International Journal of Geo-Information

MDPI

Article

Geographical Structural Features of the WeChat Social Networks

Chuan Ai [1], Bin Chen [1,*], Hailiang Chen [1], Weihui Dai [2] and Xiaogang Qiu [1]

[1] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
[2] School of Management, Fudan University, Shanghai 200433, China
* Correspondence: chenbin06@nudt.edu.cn; Tel.: +86-1378-714-8795

Received: 15 March 2020; Accepted: 22 April 2020; Published: 1 May 2020

Abstract: Recently, spatial interaction analysis of online social networks has become a big concern. Early studies of geographical characteristics analysis and community detection in online social networks have shown that nodes within the same community might gather together geographically. However, the method of community detection is based on the idea that there are more links within the community than that connect nodes in different communities, and there is no analysis to explain the phenomenon. The statistical models for network analysis usually investigate the characteristics of a network based on the probability theory. This paper analyzes a series of statistical models and selects the MDND model to classify links and nodes in social networks. The model can achieve the same performance as the community detection algorithm when analyzing the structure in the online social network. The construction assumption of the model explains the reasons for the geographically aggregating of nodes in the same community to a degree. The research provides new ideas and methods for nodes classification and geographic characteristics analysis of online social networks and mobile communication networks and makes up for the shortcomings of community detection methods that do not explain the principle of network generation. A natural progression of this work is to geographically analyze the characteristics of social networks and provide assistance for advertising delivery and Internet management.

Keywords: spatio-info networks; community detection; MDND; gibbs sampler; adjusted rand index

1. Introduction

The Internet has produced many types of information dissemination methods, including Web, FTP, E-mail, and Telnet. Based on these methods, there are many online social media applications including Facebook, Twitter, and so on, which attract many users. WeChat is the most frequently used social media software in China. The users' behaviors such as instant message, News sharing, and commenting form online social networks. The popularity of the site offers a great deal of opportunities to study the characteristics of online social networks [1]. The progress of network science has gradually brought breakthroughs to research in many fields. The work of Helbing et al. [2] introduces the implications of network science in crowd disasters, crime, terrorism, war, and disease spreading. From the perspective of spatial-networks, there are two important findings in geoscience research, which is based on mobile communication data. The first is that the Law of Universal Gravitation could be used to fit the density of interactions between urban centers [3,4]. Second, some studies have found that the detected communities match the administrative divisions very well [5–7]. These studies followed the tradition that spatial interactions can represent the theory of geographic features, which was established by Ullman [8,9] and persisted and developed by Noronha and Goodchild [10], Castells [11], among others.

The nodes classification is one important method in these studies. These studies usually extract geographic information from different networks, such as mobile phone communication networks and

user interaction networks in the social media application. Then, the networks are converted into new networks with location nodes, that is, spatial information (Spatio-info) networks. Then, the theory and method of the complex network are used to analyze the network. The community detection algorithm is the most common method; we introduce some typical community detection algorithms in Section 2.2. The most typical result is that the nodes in a community are often geographically aggregated, which is generally similar to administrative divisions. The commonly used spatial units include those based on cities, individuals [12–14], pixel grids [6], and Tyson polygons divided by the mobile phone base station [15,16].

However, the community detection algorithms can only analyze the structural characteristics of the network from the perspective of dense or sparse links in the network. These algorithms divide nodes in a network such that the ratio of links inside a community is higher than the ratio of links between communities. Methods in machine learning such as Graph Neural Networks (GNN) [17], which have been studied a lot in recent years, can also be used for node classification, and have achieved excellent performance. However, it needs much more information about the network, such as nodes' attribute, link's attribute, and timing information. These methods do not analyze the network generation process and do not dig deeply into the semantic explanation of the classification effect of network nodes in real life.

Many works care about the cause of complex networks' properties. The work of Matjaž [18] introduced that the Matthew effect reflects many phenomena in real life, and it is closely related to the concept of preferential attachment in network science, where the more connected nodes are destined to acquire many more links in the future than the auxiliary nodes.

Many statistical models about the network fundamentally analyze the network's composition methods and explain the cause for the network characteristics. They are different from community detection and machine learning methods. There are two typical kinds of statistical models for the network. The first one is called the node-exchangeable model. Its main idea is: the nodes in the same class have the equal probability of having links with other nodes. The Stochastic-Block model (SB) [19–21] is an example. The second one is edge-exchangeable models such as Dirichlet Network Distribution (DND) or Mixture of Dirichlet Network Distribution (MDND). They regard the links rather than the nodes as the data form [22]. They are different from the node-exchangeable models. In a general view, there is a strong relationship between the class of information and the person who share it. For example, a sharer of a comic is the most likely to be a comic fan in this circle. From the perspective of the network, the class of links has a strong correlation with their nodes. That is why they are named edge-exchangeable models. The Mixture of Dirichlet Network Distribution (MDND) is introduced in the work of Ghalebi et al. [22]. We also describe the conception and construction methods in detail in Section 4.3.

In this paper, a statistical model based on MDND for network rather than the community detection algorithm is used to classify nodes. The results show that the network statistical model MDND can also find the geographical aggregation phenomenon of nodes in the same class. It is similar to community detection. The MDND model's hypothesis explains the cause of the phenomena that the geographic characteristics of online social networks are strongly related to the classification of links and nodes. In social networks, it means the categories of information are strongly associated with user classification. In the network discussed in this paper, the content of information is related to the geographic location of the user who disseminates the information.

The remainder of the paper is organized as follows. Section 2 describes the preliminaries of this paper, including notations and related works on community detection. Section 3 introduces two typical statistical models for network, including the Stochastic-block model and the Dirichlet Network Distribution model. Section 4 first describes a method to build a Spatio-info network based on users' locations. Then, based on the Spatio-info network, the principle and sampling process to construct MDND are described. Section 5 gives the analysis of the model's results on the Spatio-info network.

Section 6 discusses the advantages and limitations of the model. Section 7 summarizes the content of the article and concludes.

2. Preliminaries

2.1. Notation

In this paper, M is used to indicate the number of nodes in the network, Z is used to represent the $M \times M$ network, and N is used to represent the amount of links. z_{sr} represents the connection number between node s and node r,which is a non-negative integer. Non-zero z_{sr} indicates that there is a link between node s and node r. $\Omega(.)$ indicates whether the expression between the brackets is true or not. If it is true, the value is 1; if not, the value is 0. "Spatio-info networks" in this paper refers to a city-based network constructed from a social network containing geographic information. The specific construction method is described in Section 3. The amount of interactions refers to the number of all user message propagation records between the two cities.

2.2. Community Detection

The community detection of complex networks is an important research filed. A community detection algorithm aims to get a satisfactory classification of nodes. The labels of nodes indicate the classes of nodes, which are also the results of the algorithms. A community detection algorithm usually contains many iterations, and the labels of nodes will be updated upon each iteration. Iteration will be terminated when achieving the goal of having more links within the community than between the communities. Different algorithms have different approaches to change the label of nodes in each iteration or approaches to measure whether the goal is achieved or not.

Newman and Girvan [23] proposed modularity to measure the results of community detection. Modularity is defined as the difference between the ratio of edges in communities of a network and the expected ratio in a random network.

$$Q = \sum_{k} \sum_{ij \in C} (realflow_{ijk} - estflow_{ijk}) \tag{1}$$

In Equation (1), k is the number of communities, $realflow_{ijk}$ gives the ratio of the edge between node i and j in the same community, and $estflow_ijk$ gives the expected value in a random network. The modularity of different scales of network fluctuates between 0.3 and 0.7.

Newman [24] proposed the most widely used method, which is the Newman modularity maximization method. In practical projects, there are several algorithms to support the actual network analysis in consideration of time complexity and flexibility. The fast greedy algorithm is a hierarchical agglomerative algorithm, the time complexity of which is $O(md \lg n)$, where m represents the number of edges, n is the number of nodes, and d is the tree diagram depth of community structure.

Blondel [25] applied the multilevel community detection algorithm. However, the work was based on the dataset of mobile phone communication, and the network was based on individuals devoid of spatial interaction information. Thus far, studies that try to detect the multilevel communities based on spatial interaction are rare, except for works of Sobolevsky et al.. [26], De Montis et al. [24], and Guanghua [16]. The former two works are based on the modularity maximization algorithm, and the last one is based on the Infomap algorithm. Sobolevsky et al. [26] detected deep and small communities based on the communities that formed before. Different from others, this kind of method gets large communities after those small communities are obtained. The work in this paper is based on the algorithm of Informap without detecting multilevel communities. There are many algorithms to detect communities [26]. Lancichinetti and Fortunato [27,28] found that the algorithm of Informap achieves satisfactory performance in different situations. Rosvall and Bergstrom [29] had a more detailed description about this algorithm.

In the research of Chuan et al. [30] and Guanghua [16], cities and base stations for mobile phones are regarded as nodes. The nodes can be divided into classes by community detection. The most prominent features of the community are the aggregation and continuity of the city or base station nodes. According to the general idea, the interaction in the network should not follow the law of face-to-face interaction, so that the city nodes of a community should not be geographically aggregated in one area. However, many previous results show that the geographical distance in the network interaction is preserved, and the city nodes in a community also locate together. It can also be explained that the communication of public opinion is affected by geographical factors to a large extent.

3. Statistical Model

3.1. Stochastic Block-Model

Stochastic Block-model (SB) and its associated models belong to a very important statistic model class. The basic SB assumes that each node belongs to one of the potential K classes. For any two nodes (i, j), the probability that they have connection is determined by a specific parameter θ_{c_s,c_r}. Under the condition that the classification is determined to be c_s and c_r, the probability that there is a link between node s and r obeys the distribution with θ_{c_s,c_r} as a parameter. Figure 1a is the adjacency matrix of a network generated in this way.

According to Snijder's [19] method, to reconstruct SB with the Bayesian statistics theory, a conjugate prior probability could be placed for classification and other parameters. Based on the conjugate prior probability and likelihood functions, the posterior distribution function could be written to get a new network model.

$$\begin{aligned} z_{sr} &\sim \text{Bernoulli}(\theta_{c_s,c_r}) \\ c_s &\sim \text{Discreste}(\pi), s \in \{1, ...M\} \\ \theta_{i,j} &\sim \text{Beta}(\alpha, \beta), i, j \in \{1, ..., K\} \\ \pi &\sim \text{Dirichlet}(\text{Œ}) \end{aligned} \tag{2}$$

The Beta distribution and the discrete Dirichlet distribution are used as the prior probability of the connection probability nodes classification, respectively. The SB formed by Bayesian statistics theory is shown in Equation (2).

The SB can be extended into many forms. For example, the Infinite Relational Model (IRM) [31] achieves the goal of allowing infinite classes by placing the Dirichlet process as a prior probability for the classification distribution. This eliminates the need to set an accurate number of classes in advance and allows the number of classes to increase as the number of nodes increase.

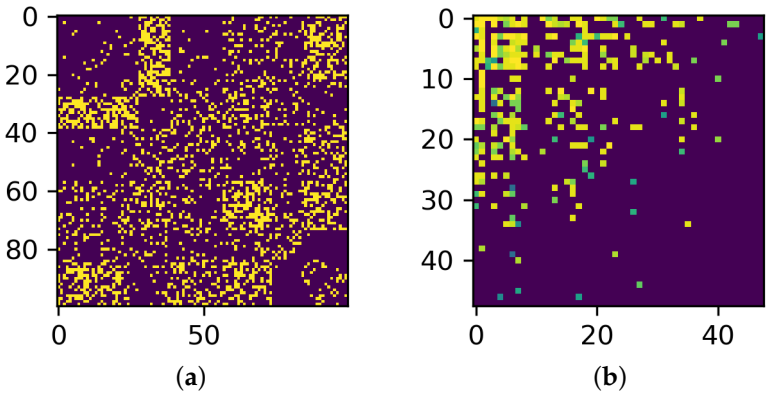

Figure 1. *Cont.*

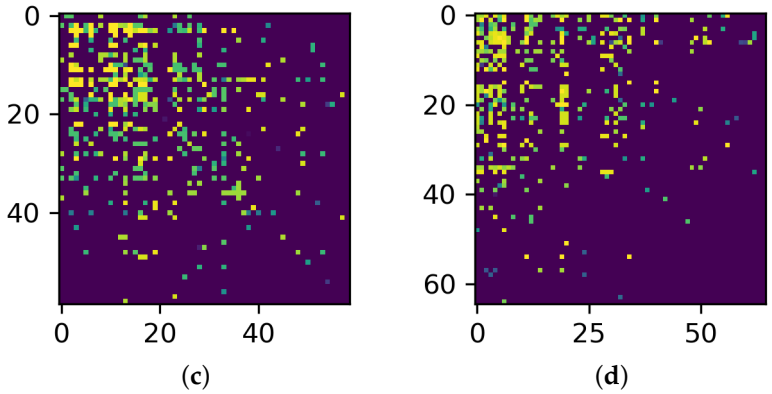

Figure 1. (**a**) The network adjacency matrix of a SB with seven blocks. (**b**) Sample from a symmetric Dirichlet network distribution ($\tau = 10, N = 500$). (**c**) Sample from an asymmetric Dirichlet network distribution ($\gamma = 50, \tau = 10, N = 500$). (**d**) Sample from an asymmetric mixture of Dirichlet network distribution ($\gamma = 20, \tau = 5, \alpha = 5, N = 1000$). The horizontal and vertical coordinates are the numbers of the nodes in the network. If there are links between the two nodes, there is a bright spot at the point with the corresponding horizontal and vertical coordinate. The number of links is reflected by the brightness of the point. In (**a**), the bright spots obviously are distributed in blocks, which is also a characteristic of the Stochastic-block model. In (**b**,**c**), the sampling results of symmetric and asymmetric Dirichlet network distributions are shown. In (**b**) there is obvious symmetry characteristic. In (**d**), there is little symmetry characteristic, but it is difficult to see the more complex nature in this figure.

Most of these models, including the SB, IRM, and MMSB, describe the structural characteristics of the network. However, they cannot describe the links of the network from a global perspective and cannot predict new links. None of these models can model the networks with increasing nodes.

3.2. Dirichlet Network Distribution

The Dirichlet Network Distribution (DND) constructs a network without the limitation of nodes in a very simple way. It generates a distribution based on a countable number of infinite nodes. Thus, a network can be represented as a sequence of (sender, receiver) pairs, and each pair corresponds to a phone, message, mail, or journey from the sender to the receiver.

The SB and associated models assume that the network is static and fully observable. This is why these models cannot predict new links. DND regards the network as a sequence of observed connections, aiming at predicting the classification of new links rather than nodes.

To achieve the purpose of generating a pair of nodes, it is possible to simply sample the sender and receiver from a distribution G based on all nodes. As in Equation (3), taking N as the total number of links, a prior probability of the Dirichlet process is placed for G. y^n represents the nth links, which is a message from sender s_n to receiver r_n, respectively.

This network construction method is called Symmetric Dirichlet Network Distribution (SDND). Figure 1b shows the network generated in this way. The features of these models are that the distribution of the sender and receiver of any link of the network is exactly the same.

$$\begin{aligned} G &\sim \mathrm{DP}(\tau, \Theta) \\ y^n &= (s_n, r_n), n = 1, ..., N \\ Y &= \{y^1, ..., y^N\} \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \Omega(s_n = i, r_n = j) \\ s_n, r_n (i.i.d) &\sim G \end{aligned} \tag{3}$$

However, there are some problems. In real-life networks, senders and receivers are not completely independent and identically distributed. For example, on Facebook or Twitter, more users are concerned about politic stars such as Trump than those who care about general users. From the perspective of national migration, with the process of urbanization, in the network of population migration, many of the migrants come from remote areas such as rural areas or townships and move to more economically developed cities or county town. That is to say, the sender and the receiver do not obey the same distribution when modeling the network.

Therefore, it is a more reasonable choice to describe the sender and receiver by using the asymmetric distributions A and B instead of the distribution G. Based on this idea, Sinead et al. [25] constructed different sender distribution A and receiver distribution B based on a discrete base measure H. It is shown in Equation (4).

$$\begin{aligned} H &:= \sum_{i=1}^{\infty} h_i \delta_{\theta_i} \sim \mathrm{DP}(\gamma, \Theta); s_n \sim A, n = 1, ..., N; \\ A &:= \sum_{i=1}^{\infty} a_i \delta_{\theta_i} \sim \mathrm{DP}(\tau, H); s_n \sim B, n = 1, ..., N; \\ B &:= \sum_{i=1}^{\infty} b_i \delta_{\theta_i} \sim \mathrm{DP}(\tau, H); z_{ij}^{(N)} = \sum_{n=1}^{N} \Omega(s_n = i, r_n = j). \end{aligned} \tag{4}$$

Figure 1c is a network generated with this method. However, this method can only solve the problem of asymmetry range between the sender and receiver. The social networks in real life are complex. For example, the followers of a certain topic in social networks are generally different from other topics. Animation, car, and fashion are communication topics that are separated and intertwined. The sender and receiver of one topic are likely to be inconsistent with other information. Generally speaking, the person who sends and receives the information is related to the topic of a message. In this case, all the models mentioned above have obvious defects and cannot apply to such networks.

4. MDND Model Based on Spatio-Info Networks

4.1. Data and Networks

We construct the Spatio-info network to analyze the geographical characteristics with the method in the work of Chuan et al. [30]. The main idea of this method is to abstract the information interaction among users in online social networks as the information transmission among cities. Then, a city's network is produced.

WeChat is the most widely and frequently used social media software for Chinese. It is mainly used on the mobile phone and can also be used on PCs and tablets. It has the function of instant messaging, photo display, and comment. It has Official Wechat Accounts, which are managed by government agencies, schools, companies, individuals, etc. The general users follow these accounts, and then receive their messages, news, reviews, articles, and so on. A record will be produced when a WeChat user clicks to browse the pages shared by others. These records became our data source. The data collection company (Fabonacci) collected a large number of historical records of pages propagated among users, forming a dataset.

Each record in the data refers to a behavior that a user clicked and browsed a webpage. A record includes the sharer ID, the viewer ID, the webpage ID, the viewer's IP, and the browsing time. To protect privacy, the data collection company makes users' IDs unique to the users' WeChat ID, but they are not the same. Through the query of the IP library, the geographical location of the user could be obtained. Song Jian et al. [32] analyzed coverage and coincidence rate of several major IP address libraries, including *IP2Location Lite*, *GeoLite2*, *Pure IP Address Library*, *Taobao IP Address Library*, *Sina IP Address Library*, and *Baidu IP Address Library*. They concluded that the *Taobao IP address library*

has the highest credibility and the highest credibility at the city level. In this paper, it is decided to use the *Taobao IP Address Library* to geographically locate the IP addresses involved in the data.

M^h indicates the number of users in the dataset and M^c denotes the number of cities in the dataset. First, the interaction records in the data can be considered as the sequence of links. N represents the number of links in the sequence. The interactive network can be represented as in Equation (5). This is an example of the online social network.

$$\begin{aligned} y^n &= (s_n, r_n), n = 1, ..., N \\ Y &= \{y^1, ..., y^N\} \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \Omega(s_n = i, r_n = j), i, j \in \{1, ..., M^h\} \end{aligned} \tag{5}$$

Each user node has a city assignment of g_i. The interactions of users between two cities are regarded as the interactions of cities. Each interaction record can be regarded as a directed edge, which is expressed as a city pair. The dataset can be seen as a sequence of city pairs, as in Equation (6). The interactive network between cities can be represented in the form of an adjacency matrix. The network generation process is also shown in Figure 2. In Figure 2a, the logo of WeChat represents the media users, and the links represent the information spreading among the users. The users and the links form a network. In Figure 2b, the users in a city are regarded as an overall entity, and the links from one city to another are regarded as a single link between the respected cities.

$$\begin{aligned} x^n &= (g_{s_n}, g_{r_n}) \\ X &= \{x^1, ..., x^N\} \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \Omega(g_{s_n} = i, g_{r_n} = j), i, j \in \{1, ..., M^c\} \end{aligned} \tag{6}$$

The data from 05:00:00 to 06:00:00 on 22 April 2015 are analyzed, and three successively enlarged regions in central China are selected to form three networks with cities as nodes. The basic statistical characteristics of the networks are shown in Table 1.

Network 2 adds city nodes of two provinces on the basis of Network 1, and Network 3 adds more city nodes of another two provinces on the basis of Network 2. A Mixture of Dirichlet network distribution model is constructed. At the same time, in Section 5, the analysis results of the MDND model and the community discovery results are compared and analyzed.

Table 1. Characteristics of three networks.

	Nodes	Links	Average Degree
Network 1	57	592	20.77
Network 2	87	784	18.02
Network 3	107	977	18.26

4.2. Mixture of Dirichlet Network Distribution

Based on the three networks obtained in Section 4.1, the MDND model of Spatio-info Networks is constructed according to the Hierarchical Dirichlet Process, and the Gibbs sampler is constructed by combining the sampling methods of the Hierarchical Dirichlet Process.

In addition to the social networks mentioned in Section 2, there are some specific issues in online social networks. Generally speaking, different topics attract different people, thus the people involved in information dissemination are different. Therefore, the distribution of senders and receivers in such networks is related to specific information topics. When modeling networks, the senders and receivers

obey different distributions corresponding to the topics. MDND can describe the case where different classes of links are strongly related to the links' nodes. This is a feature that many other statistical models of network such as SB and DND lack. Figure 2b shows the adjacency matrix of a random network generated in this way.

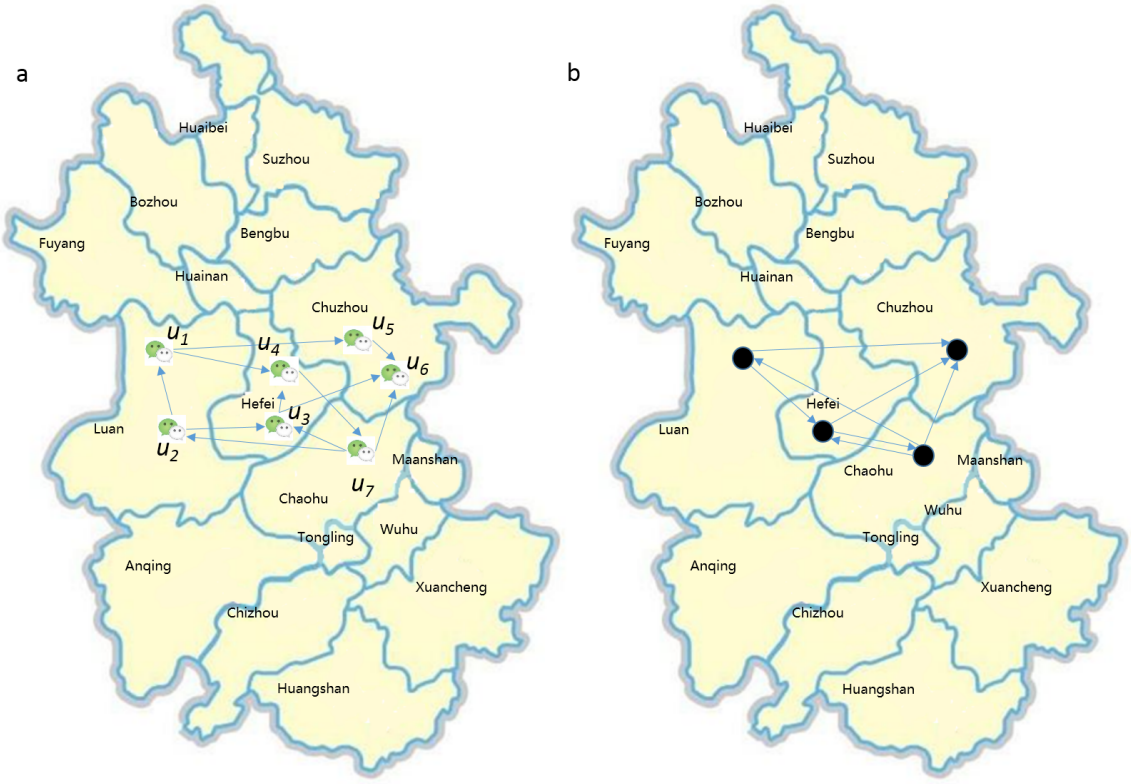

Figure 2. Schematic diagram of network construction. (**a**) The logos of WeChat represents the application users, and the links represent the messages spreading among the users. The users and the links together form a network. (**b**) The users in a city are regarded as an overall entity, and the links from one city to another are regarded as a single link between the respected cities. For example, u_3 and u_4 in (**a**) are integrated into a node *hefei* in (**b**).

According to Sinead's work, the MDND model based on Spatio-info network can be represented by Equations (7) and (8), in which α controls the number of classes, γ controls number of nodes, and τ controls the overlap between classes. The process of generating a network with MDND can be simply understood as a process of gradually generating a series of links.

$$\begin{aligned} &x^n = (g_{s_n}, g_{r_n}), X = \{x^{(1)}, ..., x^{(N)}\}; \\ &\mathrm{D} := (d_k, k \in \mathbb{N}), c_n \sim \mathrm{D}, n = 1, ..., N. \end{aligned} \tag{7}$$

$$\begin{aligned} H &:= \sum_{i=1}^{\infty} h_i \delta_{\theta_i} \sim \mathrm{DP}(\gamma, \Theta); \\ A_k &:= \sum_{i=1}^{\infty} a_{k_i} \delta_{\theta_i} \sim \mathrm{DP}(\tau, H), \qquad g_{s_n} \sim A_{c_n}; \\ B_k &:= \sum_{i=1}^{\infty} b_{k_i} \delta_{\theta_i} \sim \mathrm{DP}(\tau, H), \qquad g_{r_n} \sim B_{c_n}; \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \Omega(g_{s_n} = i, g_{r_n} = j), \quad i, j \in \{1, ..., M^c\}. \end{aligned} \tag{8}$$

4.3. Sampling Method

The symmetrical Dirichlet network model is sampled to predict new links based on the Chinese restaurant process [33]. The MDND model is based on a hierarchical Dirichlet process, thus it is appropriate to construct a collapsed Gibbs sampler [34].

First, $\eta_k = \sum_{i=1}^{N} \Omega(c_i = k)$ indicates the number of links connected to the class k. $m_{k,i}^{(1)}$ indicates the number of links in the class k that sender is i . $m_{k,i}^{(2)}$ indicates the number of links in the class k for which the receiver is i. The probability that node i is connected to class k as sender and receiver is $\rho_{i.}^{(k)}$ and $\rho_{.i}^{(k)}$. The posterior probability distribution calculation method is shown in Equation (9). $s(m_{i.}^{(k)}, \rho)$ is unsigned Stirling function. The probability of each node is determined by the probability measure of the edge to which it is connected, given by Equations (10) and (11) . $\beta_1, ..., \beta_{M^c}$ correspond to existing nodes, while β_μ corresponds to a new node.

$$P(\rho_{i.}^{(k)} = \rho | c_{1:N}, \rho_{i.}^{(k)-n} \beta_{1:N}) = \Gamma(\tau\beta_i)\Gamma(\tau\beta_i + m_{i.}^{k})^{-1} s(m_{i.}^{k}, \rho)(\tau\beta_i)^{\rho} \tag{9}$$

$$\rho_{i}^{(.)} = \sum_{k} \rho_{i.}^{(k)} + \rho_{.i}^{(k)} \tag{10}$$

$$(\beta_1, ..., \beta_{M^c}, \beta_\mu) \sim \mathrm{Dir}(\rho_1^{(.)}, ..., \rho_{M_c}^{(.)}, \gamma) \tag{11}$$

In the case where the classification of all other links are given, the distribution of the classification of the link n is shown in Equations (12) and (13) .

$$P(c_n = k | s_n, r_n, c^{-n}, \beta) \propto \eta_k^{-n} (m_{k,s_n}^{(1)-n} + \tau\beta_{s_n})(m_{k,s_n}^{(2)-n} + \tau\beta_{r_n}) \qquad \eta_k^{-n} > 0 \tag{12}$$

$$\propto \alpha\tau^2 \beta_{s_n}\beta_{r_n} \qquad \eta_k^{-n} = 0. \tag{13}$$

$P_{e(N+1)} = P(y_{N+1} = (s, r) | c_{1:N}, y_{1:N}, \beta)$ is the distribution for predicting the N+1st link, which is shown in Equation (14).

$$\begin{aligned} P_{e(N+1)} &= \sum_{k=1}^{K+} \frac{\eta_k}{N+\alpha} \frac{m_{k,s}^{(1)} + \tau\beta_s}{\eta_k + \tau} \frac{m_{k,s}^{(2)} + \tau\beta_r}{\eta_k + \tau} + \frac{\alpha}{N+\alpha}\beta_s\beta_r, && s, r \le M^c, \\ &= \sum_{k=1}^{K+} \frac{\eta_k}{N+\alpha} \frac{m_{k,s}^{(1)} + \tau\beta_s}{\eta_k + \tau} \beta_\mu + \frac{\alpha}{N+\alpha}\beta_s\beta_\mu, && s \le J, r > M^c, \\ &= \sum_{k=1}^{K+} \frac{\eta_k}{N+\alpha} \frac{m_{k,s}^{(2)} + \tau\beta_r}{\eta_k + \tau} \beta_\mu + \frac{\alpha}{N+\alpha}\beta_r\beta_\mu, && r \le J, s > M^c, \\ &= \beta_\mu^2, && s, r > M^c. \end{aligned} \tag{14}$$

5. Results and Analysis

We use the MDND model to analyze the network and obtain the links' class assignments and nodes' class assignments by sampling. We compare the node classification effect of MDND and community detection algorithms, and then the MDND model's characterization of Spatio-info networks is verified, thereby explaining the formation principle of Spatio-info networks. We focus on the characteristics of the hierarchical structure in the geographic information network. The characteristics of the hierarchical structure in the Spatio-info networks are mainly reflected in that there are more interactions between cities within the provincial administrative division and fewer interactions with the outside. This phenomenon has been confirmed by the community detection algorithm. However, from the actual situation analysis, it can be speculated that the interaction classification has a strong relationship with the regional distribution of city nodes.

5.1. Classification of Links

The three networks are analyzed separately, and the network is inferred based on the network data with the MDND. Different link's classes are denoted by different colors, as shown in Figure 3. Each subfigure of Figure 3 is similar to a matrix: the x-axis and y-axis represent city number series. The cities in the same province are placed together. The order of the city number series of the x-axis and y-axis is the same. If two cities (i and j) have a link, then there should be a point with a specific color at (i,j).

It can be clearly found that the cities in the same province are basically in the same category, and the links among different provinces also exhibit an effect of blocks. Of course, it can be clearly seen in the figure that the MDND model has a strong description ability for the structure of the network. Regardless of the size of the network, the increase of nodes does not significantly affect the quality of the model.

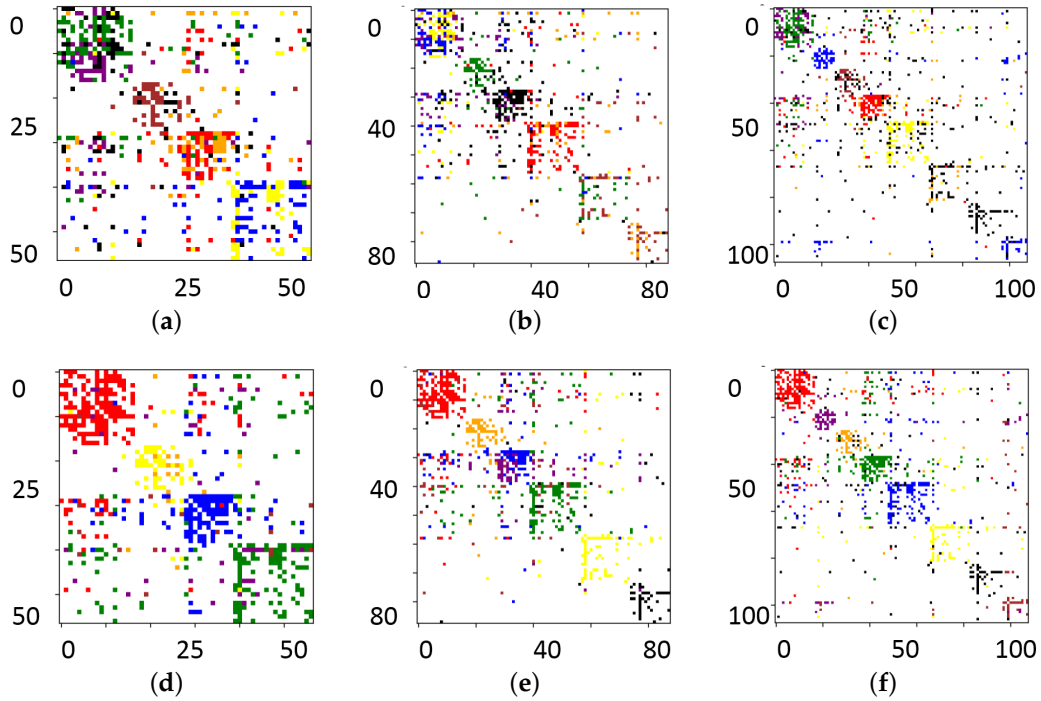

Figure 3. (**a**–**c**) The states after 10 times training of the class assignments in calculation, corresponding to Networks 1–3 in Section 3. (**d**–**f**) The states after 100 times training. The horizontal and vertical coordinates are the numbers of the nodes in the network, where different nodes represent different cities. If there are links between two nodes, there is a bright spot at the point with the corresponding horizontal and vertical coordinate. The different colors indicate the different classifications of links.

5.2. Classification of Nodes

Based on simple rules, according to the classification of links, the classification of a single node is determined by the dominant class of its links. The goal of this paper is to explore the model's ability to describe the structural characteristics of Spatio-info networks. Therefore, it is a good choice to compare the classification of nodes with the administrative division.

Figure 4 is in the form of the Spatio-info networks, in which the nodes represent cities, the links represent the links between cities, and different colors of nodes indicate the nodes are in different classes. This figure shows the differences among the classification results of administrative divisions, the community detection, and the MDND model from the perspectives of the network.

Figure 5 is in the form of maps. This figure also shows the differences among the classification results of administrative divisions, including community detection and the MDND model. Figures 4 and 5 give the same content in different forms.

It can be clearly seen in the two figures that both the community detection algorithm and the MDND model can well discover the structural characteristics of the network that are closely related to geographical factors. However, there are also some differences. Among the analysis results of Network 1, the community detection, MDND model, and administrative division are completely consistent. In the result of Network 2, the results of community detection are consistent with administrative divisions, but the results of MDND models are obviously inconsistent. In the result of Network 3, the community detection algorithm maps the cities of Jiangxi and Fujian provinces into the same category, and the MDND model successfully separates the cities of the two provinces.

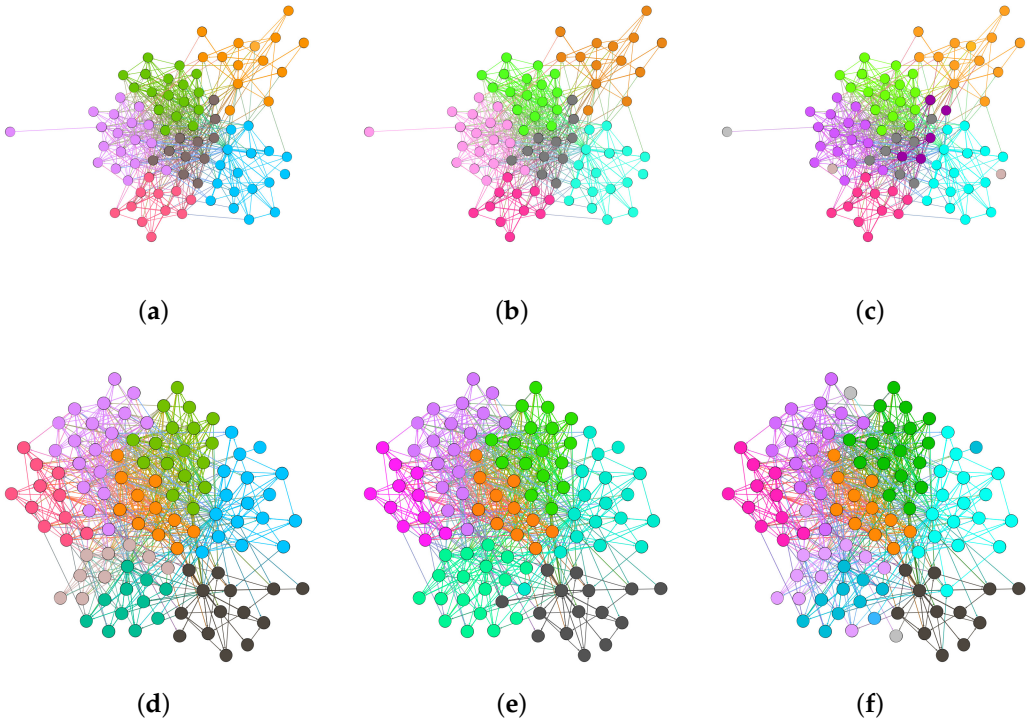

Figure 4. The nodes' classification of Spatio-info network, where different colors represent different class: (**a**–**c**) Network 2; and (**d**–**f**) Network 3. (**a**,**d**) The administrative division; (**b**,**e**) the result of the community detection algorithm; and (**c**,**f**) the classification result of the MDND. This is in the form of the Spatio-info Networks, in which the nodes represent cities, the links represent the links among cities, and different colors of nodes indicate the nodes are in different classes.

To better analyze the classification results of MDND model, this paper introduces the Adjusted Rand Index (ARI) [16], which measures the similarity between two division results. It is needed to draw a contingency table and calculate the ARI according to the contingency table as shown in the matrix in Equation (15). According to the calculation method shown in Equation (16), the ARIs between any two divisions are shown in Table 2, noting that *AD, MC, and CR* mean administrative division, modularity class, and classification results of MDND model, respectively. It can be seen that both the MDND and community detection results are nearly the same as administrative divisions. With the increase of network scale, the similarity is lower for both MDND and community detection results. Based on this, it can be known that the MDND can model such Spatio-info networks, which are obviously affected by geographical factors, and accurately reflect the structural information related to geographical factors in the network.

$$\begin{bmatrix} X/Y & Y_1 & Y_2 & \cdots & Y_s & Sum_s \\ X_1 & n_{11} & n_{12} & \cdots & n_{1s} & a_1 \\ X_2 & n_{21} & n_{22} & \cdots & n_{2s} & a_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_r & n_{r1} & n_{r2} & \cdots & n_{rs} & a_r \\ Sum_r & b_1 & b_2 & \cdots & b_s & \end{bmatrix} \tag{15}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \binom{n}{2}^{-1}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \binom{n}{2}^{-1}} \tag{16}$$

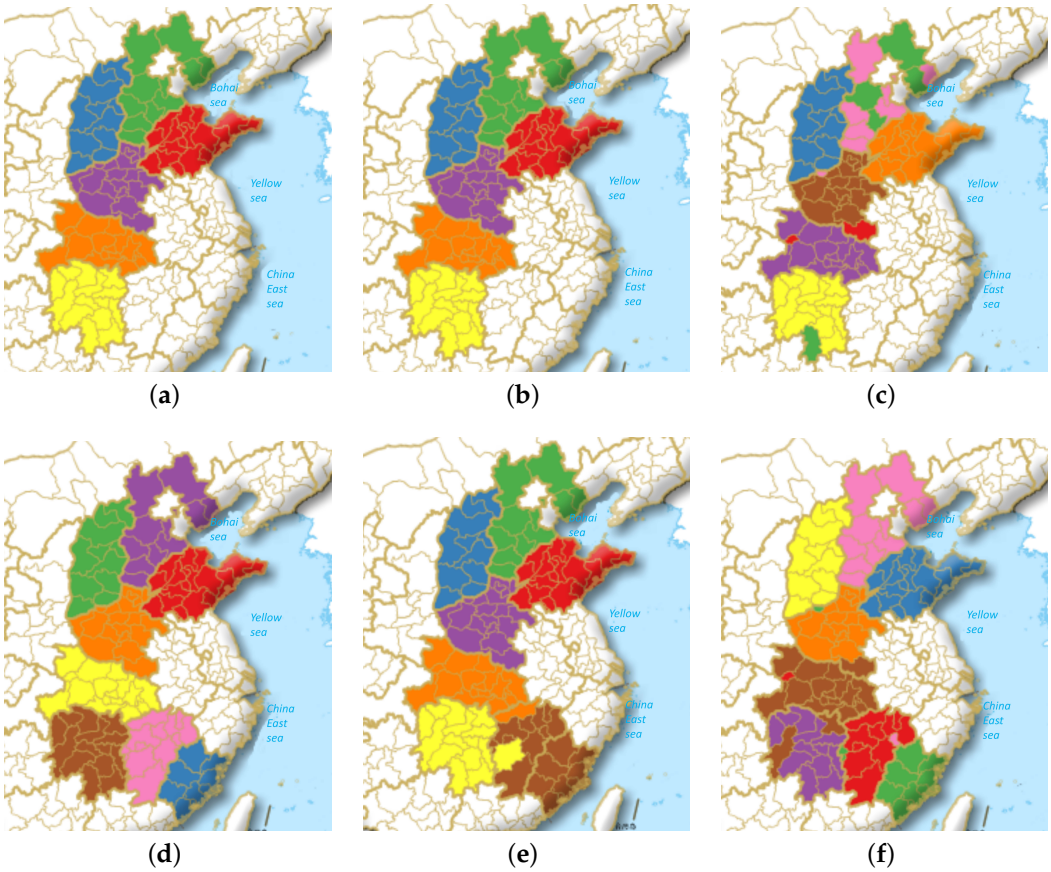

Figure 5. The different divisions displayed geographically. The same color means that these city nodes are assigned to the same class: (**a**,**d**) city divisions; (**b**,**e**) the community detection result; and (**c**,**f**) the calculation result of the MDND model. (**a**–**c**) The analysis results of Network 2; and (**d**–**f**) the analysis results of Network 3. These figures are in the form of maps. The cities are drawn in the same color in each figure if they are in the same class. It can be seen that there is little difference between any two of the three classifications.

Table 2. Comparison of the analysis results of the MDND on the three networks with the administrative divisions and community detection algorithm.

	Network 1	Network 2	Network 3
AD-MC	1	0.98	0.89
AD-CR	1	1.00	0.98
CR-MC	1	0.98	0.83

5.3. Computational Performance Analysis

According to the Gibbs sampler of the MDND, the calculation results will be different as the number of iteration increases. Figure 6 shows the similarity index ARI of the MDND result and the administrative division as the number of iteration increase. It can be seen that as the number of iteration increases, the ARI value gradually approaches 1, which means that the classification results of MDND are consistent with the administrative division. It shows that the number of iterations increases, the ARI value increases rapidly in the initial stage, and then the increasing speed gradually slows down. At the same time, as the network scale increases, the speed at which the ARI value increases to 1 is also different. In Table 2, it can be seen that the final results of Networks 1 and 2 will both be 1. However, as the scale grows, the increasing speed will slow down.

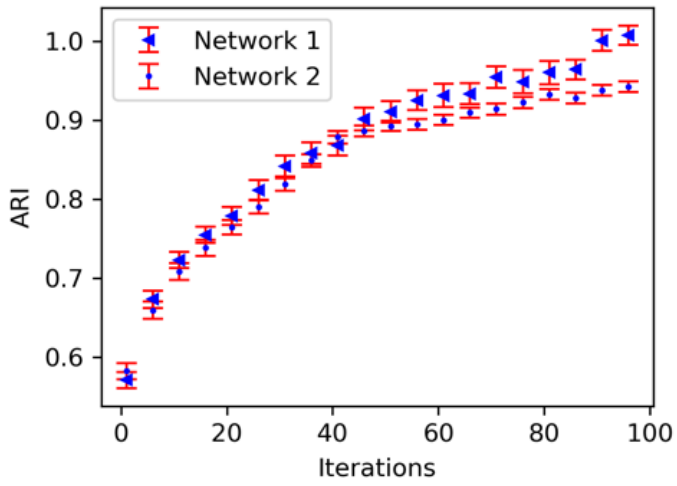

Figure 6. The classification result with MDND of Networks 1 and 2 with the increase of the iteration. It can be seen that as the iteration times increases, the ARI value gradually approaches 1, and the volatility is small.

6. Two Special Cases

There are some special circumstances in the Spatio-info network with cities as nodes. First, there may be cases where the network is extremely sparse. For example, the economic situation in some areas is not good, so the Internet infrastructure is relatively poor. In some areas, because of the strong protection of data privacy, only a small amount of information can be collected. Further analysis is needed to determine whether the MDND is suitable in the case where only minimal interaction information can be gathered. In addition, many neighboring regions have different levels of economic and Internet development. For example, Guangdong and Guangxi provinces in China border each other, but the economic situation and the Internet infrastructure of Guangdong province are much better. This fact causes the unbalanced nature of the Spatio-info network in the scope of Guangdong and Guangxi, and whether this feature will be reflected in the MDND is of great significance.

6.1. Sparse Network

For the sake of comparison, the Spatio-info Network 1 formed in Section 5 is sampled to obtain a sparse network. The basic information of the sparse network is shown in Table 3. The average degree of the network is 3.24. It is modeled with the MDND. The links and nodes are classified. We then compare the classification assignments of nodes with administrative divisions and calculate ARI. The ARI value changes with the increase of iterations, which is shown in Figure 7.

Table 3. Interaction amounts.

	GD to GD	GX to GX	GD to GX	GX to GD
Link amounts	548	50	22	20

Notes: GD, Guangdong Province; GX, Guangxi Province.

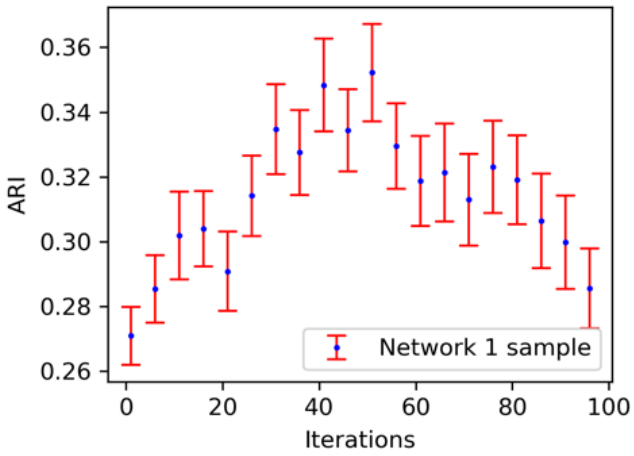

Figure 7. The similarity (ARI) varies with the increasing iteration times between the classification results of the MDND on and the administrative divisions. It can be seen that the ARI value does not increase obviously with the increase of iterations, and the volatility is very large.

It can be seen in Figure 7 and Table 4 that, under the condition of a sparse network, MDND is not good for geo-related structural information mining in social networks. The ARI value does not increase significantly with the increase of iteration times, and the volatility is very large. It can be seen that, under the condition of an extremely sparse network, the classification assignments of the two methods are quite different from the administrative division, but the community detection results are better. Therefore, the MDND model is not stable enough to deal with the sparse Spatio-info network.

Table 4. Comparison of the analysis results of MDND on the sparse network with the administrative divisions and community detection algorithm.

	Network 1	Network 1 Sample
AD-MC	1	0.34
AD-CR	1	0.59
CR-MC	1	0.20

Notes: AD, Administrative Division; MC, Modularity Class; CR, classification Results of MDND.

6.2. Unbalanced Network

Imbalance in this paper refers to the fact that, in a Spatio-info network, the interaction amount of an area is significantly less than that of another area. The geographic information interaction networks with cities as nodes in Guangdong and Guangxi provinces are obtained in the same way as

the networks in Section 5. First, 640 records are obtained. There are 548 in Guangdong Province and 50 in Guangxi. The details are shown in Table 4. The interactions in Guangdong is significantly more than that in Guangxi.

MDND is used to model and analyze the Spatio-info network, and the comparison between the classification result and the administrative division is shown in Figure 8 and Table 5. It can be seen that the increase of ARI value is obvious with the increase of the number of iteration , and the volatility is very large. Moreover, neither the classification of MDND nor the community detection result is similar to the administrative division. The results show that it is difficult for MDND to mine the geographically related structural information in the Spatio-info network in this special case. A new model should be built when dealing with unbalanced geographic information networks.

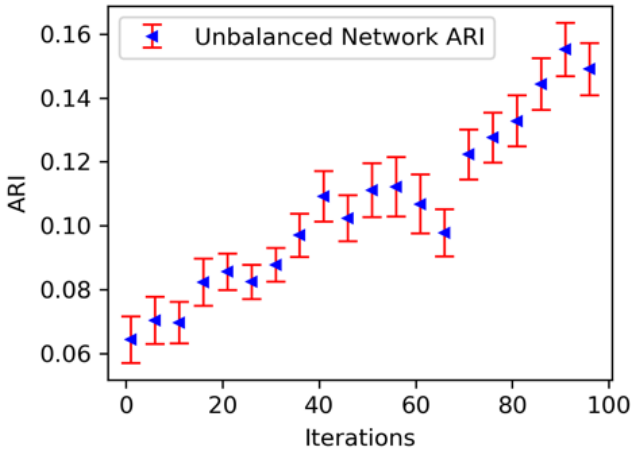

Figure 8. The similarity (ARI) varies with the increasing iteration times between the classification results of the MDND and the administrative divisions. It can be seen that the increase of ARI value with the increase of the iteration times is obvious, and the volatility is very large.

Table 5. Comparison of the analysis results of the MDND on the sparse network with the administrative divisions and community detection algorithm.

	Network 1	**Unbalanced Network**
AD-MC	1	0.11
AD-CR	1	0.10
CR-MC	1	0.27

Notes: Ad, Administrative Division; MC, Modularity Class; CR, classification Results of MDND.

7. Conclusions

This paper focuses on the analysis of geographic characteristics in various networks with complex network methods. These studies usually extract geographic information from different networks, such as mobile phone communication networks and user interaction networks in the social media application. Then, the networks are converted into networks with location nodes, that is, spatial information networks. The theory and method of the complex network are used to analyze the network. The most typical result is that the nodes in a community are often geographically aggregated, which is generally similar to administrative divisions. The principle of community detection is that there are more links within the community. This explains, to a certain extent, the phenomenon that the nodes in the community in the network also aggregate geographically. It is said that, whether online social network or phone-call communication networks, there will be more communication between geographically close regions than other regions.

This paper tries to analyze spatial information networks with network statistical models other than community detection. The principles, advantages, and disadvantages of a series of network statistical models are analyzed. The MDND model is chosen in this paper. The construction and implementation details of the MDND model are also introduced. Finally, the analysis results are presented in the form of nodes and links classification. The results are compared with the results of the community detection algorithm and administrative division. It is found that the three types of divisions show a high degree of agreement. It shows that the MDND model can also analyze the geographically aggregating phenomena of nodes in the same community.

At the same time, there are hypotheses and mathematical basis on the construction process of the MDND model. Therefore, the hypothesis of the model explains the network constructed in the paper to a degree. That is to say, the geographically aggregating of nodes in this network may not only be due to the fact that there are more links within the community, but also because the classification of links is correlated with the nodes. Of course, such hypotheses and explanation require more empirical data to analyze and confirm. However, the classification ability of the MDND model for links and nodes of such networks can be used for further analysis and research.

The nodes grouping ability in the network of the MDND model is similar to the community detection method used to group network nodes, and the link classification ability can be used to classify information in social media networks in the future. As for Spatio-info networks, i.e., the MDND model, similar to other statistical models, can predict the rest of the network by building the model of the network with a part of data. For example, 90% of the links in the network are known, and the remaining 10% can be predicted with these methods. This is a typical task in link prediction. Therefore, in the future, with the help of the MDND model, link prediction methods in complex networks will be used to analyze geographic information-related networks, and more valuable results can be obtained. The link prediction methods and community detection algorithms could be used in advertising delivery and Internet management.

The significance of the work is to analyze the spatial structure characteristics of social networks from the perspective of statistical models for the first time. The performance of its node classification can achieve the performance of community detection. The model can mathematically reflect the spatial structure features contained in the network. At the same time, the network model has the great advantage of predictability as it is a statistical model for the network. Naturally, this work can be carried out in the future, and it will play a role in the prediction of the geographical scope of information dissemination and provide assistance for many aspects such as advertising delivery, Internet management, and so on.

Author Contributions: Conceptualization, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Data Curation, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Formal Analysis, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Funding Acquisition, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Investigation, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Methodology, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Project Administration, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Resources, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Software, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Supervision, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Validation, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Visualization, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; Writing—Original Draft, Chuan Ai, Bin Chen, Hailiang Chen, and Xiaogang Qiu; and Writing—Review and Editing, Chuan Ai, Bin Chen, Hailiang Chen,Weihui Dai, and Xiaogang Qiu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key Research & Development (R&D) Plan under Grant No. 2018YFC0806900; the National Natural Science Foundation of China under Grant Nos. 71673292 and 71673294; National Social Science Foundation of China under Grant No. 17CGL047; and Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion.

Acknowledgments: Thanks to the Shenzhen Fibonacci Data Consulting Co., Ltd. for the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mislove, A.; Marcon, M.; Gummadi, K.P.; Druschel, P.; Bhattacharjee, B. Measurement and Analysis of Online Social Networks. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, San Diego, CA, USA, 24–26 October 2007; ACM: New York, NY, USA, 2007; pp. 29–42.
2. Helbing, D.; Brockmann, D.; Chadefaux, T.; Donnay, K.; Blanke, U.; Woolley-Meza, O.; Moussaid, M.; Johansson, A.; Krause, J.; Schutte, S.; et al. Saving Human Lives: What Complexity Science and Information Systems can Contribute. *J. Stat. Phys.* **2015**, *158*, 735–781. [CrossRef] [PubMed]
3. Krings, G.; Calabrese, F.; Ratti, C.; Blondel, V.D. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech. Theory Exp.* **2009**, *2009*, L07003. [CrossRef]
4. Roth, C.; Kang, S.M.; Batty, M.; Barthélemy, M. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE* **2011**, *6*, 5923. [CrossRef] [PubMed]
5. Vincent, G.K.; Thomas, I. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Bruss. Stud.* **2010**, *2*. [CrossRef]
6. Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R.; Strogatz, S.H. Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE* **2010**, *5*, 4248. [CrossRef]
7. Moyano, L.G.; Thomae, O.R.M.; Frias-Martinez, E. Uncovering the Spatio-temporal Structure of Social Networks Using Cell Phone Records. In Proceedings of the IEEE International Conference on Data Mining Workshops, Brussels, Belgium, 10 December 2012.
8. Siddall, W.R. Geography as Spatial Interaction. *Econ. Geogr.* **1981**, *57*, 270–271. [CrossRef]
9. Chen, B.; Wang, Y.; Wang, R.; Zhu, Z.; Ma, L.; Qiu, X.; Dai, W. The Gray-Box Based Modeling Approach Integrating Both Mechanism-Model and Data-Model: The Case of Atmospheric Contaminant Dispersion. *Symmetry* **2020**, *2*, 254. [CrossRef]
10. Noronha, V.T.; Goodchild, M.F. Modeling Interregional Interaction: Implications for Defining Functional Regions. *Ann. Assoc. Am. Geogr.* **1992**, *82*, 86–102. [CrossRef]
11. Kumar, K. The Information Age: Economy, Society and Culture. *Rise Netw. Soc.* **2009**, *1*, 132–134.
12. Onnela, J.P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; de Menezes, M.A.; Kaski, K.; Barabási, A.L.; Kertész, J. Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.* **2007**, *9*, 179. [CrossRef]
13. Palla, G.; Barabäsi, A.L.; Vicsek, T. Quantifying social group evolution. *Nature* **2007**, *446*, 664. [CrossRef] [PubMed]
14. Lambiotte, R.; Blondel, V.D.; de Kerchove, C.; Huens, E.; Prieur, C.; Smoreda, Z.; Dooren, P.V. Geographical dispersal of mobile communication networks. *Phys. Stat. Mech. Its Appl.* **2008**, *387*, 5317–5325. [CrossRef]
15. Gao, S.; Liu, Y.; Wang, Y.; Ma, X. Discovering Spatial Interaction Communities from Mobile Phone Data. *Trans. GIS* **2013**, *17*, 463–481. [CrossRef]
16. Chi, G.; Thill, J.C.; Tong, D.; Shi, L.; Liu, Y. Uncovering regional characteristics from mobile phone data: A network science approach. *Pap. Reg. Sci.* **2016**, *95*, 613–631. [CrossRef]
17. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
18. Perc, M. The Matthew effect in empirical data. *J. R. Soc. Interface* **2014**, *11*, 20140378. [CrossRef] [PubMed]
19. Snijders, T.A.; Nowicki, K. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *J. Classif.* **1997**, *14*, 75–100. [CrossRef]
20. Wang, Y.J.; Wong, G.Y. Stochastic Blockmodels for Directed Graphs. *J. Am. Stat. Assoc.* **1987**, *82*, 8–19. [CrossRef]
21. Holland, P.W.; Laskey, K.B.; Leinhardt, S. Stochastic blockmodels: First steps. *Soc. Netw.* **1983**, *5*, 109–137. [CrossRef]
22. Ghalebi, E.; Mirzasoleiman, B.; Grosu, R.; Leskovec, J. Dynamic Network Model from Partial Observations. *arXiv* **2018**, arXiv:1805.10616.
23. Newman, M.; Girvan, M. Find and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 292–313.
24. Montis, A.D.; Barthelemy, M.; Chessa, A.; Vespignani, A. The structure of Inter-Urban traffic: A weighted network analysis. *Physics* **2005**. [CrossRef]

25. Williamson, S.A. Nonparametric Network Models for Link Prediction. *J. Mach. Learn. Res.* **2016**, *17*, 1–21.
26. Sobolevsky, S.; Szell, M.; Campari, R.; Couronné, T.; Smoreda, Z.; Ratti, C. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE* **2013**. . [CrossRef] [PubMed]
27. Fortunato, S.; Lancichinetti, A. Community Detection Algorithms: A Comparative Analysis: Invited Presentation, Extended Abstract. In Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, Pisa, Italy, 20–22 October 2009; ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering): Brussels, Belgium, 2009.
28. Fortunato, S. Community Detection in Graphs. *arXiv* **2009**, arXiv:0906.0612.
29. Rosvall, M.; Bergstrom, C.T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1118–1123. [CrossRef]
30. Ai, C.; Chen, B.; He, L.; Lai, K.; Qiu, X. The national geographic characteristics of online public opinion propagation in China based on WeChat network. *GeoInformatica* **2018**, *22*, 311–334. [CrossRef]
31. Kemp, C.; Tenenbaum, J.B.; Griffiths, T.L.; Yamada, T.; Ueda, N. Learning systems of concepts with an infinite relational model. In Proceedings of the National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 July 2006.
32. Song, J.; Xu, K.; Song, M.; Zhan, X. Method for evaluating credibility of domestic IP address library. *Comput. Appl.* **2014**, *105*, 4–6.
33. Neal, R. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. Comput. Graph. Stat.* **2000**, *9*, 249–265.
34. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet Processes. *Publ. Am. Stat. Assoc.* **2006**, *101*, 1566–1581. [CrossRef]