groups in which there is a higher density of edges within groups than among groups. Detecting the community structure of networks can help us discover hidden relations between vertices. The approach has been used in exploring data of e-mail communities, metabolic networks, web communities, literature co-citations, and project cooperation, among others (Flake *et al.* 2002, Ravasz *et al.* 2002, Palla *et al.* 2005, Tyler *et al.* 2005, Barber *et al.* 2011). Many approaches have been developed for finding communities; most of them segment networks based on vertex connectivity, namely, topological structure, and can be divided into two types. One type is based on graph partitioning, such as the Kernighan–Lin algorithm (Kernighan and Lin 1970), the algorithm based on minimum cut or normalised cut (Wu and Leahy 1993, Shi and Malik 2000), and the spectral partitioning method (Pothen *et al.* 1990, Seary and Richards 2003). The other type is based on hierarchical clustering, such as the Girvan–Newman algorithm (Girvan and Newman 2002), the CNM algorithm (Clauset *et al.* 2004, Newman 2006), and the Structural Clustering Algorithm for Networks algorithm (Xu *et al.* 2007). Additionally, there are some approaches to segment networks according to the attribute similarity of vertices, such as the Summarization by Grouping Nodes on Attributes and Pairwise Relationships algorithm (Tian *et al.* 2008). In many applications, both the topological structure and the vertex properties are important. Zhou *et al.* (2009) proposed the SA-Cluster algorithm, which is based on both structural and attribute similarities through a unified distance measure.

However, none of the preceding methods considers spatial characters of networks. Actually, many complex networks are spatially constrained in which the vertices and edges are embedded in space (Barthélemy 2011). The distance between vertices is related to the strength of the connection and thus has important effects on topological properties. Therefore, community partitioning by topology alone is not enough to disclose the relations among vertices. Based on the CNM algorithm, we proposed a geo-distance-based method of detecting communities in spatially constrained networks. The method can find geo-communities that are both highly topologically connected and spatially clustered. A geo-community is like a community which is a group of intensely connected nodes being loosely connected with others (Girvan and Newman 2002), but it is more compact in space. So the nodes in a geo-community not only share some network properties but also have stronger spatial correlations than those in other communities. The purpose is to disclose the regional characteristics of the network connections and to identify which is the most impact acting on an entity on the network, spatial interaction or

network interaction. The algorithm is tested with the air transport network of China and the BrightKite social network.

The remainder of this article is organised as follows: Section 2 introduces spatial issues of complex networks and related work; Section 3 describes the geo-distance-based method of detecting communities and the testing results after a brief review of the definition of modularity and the CNM algorithm; Section 4 presents two case studies and their results; finally, the limitations of the study are discussed in Section 5, and our conclusions are given in Section 6.

2. Spatially constrained complex network

2.1. Why space is concerned

Many complex networks, such as transportation networks, power grids, and water distribution networks, are constructed in space; therefore, they must be subject to spatial constraints. For example, vertices only connect to their neighbours in road or railway networks such that the links are spatially embedded; sensors, ad hoc, and wireless networks have short-range connections; power grids, communication networks, and neural networks depend on distance; human social interaction may also rely on distance. All of these examples are all spatially constrained. In spatially constrained networks, vertices are embedded in two- or three-dimensional space with metrics; edges indicate spatial distance information other than interactions among vertices, so space is relevant (Barthélemy 2011).

The networks of geographical phenomena are all spatially constrained. A vertex that represents a geographical object in a geographical network may have connections with distant objects; however, it must have some interactions with closer objects that are stronger than with farther objects due to the distance decay principle. As Tobler's first law states, near things are more related than distant things. Therefore, the relations among vertices are not only contained in the links of the network but also implicated by the distances among them. For example, in an air transport network, if each vertex represents a city or an airport, then the linkage between two vertices means they have flight connections. Additional flights between two cities only indicate more aviation or traffic interactions between them. Does this mean that these two cities have more interactions overall? Can the city groups segmented by topographic linkage alone truly represent the communities of cities? According to Tobler's law and the distance decay principle, the interaction between two cities declines as the distance between them increases. Close cities must have more interactions. However, distance is not the only factor. Cities far apart can increase their interactions through other connections, such as aviation or other activities. If we want to find the city communities that truly reflect the relations of cities, we should most likely consider both distance and other interactions.

Our social activities are also constrained by geography. People are more likely to have interactions with nearby friends. Specifically, when people move to another place, they may make new friends in that city and participate in more activities with them than with their previous friends; thus, the social circle changes. As the distance between two persons increases (decreases), the strength of their connection becomes weaker (stronger). So, the strength of a connection is elastic. The elastic social network was first proposed by the makers of an iPhone app called 'Color', which means the strength of a connection is based on the proximity of the nodes (Carr 2011, McCarthy 2011). In the elastic network, a person's location at a moment has a great effect on their following actions and

movements. Imagine that you are back in your hometown, and you have the opportunity to meet up with your childhood friends: who and where will you meet? At present the community you belong to is entirely different from the one at ordinary times. Thus, the geo-communities are dynamic due to the elastic strength between vertices. There is no doubt that we should consider the dynamic locations of vertices in the process of community detection.

To illustrate the basic idea of this study, we would like to explain the important role location plays in community detection of spatial networks. We assume that Figure 1a is a small social network in a two-dimensional space; a node represents one person in the space at a specific time and an edge represents a friendship. If only the network topology is considered, then three communities can be detected where there is a high density of edges within communities and only four links bridging the communities (Figure 1b). If only location is taken into account, then we can obtain three clusters (Figure 1c) using a spatial clustering algorithm (e.g., K-means algorithm). However, if both locations and network topology are considered, then we will most likely obtain different geo-communities, as Figure 1d shows. Thus, supposing that people are moving in the space, we may obtain completely different communities at different times. According to the

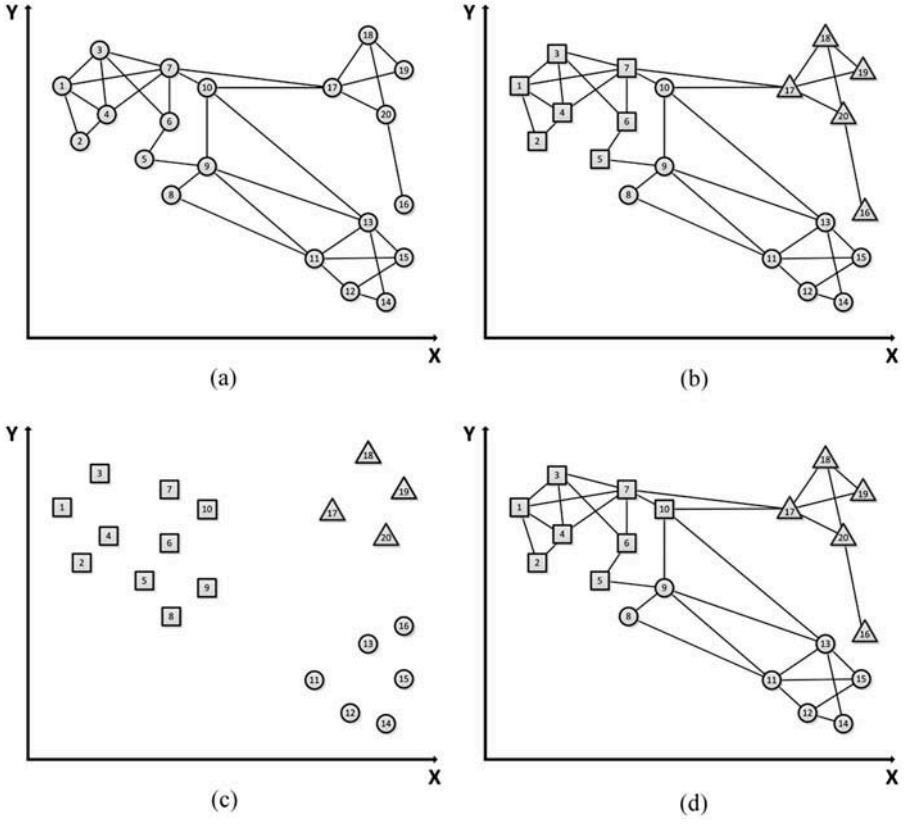

Figure 1. An example of a social network in space and its partitions with different methods: (a) original social network in space; (b) the community partition result with the CNM algorithm; (c) the community partition result with the spatial clustering algorithm; (d) the probable community partition result with the geo-distance-based community detection method.

preceding discussion, we consider that the existing methods for community detection are deficient in spatial network analysis. We need to reconstruct the community detection algorithm for spatially constrained networks.

2.2. Related work

In recent years, many researchers have realised that space plays a crucial role in spatial networks. They attempted to discuss the direct or indirect effect of space on network topology. Many studies on social networks have found that the probability of two persons being friends or having interactions depends on geographical distance. That is, the closer the distance, the greater possibility that two people have social interaction. As the distance increases, social interaction becomes less likely. Let $P(d)$ be the probability of having social interactions; its relation with distance d fits the distance decay function:

$$P(d) \sim d^{-\alpha} \tag{1}$$

However, the value of α changes in different data-sets. Liben-Nowell *et al.* (2005) explored more than one million nodes of the online network LiveJournal and obtained an α value of approximately 1. Lambiotte *et al.* (2008) analysed statistical properties of mobile phone communication in Belgium and obtained an α value of 2. In the work of Onnela *et al.* (2011) on mobile phone calls and text messaging data in a European country, α is approximately 1.5. They studied the topological positions and geographical positions of social groups and found that small social groups are geographically very tight but suddenly spread out when the group size exceeds approximately 30 members. Scellato *et al.* (2010) analysed four large online social networks with new geo-social metrics, and the results show that geographic distance influences some properties of a social structure. Liu *et al.* (2014) analysed the toponym co-occurrences of Chinese provinces on web pages. They found that the frequency of co-occurrences exhibits a power law distance decay effect with an exponent of 0.2 and that spatially close provinces generally have similar co-occurrence patterns.

Regarding the community detection of a spatially constrained network, some researchers have addressed the issues of geographic constraint. Current studies generally opt for three approaches:

First, most studies adopt the classic community detection algorithm without considering spatial factors. However, because nodes in close proximity are more likely to be connected than those far apart, the community structures of some networks already exhibit regionalised properties (Guimera *et al.* 2005, Barber *et al.* 2011).

The second approach is taking spatial constraint into account with graph partitioning. Guo (2008) proposed the average linkage (ALK) method, a regionalisation method based on spatially constrained hierarchical clustering. Using the ALK method and adopting modularity as the connection strength of each pair of nodes, he constructed a spatially continuous tree and realised hierarchical regionalisation and visualisation of flow data (Guo 2009). Using the same regionalisation method, Liu *et al.* (2014) showed the spatial organisation of China in different levels based on co-occurrences of province names on web pages.

The third perspective is proposed by Expert *et al.* (2011), who eliminates the effect of space in graph partitioning to reveal the hidden structural similarities between the nodes. In their opinion, communities are strongly determined by geographic factors in spatial networks; thus, community detection without considering spatial anomalies will fail to

discover the other factors. They proposed a modified function of modularity based on connective probability at different distances to discover community patterns that are outside of spatial influence. They tested their new method with mobile phone data and found it works well to reveal more clearly hidden structural similarities between the nodes. Gao *et al.* (2013) also proposed an alternative modularity function incorporating a gravity model to remove spatial influence in discovering the communities of mobile phone users.

Unlike Expert's experiment, we believe that vertices nearby are more likely to have relations, so we choose to strengthen the effect of space in exploring communities instead of factoring out it. Guo's regionalisation method is based on ALK stresses space, and it results in continuous regions. Many networks involve discrete phenomena with no spatial continuity. Sometimes, if two vertices have strong interactions despite great distances, they most likely belong to one community. In this article, we synthesise the effect of distance and linkage in network community detection by gathering vertices that are both tightly linked and spatially clustered into one group.

3. Methodology

Among the algorithms of community detection, the fast modularity maximisation algorithm (CNM algorithm) proposed by Clauset *et al.* (2004) is widely used. By defining the weights of the edges in the network as the function of the distance between two directly connected vertices, we modify the fast modularity maximisation algorithm. The weight is defined such that the closer the distance, the greater the weight.

3.1. Modularity and CNM algorithm

The CNM algorithm (Clauset *et al.* 2004) is a community detection algorithm based on hierarchical clustering. The basic concept of the CNM algorithm is modularity (Newman and Girvan 2004). It is a measure to evaluate the result of the network partitioning, which computes the difference between the number of links within communities and the expected number. A good partition of a network should result in a significantly greater number of edges within communities than expected. The mathematical expression of modularity is:

$$Q = \frac{1}{2m} \sum_{C \in P} \sum_{v,w \in C} \left(A_{vw} - \frac{k_v k_w}{2m} \right) \tag{2}$$

where C is a community, P is the community set of the network, and v, w are nodes in the community C. A_{vw} is the element of the adjacency matrix of the network. If there is an edge between v and w, then $A_{vw} = 1$; otherwise, $A_{vw} = 0$. m is the number of edges in the network, and k_v is the degree of node v. The higher the value of Q, the better the community structure is. To simplify the description, two variables are introduced:

$$e_{ij} = \frac{1}{2m} \sum_{v \in i, w \in j} A_{vw} \tag{3}$$

$$a_i = \sum\nolimits_j e_{ij} \tag{4}$$

where e_{ij} is the fraction of edges that join vertices in community i to vertices in community j, and a_i is the fraction of edges that are attached to vertices in community i. Thus, the function of Q can be transformed into:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{5}$$

The CNM algorithm initially regards every vertex as a community and then merges them step-by-step. When each pair of communities is merged, an increase of modularity ΔQ results from the amalgamation. At each time, it chooses the pair of communities that generates the maximum ΔQ to join a new community. When the maximum ΔQ becomes negative, the process is stopped, and the community structure of the network is revealed.

3.2. *Geo-modularity and modified algorithm*

In this article, we modify the CNM algorithm to detect a geo-community. Because the geographical distance will affect the strength of the connection among vertices, we propose a geo-modularity, which adds a weight of distance to edges when calculating modularity. According to the effect of distance decay, the weight of edges is defined as

$$\text{weight}_{vw} = 1/d_{vw}^n \tag{6}$$

where d is the normalised distance between vertex v and vertex w. Then, according to Equation (3), we can obtain the weighted matrix of $\mathbf{e}^{\text{geo}}$:

$$\mathbf{e}_{ij}^{\text{geo}} = \frac{\sum\limits_{v\in i, w\in j} \text{weight}_{vw}}{2W} = \sum_{v\in i, w\in j} 1/d_{vw}^n \Big/ \sum_{v,w} 1/d_{vw}^n \tag{7}$$

where W is the weighted sum of edges in the network. Let a_i^{geo} be the sum of elements in each line of the matrix,

$$a_i^{\text{geo}} = \sum\nolimits_j \mathbf{e}_{ij}^{\text{geo}} \tag{8}$$

the geo-modularity can be defined as:

$$Q^{\text{geo}} = \sum_i \left[\mathbf{e}_{ii}^{\text{geo}} - (a_i^{\text{geo}})^2\right] = \sum_i \left[\frac{\sum\limits_{v,w\in i} \text{weight}_{vw}}{2W} - \left(\frac{\sum\limits_{v\in i} \text{weight}_{vw}}{2W}\right)^2\right] \tag{9}$$

As the CNM algorithm, the geo-distance-based community detection algorithm starts with each vertex as a single community. It calculates the increase of geo-modularity ΔQ^{geo} of joining each pair of communities, and the pair generating the maximum ΔQ^{geo} is merged. In this way, we can find geo-communities considering both geographic distance and links.

Choosing a different value of the nth power in Equation (6), the geo-distance-based community detection algorithm can obtain different community structures. With the increase of the n value, the geo-modularity of the network always increases. There is no way to choose n by the topology of the community structure. Thus, we propose the concept of a spatial clustering coefficient to measure the degree to which vertices tend to cluster together spatially; then, we choose the value of n by the degree of clustering in space. First, we obtain the diameter of the network O, which is the farthest distance between the nodes in the network, and let d_{vw} be the distance between vertex v and vertex w. Thus, we define δ_{vw} as the spatial closeness between v and w:

$$\delta_{vw} = 1 - d_{vw}/O \tag{10}$$

Then, we obtain the distance matrix $\mathbf{D}$. $\mathbf{D}_{ij}$ is the element this matrix, which is the ratio of spatial closeness between community i and j to the sum of spatial closeness of all vertices in the network.

$$\mathbf{D}_{ij} = \sum_{v \in i} \sum_{w \in j} \delta_{vw} \Big/ \sum_{v,w \in U} \delta_{vw} \tag{11}$$

where U is the whole vertex set.

Let c_i be the sum of the elements in each line of $\mathbf{D}_{ij}$, which represents the ratio of spatial closeness between community i and any other communities to the sum of spatial closeness of all vertices in the network.

$$c_i = \sum_j \mathbf{D}_{ij} \tag{12}$$

Finally, spatial clustering coefficient can be calculated as follows:

$$S = \sum_i [D_{ii} - c_i^2] \tag{13}$$

As the formula of modularity, S measures the difference between the closeness of vertex within communities and the expected closeness. The higher S is, the more significantly the vertices cluster in space.

3.3. *Test result*

The small network in Figure 1a is used to test the geo-distance-based community detection algorithm. Figure 2 shows the modularity of the CNM algorithm, which ignores the geographic distance and geo-distance-based algorithm with a different power of n at different steps of joins. All methods stop at the 15th–17th steps. The higher the power value is, the higher the peak value of the geo-modularity is and the faster the algorithm reaches the peak value and stops. Table 1 compares the modularity, geo-modularity, and spatial clustering coefficient of the community divisions with different methods. After taking distance into consideration, the network is divided into more communities, and the modularity becomes smaller. Table 1 shows that when the value of the nth power increases, Q^{geo} also increases, while Q decreases. However, the spatial clustering

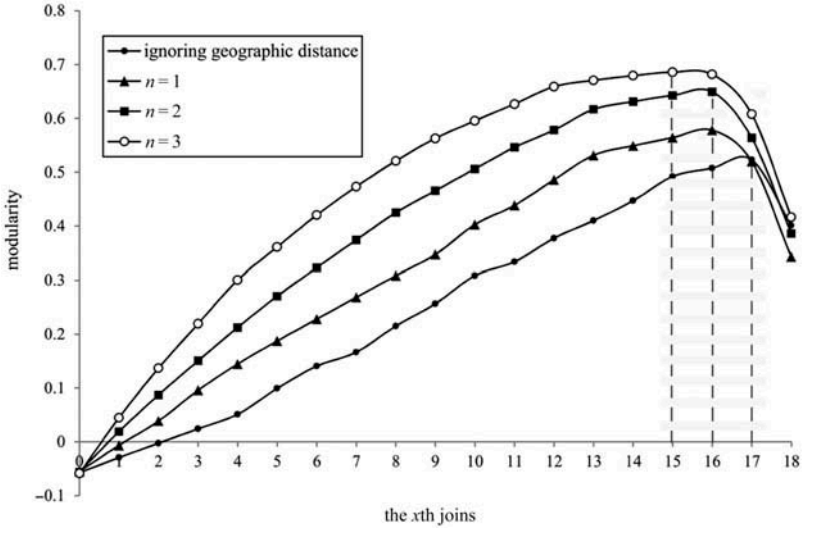

Figure 2. The change of modularity while applying the different algorithms.

Table 1. The measurements of community divisions with different algorithms.

Algorithm		The number of communities	Q	Q^{geo}	S	The steps of recursive
CNM algorithm		3	0.52	/	0.067	17
Modified algorithm	$1/d$	4	0.44	0.58	0.069	16
	$1/d^2$	4	0.44	0.65	0.069	16
	$1/d^3$	5	0.43	0.69	0.059	15

coefficient S first increases and then decreases. When n is equal to 1 or 2, S has the highest value, which means the resulting community divisions are mostly clustered in space. Considering the three measurements, the best choice of n is 2.

Figure 3 shows the community divisions with different value of the nth power. When n equals 1 or 2, four communities are found (Figure 3a). When n is 3, five communities are found (Figure 3b). Compared with the result of the CNM algorithm in Figure 1b, we can easily realise that the geo-distance-based algorithm can reveal more communities, and the vertices in the same community are much closer to each other.

4. Case studies

In this section, two experiments with the geo-distance-based community detection algorithm are conducted using the China air transport network and BrightKite social network, and the results are analysed.

4.1. China air transport network

The study area includes all cities with airports in mainland China (excluding Hong Kong, Macao and Taiwan). Each pair of airports with connecting flights is linked to form a

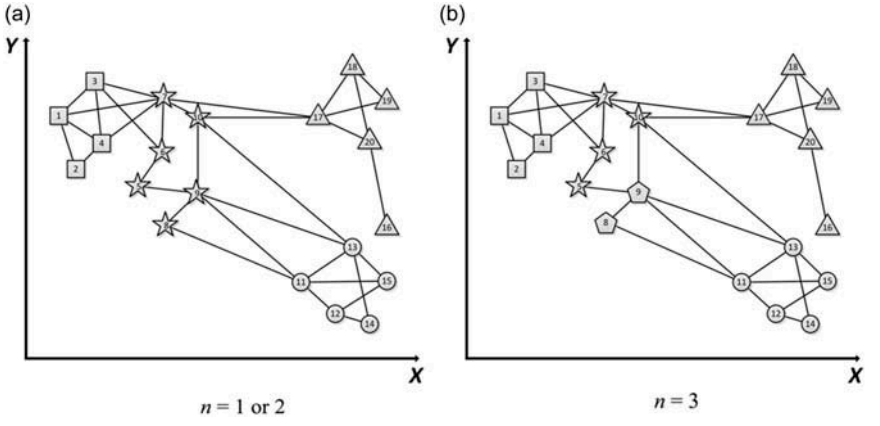

Figure 3. The community divisions by the geo-distance-based algorithm with different power values of *n*.

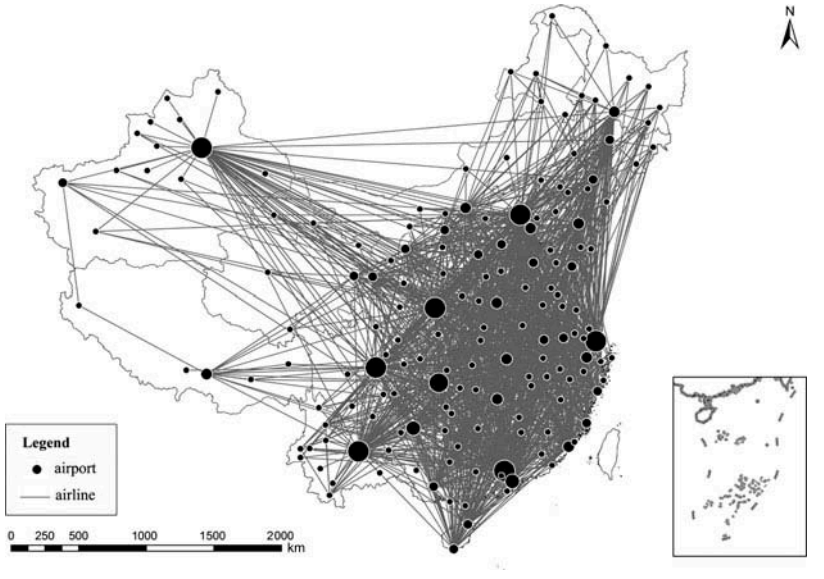

Figure 4. China air transport network.

network. For cities that have more than one airport, the airports are merged as one node, and the flights of these airports are all set to this node. Basically, a node in the network represents a city in China, and the links between nodes represent direct flights between these cities. There are 167 nodes and 1386 links in the network (Figure 4).

4.1.1. Community detection

Many studies have focused on the properties of air transportation networks. Our study focuses on the community structure of the China air transportation network. Air transportation between two cities reflects a type of economic interaction. More flights mean

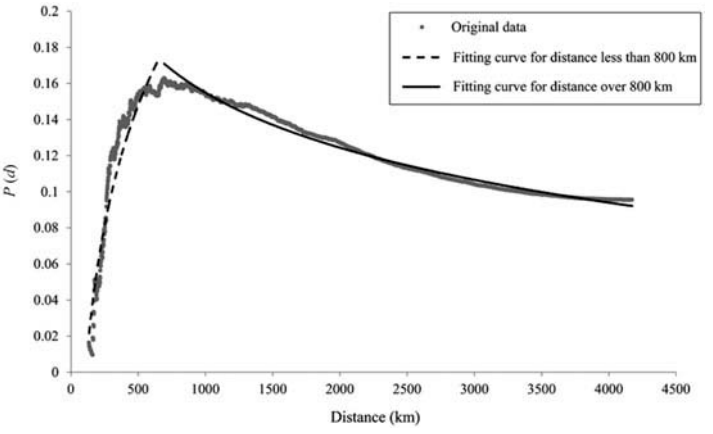

Figure 5. The relationship between flight probability and geographic distance of the China air transport network.

that there is a stronger interaction. However, air transportation is useful for travelling long distances. Close cities may not have air transportation between them, even if they have strong economic interaction. As we can see in Figure 5, the relationship between flight probability and geographic distance of the China air transport network has two components. When the distance is over 800 km, the distribution of flight probability $P(d)$ fits the distance decay function $P(d) \sim d^{-0.35}$. However, when the distance is less than 800 km, the flight probability $P(d)$ decreases when the distance decreases. Under such circumstances, when dividing the network without considering the distance, the regional properties will be neglected.

The CNM algorithm and geo-distance-based algorithm are used for detecting the communities. Here, we consider the number of flights in one week N_{ij} as the weight of the link. In the geo-distance-based algorithm, we computed the weight as:

$$\text{weight}_{ij} = \frac{N_{ij}}{d_{ij}^{n}} \tag{14}$$

Table 2 shows the measurements of the final divisions with different value of the nth power from 1 to 3. When n is approximately 2, S is maximum. Therefore, we take $n = 2$ as the power of the distance.

Table 2. The measurements of community divisions of the China air transport network.

Algorithm		The number of communities	Q	Q^{geo}	S
CNM algorithm		4	0.148	/	0.141
Modified algorithm	$n = 1.0$	6	0.100	0.256	0.277
	…	…	…	…	
	$n = 2.0$	8	0.081	0.442	0.290
	…	…	…	…	…
	$n = 3.0$	10	0.084	0.605	0.263

4.1.2. Results and analysis

To compare the results of different algorithms and parameters, we choose the results for n = 1, 2, 3 and the CNM algorithm (Figure 6). Four communities are found by the CNM algorithm (Figure 6a). The four communities intersect each other in space. There is no obvious regional characteristic of the communities, except in the northwest part of China. The result for n = 1 shows slight regionality, but some communities disintegrate in space. For example, community 5 includes five cities; two are near the east coast, while the other three are in Central China (Figure 6b). Because there are more flights among these cities to produce strong connections, they are incorporated into one community despite their distances. When n = 2, the result shows good regionality (Figure 6c). Due to the increased power of the distance, the three cities in community 5 of the previous result are now members of the community in southeastern China. However, there is one city, i.e., Foshan City, near the southeastern coast of China that remains a member of the distant group. It is linked only to Beijing, so it is a member of the group to which Beijing belongs. The same is true for the city of Guangyuan. Comparing Figure 6a and c, we can identify three kinds of nodes. The first kind of nodes is more affected by spatial relation than by network connection. For example, the four cities in the middle of China, Changsha, Zhangjiajie, Huaihua and Yongzhou, have both long-distance and short-distance flights to other cities. They

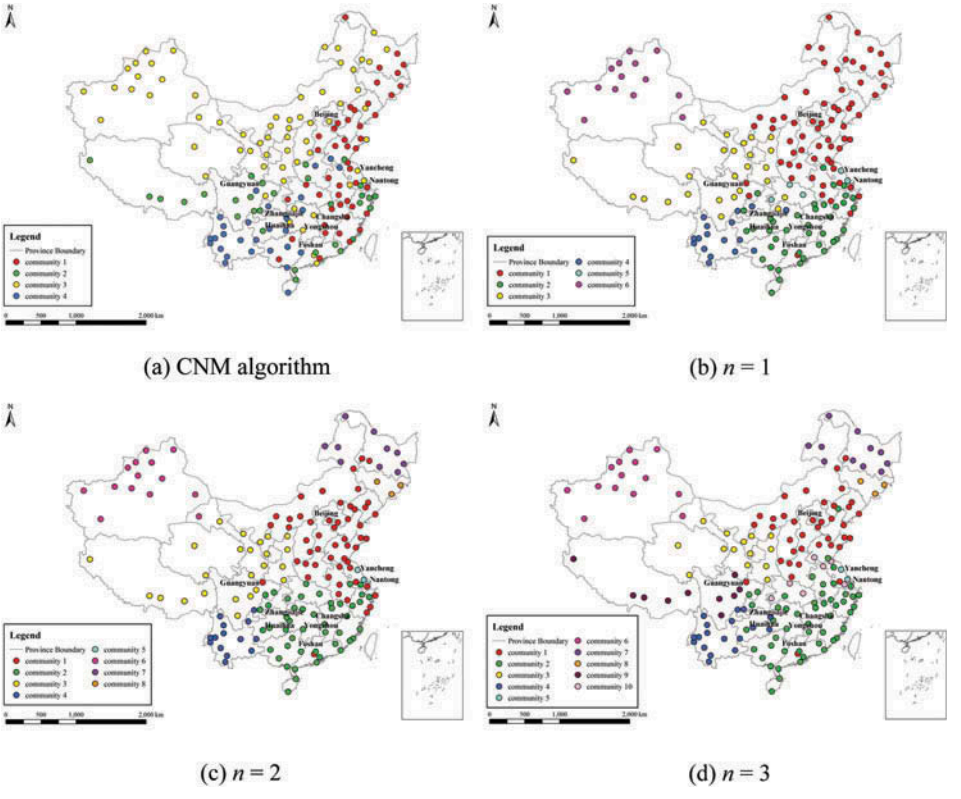

Figure 6. Communities of the China air transport network found by different algorithms. (a) CNM algorithm; (b) n = 1; (c) n = 2; (d) n = 3.

belong to a community occupying most part of North China in Figure 6a, but when distance factor is considered, they are merged to a community with surrounding cities. The second kind of nodes is more affected by network connection than by spatial relation, such as Foshan and Guanyuan mentioned earlier. Even when distance factor is considered, they still belong to a community faraway from them. The third kind of nodes has both strong spatial relation and network connection. For example, there are a lot of flights between the two cities near the east coast, Yancheng and Nantong, and they are geographically close to each other, so these two cities form a separate community. With the increased distance constraint, the result of $n = 3$ becomes more divided. Some communities from $n = 2$ break into two components (Figure 6d). The three cities of community 5 in Figure 6a once again break out from the spatially adjacent communities to form a separate community. In general, the result of the geo-distance-based algorithm shows more regionality and retains the topological properties of the network. The geo-distance-based algorithm reveals more communities than the CNM algorithm. The small communities are more spatially congested.

Figure 7 shows the seven main economic regions in China. Compared with Figure 6c, we can see the community division with the geo-distance-based algorithm matches the economic regional division. Communities 3, 4, and 6 in Figure 6c are equivalent to the northwest and southwest regions in Figure 7, while community 2 in Figure 6c represents the east, middle, and southeast regions in Figure 7. The northeast economic region is split into two communities in Figure 6c. The economic regions are divided by provinces, while the communities are composed of cities; thus, the boundaries of the two divisions are slightly different. The result shows that the community division considering both topology and distance can disclose regional characteristics of Chinese cities and reflect some economic phenomena.

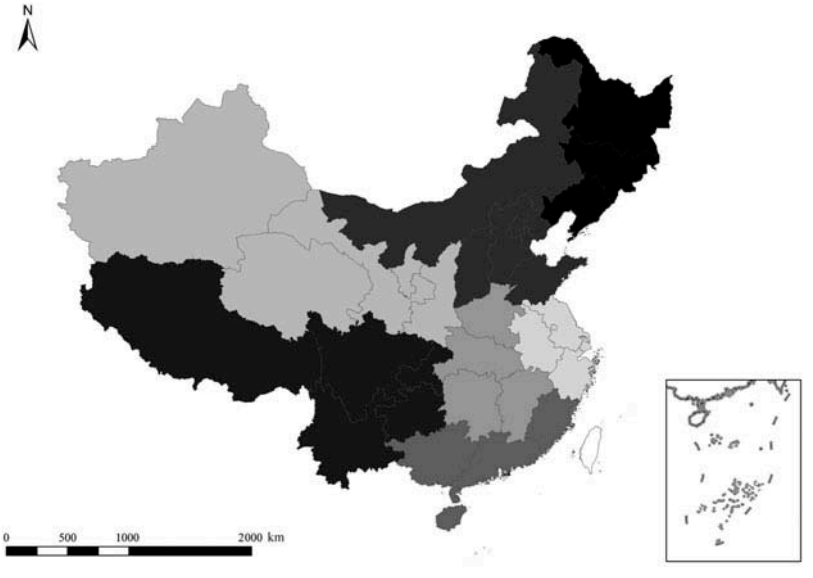

Figure 7. The seven main economic regions and the corresponding provinces of China.

4.2. *BrightKite social network*

Many online social networks provide a check-in service through which users can share their locations. If we can identify social communities that strongly interact and are spatially closed through this information, we can provide location-based services to a group of people, rather than only a single person. In this study, we use the BrightKite data-set. BrightKite was once a location-based social networking service provider that allowed users to share their locations; it was closed in 2012. The data we use are from the Stanford Network Analysis Platform (http://snap.stanford.edu/data/loc-brightkite.html). The data-set is very large, so we choose only the users in the US mainland. We collected the check-in data from April 2008 to October 2010. There are a total of 28,074 users and 100,651relationships in the subset.

4.2.1. *Community division of the BrightKite social network*

First, the users' permanent locations are estimated using the method introduced by Scellato *et al.* (2011) and Cho *et al.* (2011). The world is divided into 25-by-25-km cells, and the permanent location of a person is defined as the average position of check-ins in the cell with the most check-ins. According to Cho *et al.* (2011), manual inspection showed that the accuracy of this method is as high as 85%. Taking users as nodes and friendships as edges, a network is constructed (Figure 8).

Six hundred and twenty-eight communities are found by the CNM algorithm. For visualisation, we only present the three largest communities in Figure 9a. Unfortunately, no geographic effect can be found in this graphic. To compare our geo-distance-based algorithm with CNM, we consider the permanent location of each node as the source for computing the geographic distance between nodes. There are more than 1200 communities found. Figure 9b shows the distribution of geo-communities when the value of power n is 2 (for the effect of visualisation, only the 10 largest communities are shown in Figure 9b). The communities are more congested in space,

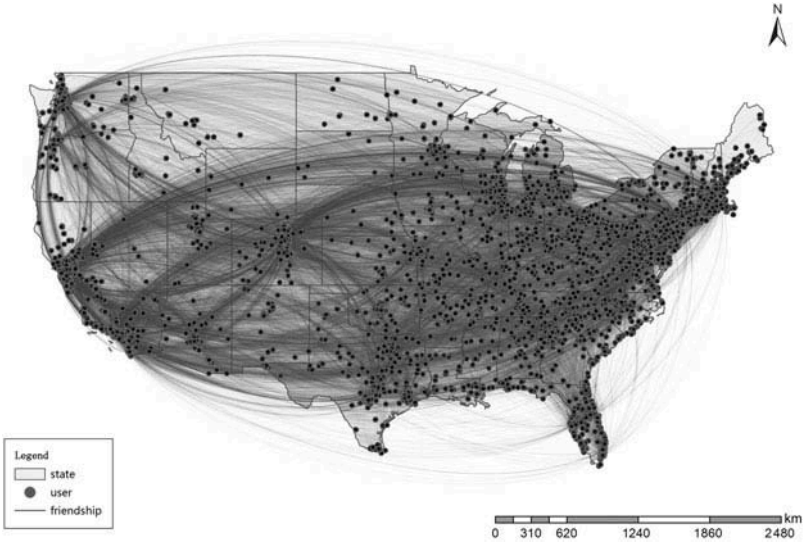

Figure 8. BrightKite social network in the US mainland.

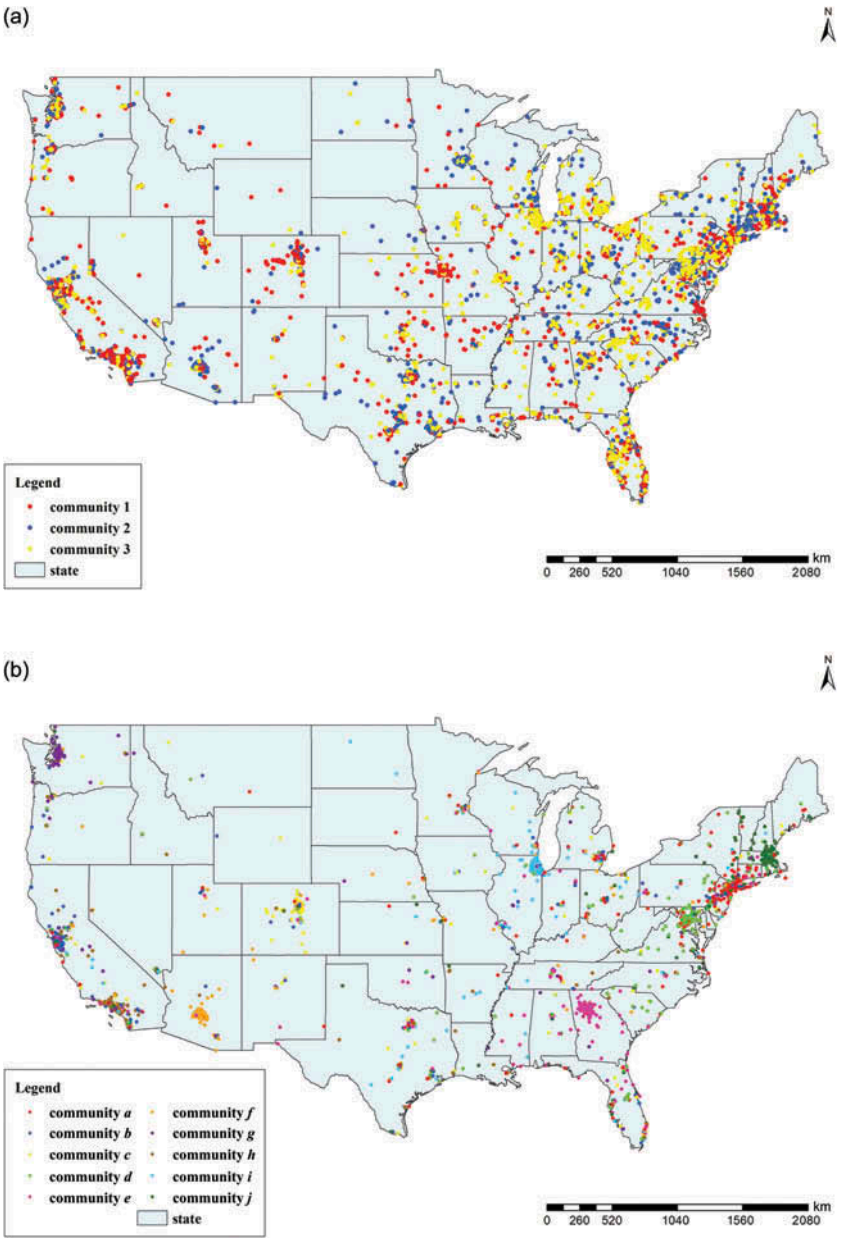

Figure 9. Communities of the BrightKite social network in the US mainland by different algorithms. (a) CNM algorithm; (b) geo-distance-based algorithm.

although each community still has some nodes scattered throughout space, which means the relations between them are too strong to be broken by distance. Considering that location produces twice the number of communities as the CNM algorithm, the geo-communities are much smaller than the structure communities.

4.2.2. Visualisation

To analyse the closeness between communities intuitively, we apply a spectral graph method to visualise the results. The spectral method is a component of graph theory. The eigenvalues of a network are closely related to important topological features of the network. If using eigenvectors as a coordinate system, the features can be visualised as a spectral graph (Seary and Richards 2003, Barber *et al.* 2011). Here, we unite the nodes in a community into one node and simplify the whole network into a network of communities. Then, the relations of the communities can be visualised in a Laplacian spectral graph. The Laplacian matrix of a network is:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{15}$$

Here, **A** is the adjacency matrix of the network, and **D** is the diagonal matrix in which the elements on the diagonal are the degrees of the nodes. Therefore, the elements in the Laplacian matrix are:

$$l_{ij} = \begin{cases} \mathrm{k_i} & \text{if i = j and } \mathrm{k_i} \neq 0 \\ -1 & \text{if i} \neq \text{j and i adjacent to j} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Because the minimum eigenvalue of the Laplacian matrix is 0, the eigenvectors corresponding to the second and third smallest eigenvalues are used as the x and y coordinates of the graph.

Figure 10 shows the Laplacian spectral graph of the 10 largest communities of the BrightKite social network. The coordinates of the nodes represent the locations of the communities in eigenvector space, and the close distance between communities means there are more interactions. In Figure 10, we can see that community a and community g are very close to each other. Figure 11 shows their locations. Community a is basically distributed around New York City, and community g is generally distributed around Seattle. Therefore, people in these two cities have more online interactions. When only topology is considered, more than 57% of the people in these two communities belong to one community. When distance is also taken into account, they are separated into two communities. Because people in the same city could have more offline interactions, the relations in one city could be stronger than in multiple cities. This result is reasonable.

4.2.3. Dynamic community detection

Because most social networks are actually dynamic, when a person's location changes, his or her social group also changes. To show the geo-distance-based algorithm's application in an elastic social network, we explore the dynamic communities according to the check-in activities at different times.

This study chooses a very active user X in Denver, Colorado, as the case study. This user checks in four times a day on average. User X and his 97 online friends form a small network with 98 nodes and 638 links. Most of his friends live in Denver, and some of them live in San Francisco and elsewhere. The geo-distance-based algorithm is used only on this small network to show the dynamic changes of his social group. First, communities are detected based on their permanent locations. The result is shown in Figure 12. Seven communities are found. The one to which user X belongs is denoted in

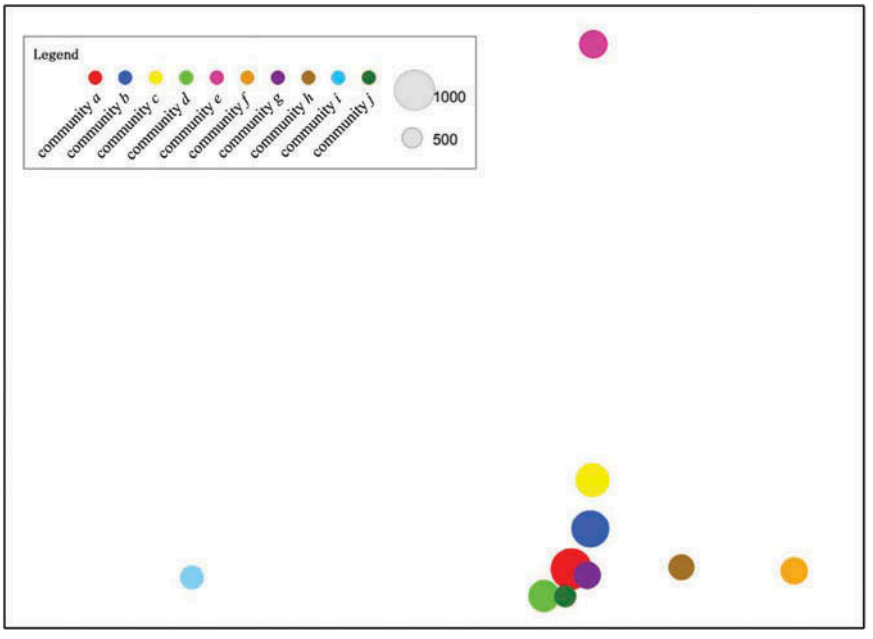

Figure 10. Laplacian spectral graph of the 10 largest communities of the BrightKite social network. The node size represents the number of people of the community.

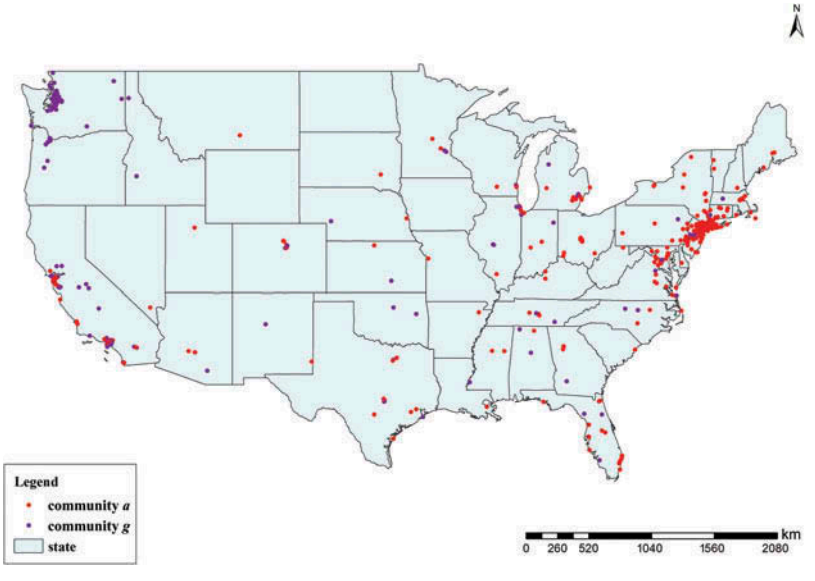

Figure 11. The spatial distribution of community *a* and community *g*.

red. The areas of Denver and San Francisco are enlarged on the left. Because *X* is the central node of the network, we can see in Figure 12 that the communities in Denver are spatially clustered, while those in San Francisco show no spatial regularity due to the long distance to *X*.

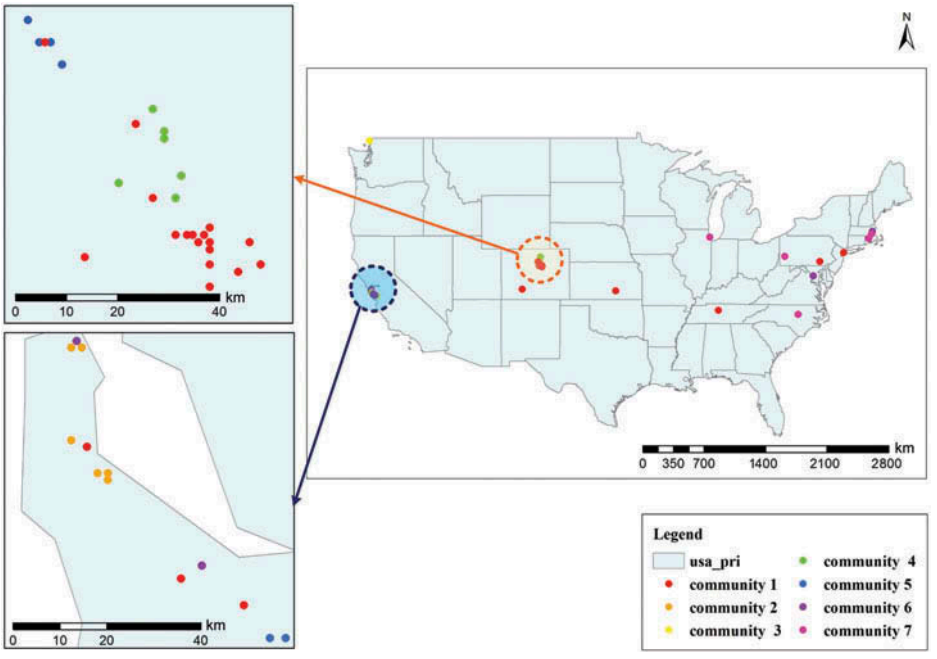

Figure 12. User *X*'s social network at his permanent location. The one to which user X belongs is denoted in red. The areas of Denver and San Francisco are enlarged on the left.

It is known from the check-in information that *X* was in San Francisco on 25 May 2009. We use the geo-distance-based algorithm to find the temporal social communities of *X* on that day (Figure 13). Now we can see that people in San Francisco obviously form a community due to his arrival, while the communities in Denver show no spatial regularity. We expect that there would be a large welcome party for *X* in San Francisco if BrightKite could inform everyone about the location change.

5. Discussion

Although the validity of the geo-distance-based algorithm has been tested in case studies, there are still some aspects of the algorithm that need to be discussed.

5.1. The adaptability of the algorithm

Generally, the communities detected are characterised as regional while still keeping the topological relations of the complex network after taking space into account. The geo-distance-based algorithm produces more communities than previous algorithms. These smaller communities synthetically reflect the spatial closeness and topological connection of the nodes. However, we must consider carefully whether distance should be involved in community detection. For different purposes and applications, the answer would be different. If the purpose is to study the topology of a spatial network, then there is no need to consider the geographic distance. If the phenomenon under research is related to spatial distribution, then the geographic distance must be taken into consideration. For instance, consider the cases in this study. If we only pay attention to the air transport

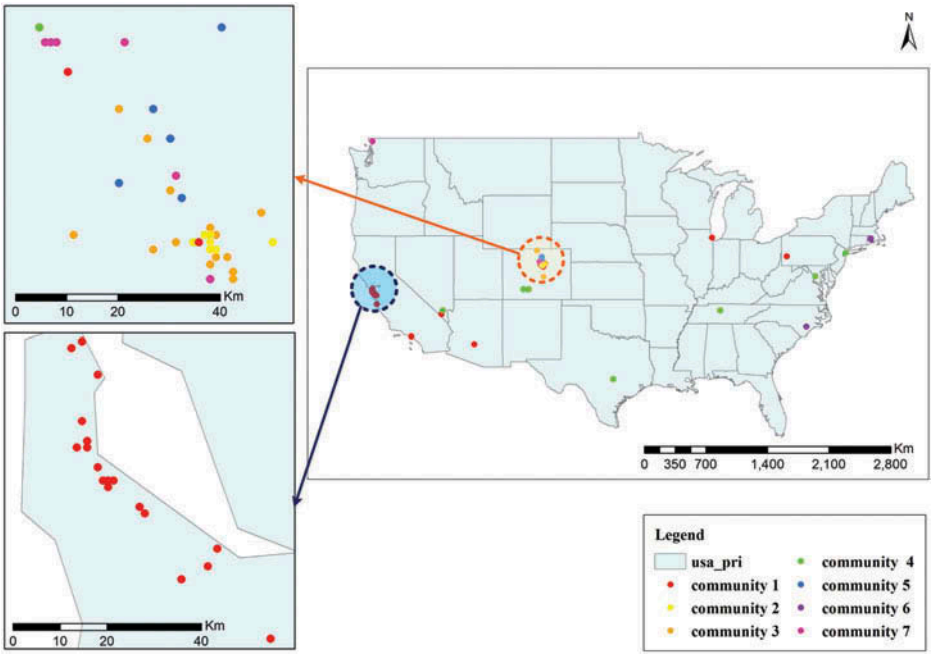

Figure 13. User *X*'s temporary social network on 25 May 2009. The one to which user *X* belongs is denoted in red. The areas of Denver and San Francisco are enlarged on the left.

convenience between cities or online social interaction of people, we do not need to take geographic distance into consideration. However, when inter-city interaction that depends on spatial distance to a certain extent is studied or real-world activities of online social networks are involved, such as assemblies or conventions, the geo-distance-based algorithm is useful in identifying the tightly connected and spatially clustered groups of nodes. The geo-distance-based algorithm is actually a community detection algorithm based on a weighted network, except that the weigh of links on a spatial network is the proximity between nodes. So, a node is incorporated into a community only if it has connection with other nodes in the community; otherwise, it will not be incorporated even if it is very close to the community, such as the city of Foshan in Figure 6. On the contrary, if nodes with intensive connection are far apart in distance, if the distance is long enough to weaken the connection, they will not be incorporated into one community.

Even when distance is considered, there are different ways to use it in solving different problems. When the purpose is to eliminate the effect of distance to explore non-spatial factors that effect the distribution of the communities, Expert's *et al.* (2011) method should be used. However, if spatial factor is important in effecting the interactions between nodes, and clusters with strong interactions are to be found, then the geo-distance-based community detection algorithm should be used. According to Equation (1), the intensity of connection is affected by the distance, so the result based on CNM algorithm might show a little regionality, such as the cases in Expert *et al.* (2011) and Gao *et al.* (2013). But when we strengthen the distance more, the regionality will become more evident, such as the case of BrightKite dataset (Figure 9). In the case of BrightKite social network, people are moving in the space. If the distance is strengthened, the dynamic communities reflecting human mobility can be detected, otherwise only one fixed

community division can be acquired. As to the air transport networks, the flight probability does not always follow the distance decay function (Figure 5); however, the interaction between cities still follows the first law of geography. In that case, taking distance into account will find more regional communities.

Both space relation and network relation have impacts on the interactions between nodes in a network. If we assume spatial relation as an interaction and add it to the network, then the spatially constrained networks could be treated as heterogeneous networks. A heterogeneous network, or multidimensional network, is a network having different kinds of interactions, where each interaction between two entities is a dimension (Yang *et al.* 2012, Berlingerio *et al.* 2013). In a spatially constrained network, each node is under the action of at least two interactions: one is the network connection, and the other is the spatial interaction. However, the effects that the two kinds of interactions act on a node are different. Which one is dominant on a node depends on the character of the node. As in the case of China air transport network, three kinds of cities can be identified. The geo-distance-based algorithm can discover the dominant interaction of the nodes in a network.

5.2. *The power of the distance weight*

Obtaining the best nth power of distance weight is the key problem in the geo-distance-based algorithm. In this article, an experimental method is used for determining the value of n. First, we propose spatial clustering coefficient S to estimate the clustering degree of the divided communities. Then, we consider several values of n for testing, and value that produces the best spatial clustering coefficient S is chosen as the power of the weight; however, this method is tedious. Is there a relationship between links and geographic distance that can help ascertain the power of n? Is the best value of n of a network related to any characteristic of itself, such as the power of distance decay or other properties? We wanted to determine the value of n by fitting the parameter within data-sets. Because only two networks are analysed in this article, it is impossible to find the correlation. This could be the target of further researches. In addition, the chosen n can only be theoretical at best. After all, is there a best value of n for real-world applications? We probably need different n values to find clustered communities at different scales. For example, in a social network, we should most likely use different n values when we attempt to obtain the national scale of communities and the city scale of communities.

6. Conclusion

This article seeks a way to combine the physical distance and the topological structure of a network in exploring community structures of a spatially constrained network because the strength of an inter-node connection is concerned with geographic distance and topological linkage in such a network according to the first law of geography (i.e., near things are more related than distant things). Thus, a geo-distance-based method of detecting communities in spatially constrained networks is proposed. The algorithm is based on the CNM algorithm. It considers the geographic distance as the weight of links in calculating the geo-modularity Q^{geo} of the network. The weight of a link is defined as $1/d^n$ (d is the geographic distance between two vertices). Then, the shorter the distance between vertices becomes, the greater the weight of an edge is. We also define the spatial clustering coefficient as a measure of clustering of the network to determine the power n of the distance. The algorithm is tested using the air transport network of China and BrightKite

social network data-sets. The results show that this method can combine both the geographic distances and links among vertices well. Specifically, the partition result of the China air transport network is similar to the seven economic regions of China. The result of the BrightKite social network shows the regional properties of social groups. The algorithm can find dynamic social groups caused by the location changes of users as well.

Our method of discovering geographically constrained communities is significant for analysing the network's topological structure and space distribution and discovering which one of the spatial and network interactions is dominant on an entity. It would be helpful to study the clustering property, the information dissemination of inter-community or inner community, and the behaviour prediction of spatially constrained networks. Particularly, it can dynamically divide the communities according to the real-time location change of nodes. It is very useful in location-based services for providing timely recommendations for a group of people. The adaptability of the algorithm and its remaining problems are also discussed for future study.

Acknowledgements

The authors thank the anonymous reviewers whose remarks greatly improved the presentation of this work. This research was supported by the National Natural Science Foundation of China [Grants 41171296 and 41371380].

References

Adamic, L.A., *et al.*, 2000. Power-law distribution of the World Wide Web. *Science*, 287, 2115. doi:10.1126/science.287.5461.2115a

Bagler, G., 2008. Analysis of the airport network of India as a complex weighted network. *Physica A: Statistical Mechanics and Its Applications*, 387 (12), 2972–2980. doi:10.1016/j.physa.2008.01.077

Barabási, A.-L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5 (2), 101–113. doi:10.1038/nrg1272

Barber, M.J., Fischer, M.M., and Scherngell, T., 2011. The community structure of research and development cooperation in Europe: evidence from a social network perspective. *Geographical Analysis*, 43 (4), 415–432. doi:10.1111/j.1538-4632.2011.00830.x

Barthélemy, M., 2011. Spatial networks. *Physics Reports*, 499 (1–3), 1–101. doi:10.1016/j.physrep.2010.11.002

Berlingerio, M., *et al.*, 2013. Multidimensional networks: foundations of structural analysis. *World Wide Web*, 16 (5–6), 567–593. doi:10.1007/s11280-012-0190-4

Carr, A., 2011. *Highlight CEO Paul Davison launches elastic network of a different color* [online]. Available from: http://www.fastcompany.com/1825725/highlight-ceo-paul-davison-launches-elastic-network-different-color [Accessed 9 December 2014].

Chen, C., 2003. *Mapping scientific frontiers: the quest for knowledge visualization*. London: Springer-Verlag.

Chen, C., 2006. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57 (3), 359–377. doi:10.1002/asi.20317

Cho, E., Myers, S.A., and Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. *In*: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 21–24 August, San Diego, CA. New York: ACM. 1082–1090. doi:10.1145/2020408.2020579

Clauset, A., Newman, M.E.J., and Moore, C., 2004. Finding community structure in very large networks. *Physical Review E*, 70 (6), 066111. doi:10.1103/PhysRevE.70.066111

Crucitti, P., Latora, V., and Porta, S., 2006. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73 (3), 036125. doi:10.1103/PhysRevE.73.036125

Expert, P., *et al.*, 2011. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108 (19), 7663–7668. doi:10.1073/pnas.1018962108

Flake, G.W., *et al.*, 2002. Self-organization and identification of web communities. *Computer*, 35 (3), 66–70. doi:10.1109/2.989932

Gao, S., *et al.*, 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17 (3), 463–481. doi:10.1111/tgis.12042

Girvan, M. and Newman, M.E.J., 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99 (12), 7821–7826. doi:10.1073/pnas.122653799

Guimera, R., *et al.*, 2005. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102 (22), 7794–7799. doi:10.1073/pnas.0407994102

Guo, D., 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22 (7), 801–823. doi:10.1080/13658810701674970

Guo, D.S., 2009. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics (TVCG: Proc. of InfoVis'09)*, 15 (6), 1041–1048. doi:10.1109/TVCG.2009.143

Jiang, B., 2007. A topological pattern of urban street networks: universality and peculiarity. *Physica A: Statistical Mechanics and its Applications*, 384 (2), 647–655. doi:10.1016/j.physa.2007.05.064

Jiang, B., 2009. Street hierarchies: a minority of streets account for a majority of traffic flow. *International Journal of Geographical Information Science*, 23 (8), 1033–1048. doi:10.1080/13658810802004648

Kernighan, B.W. and Lin, S., 1970. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49 (2), 291–307. doi:10.1002/j.1538-7305.1970.tb01770.x

Lambiotte, R., *et al.*, 2008. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387 (21), 5317–5325. doi:10.1016/j.physa.2008.05.014

Liben-Nowell, D., *et al.*, 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102 (33), 11623–11628. doi:10.1073/pnas.0503018102

Liu, Y., *et al.*, 2014. Analyzing relatedness by Toponym co-occurrences on web pages. *Transactions in GIS*, 18 (1), 89–107. doi:10.1111/tgis.12023

McCarthy, C., 2011. *New photo app color hopes to see the future* [online]. Available from: http://www.cnet.com/news/new-photo-app-color-hopes-to-see-the-future/ [Accessed 9 December 2014].

Michalis, F., Petros, F., and Christos, F., 1999. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, 29 (4), 251–262. doi:10.1145/316194.316229

Newman, M.E.J., 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103 (23), 8577–8582. doi:10.1073/pnas.0601602103

Newman, M.E.J. and Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69 (2), 026113. doi:10.1103/PhysRevE.69.026113

Onnela, J.-P., *et al.*, 2011. Geographic constraints on social network groups. *PLoS ONE*, 6 (4), e16939. doi:10.1371/journal.pone.0016939

Palla, G., *et al.*, 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 (7043), 814–818. doi:10.1038/nature03607

Pothen, A., Simon, H., and Liou, K., 1990. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11 (3), 430–452. doi:10.1137/0611030

Ravasz, E., *et al.*, 2002. Hierarchical organization of modularity in metabolic networks. *Science*, 297 (5586), 1551–1555. doi:10.1126/science.1073374

Scellato, S., *et al.*, 2010, Distance matters: geo-social metrics for online social networks. *In*: *Proceedings of the 3rd conference on online social networks*. Boston, MA: USENIX Association, 8.

Scellato, S., *et al.*, 2011, Socio-spatial properties of online location-based social networks. *In*: *5th International AAAI conference on weblogs and social media (ICWSM)*, 17–21 July, Barcelona. Menlo Park, CA: The AAAI Press, 329–336.

Seary, A.J. and Richards, W.D., 2003. Spectral methods for analyzing and visualizing networks: an introduction. *In*: R. Breiger, K. Carley, and P. Pattison, eds. *Dynamic social network modeling and analysis*. Washington, DC: The National Academies Press, 209–228.

Sen, P., *et al.*, 2003. Small-world properties of the Indian railway network. *Physical Review E*, 67 (3), 036106. doi:10.1103/PhysRevE.67.036106

Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (8), 888–905. doi:10.1109/34.868688

Strogatz, S.H., 2001. Exploring complex networks. *Nature*, 410 (6825), 268–276. doi:10.1038/35065725

Tian, Y., Hankins, A.R.A., and Patel, A.J.M., 2008, Efficient aggregation for graph summarization. *In*: *Proceedings of the* 2008 *ACM SIGMOD international conference on Management of data*, 9–12 June, Vancouver. New York: ACM, 567–580. doi:10.1145/1376616.1376675

Tyler, J.R., Wilkinson, D.M., and Huberman, B.A., 2005. E-mail as spectroscopy: automated discovery of community structure within organizations. *The Information Society*, 21 (2), 143–153. doi:10.1080/01972240590925348

Wang, J., *et al.*, 2011. Exploring the network structure and nodal centrality of China's air transport network: a complex network approach. *Journal of Transport Geography*, 19 (4), 712–721. doi:10.1016/j.jtrangeo.2010.08.012

Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442. doi:10.1038/30918

Wu, Z. and Leahy, R., 1993. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (11), 1101–1113. doi:10.1109/34.244673

Xu, X., *et al.*, 2007, SCAN: a structural clustering algorithm for networks. *In*: *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, 12–15 August, San Jose, CA. New York: ACM, 824–833. doi:10.1145/1281192.1281280

Yang, Y., *et al.*, 2012. Predicting links in multi-relational and heterogeneous networks. *In*: *2012 IEEE 12th international conference on data mining (ICDM)*, 10–13 December, Brussels. IEEE, 755–764. doi:10.1109/ICDM.2012.144

Zhou, Y., Cheng, A.H., and Yu, A.J.X., 2009. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2 (1), 718–729. doi:10.14778/1687627.1687709