Article

JIS

Journal of Information Science
2014, Vol. 40(1) 3–14

DOI: 10.1177/0165551513512251
jis.sagepub.com
SAGE

Static analysis and exponential random graph modelling for micro-blog network

Dong-Hui Yang
School of Management, Harbin Institute of Technology, People's Republic of China

Guang Yu
School of Management, Harbin Institute of Technology, People's Republic of China

Abstract

Social network analysis has been used to study complex networks by analysing their static structure and the dynamic changes. Although one of the newer forms of social media, micro-blogs have quickly become one of the most popular communication platforms. This popularity accounts for, in part, an increase in the scientific interest in micro-blogs and their users. In this paper, we chose as our test bed diabetes-related posts from the Chinese micro-blog Sina Weibo. We calculated the degree, average shortest path, betweenness and clustering coefficient of the Sian Weibo network to analyse its static structure. We demonstrate the characteristic results of average degree, diameter and clustering coefficient of diabetes micro-blog static structure. More importantly, we introduce a general model for micro-blog with directed network data, Exponential-family Random Graph Models (ERGMs). Meanwhile, we illustrate the utility for estimating, analysing and simulating micro-blog network. We also provide a goodness-of-fit approach to capture and reproduce the structure of the fitted micro-blog network. Parameter estimation of the model, similarity results of simulated networks and observed networks, and goodness of fit analysis for the micro-blog network all illustrate that ERGMs are excellent methods for deeply capturing complex network structures.

Keywords

Directed network; exponential-family random graph models; goodness-of-fit; micro-blog

1. Introduction

Social network theory provides explanations for myriad social phenomena, from individual creativity to corporate profitability [1]. Social networks describe the relationships between participating social actors. Facebook, Twitter, LinkedIn and MySpace are examples of social networks in which two actors are linked if there are interactions between them [2, 3]. Nodes and links are essential elements of social networks whether they are directed or undirected networks. Generally, there are two types of social network analysis: static and dynamic network analysis. Static analysis is used to discover the structural regularities of the nodes and links at the time. Dynamic analysis is used to find patterns of changes in the network over time [4].

Normally, in static structural analysis, several network properties should be included, such as degree centrality, betweenness, closeness and eigenvector centrality, to reflect the importance of actors and links in the network [5]. Most previous works focus on figuring out the key player or average shortest path to show the most important nodes and their distances in the network. In contrast, dynamic analysis aims to find the evolutionary process of a network structure. Therefore, how we describe, model and predict the dynamics is of vital importance. The previous descriptions of the changes in a network over time are relatively simple. Researchers use topological statistics such as the changes in average degree and clustering coefficient to express the changes. In real networks, many of them are scale-free topologies which show power-law distribution in degree and the preferential attachment mechanism [6]. Many works focus on

Corresponding author:
Dong-Hui Yang, School of Management, Harbin Institute of Technology, Harbin 150001, People's Republic of China.
Email: ydh95130@gmail.com

those problems. However, it is still much more challenging to more precisely and effectively model and predict the structural dynamics of a social network.

Micro-blogs are increasingly becoming critical platforms for individuals and organizations to seek and share real-time news updates. On the platform of micro-blogs, users are more active in renewing their messages in a short period of time. Twitter, with more than 140 million active users, is the best-known micro-blog worldwide [7]. Many works have shown how to use Twitter as a corpus for sentiment analysis [8–10]. However, it is blocked in China. As a substitute, Sina Weibo is a local micro-blog that has over 250 million users. Its users and the relationships among users change instantly. Moreover, according to the statistics of Hitwise, the utilization rate and user loyalty of Sina Weibo surpassed those of Twitter in April, 2011 [11]. Therefore, it would be interesting to understand the dynamic structure of its social networks.

Exponential-family random graph models (ERGMs), also known as p*class models, are a set of powerful tools to study complex network data. The importance of this framework lies in its capacity to represent social structural effects commonly observed in many human social networks [12]. They are statistical models to estimate the effects of covariates and simulate common features in social networks. An ERGM is used to address the complex dependencies within relational data structures and provide a flexible framework to represent them. It is a powerful tool for formulating theoretical models and learning properties of empirical networks [13]. Because of those characteristics, ERGM is a suitable approach to analyse, simulate and visualize the micro-blogging data. We could use ERGM to find out what kind of dependence assumptions and parameters are good in the network of micro-blog.

The remainder of this paper is organized as follows. We first discuss the Sina Weibo data on the topic of diabetes in Section 2 and present the methods containing both static network analysis and exponential random graph models in Section 3. Section 4 introduces the experiments and results of static analysis, estimation, simulation and goodness of fit by using ERGM. Finally, we outline the conclusions of this work in Section 5.

2. Related work

Social behaviour is defined as activities among members of social groups. In real world, social behaviour produces complicated networks. Also, in social media, people connect to each other online and build complex networks. To understand network evolution or structures, models can be of great value in achieving efficient representation. Many models have been proposed that are useful tools for assumption and simulation. What we need is to estimate model parameters from data and evaluate how adequately the model represents the network. An exponential-family random graph model can achieve this aim. The ERGM simultaneously allows arbitrarily complex network structures to be modelled.

According to different dependence assumptions, there are diverse expressions of exponential-family random graph models. At an earlier stage, Bernoulli random graph distributions were generated when they assumed that edges were independent. Dyadic models were built for directed networks which dyads were assumed to be independent of each other. A much more realistic assumption empirically was Markov random graphs, in which two possible network ties that had a common actor were conditionally dependent. Based on realization-dependence structures, Snijders et al. developed new specifications for exponential random graph models that included new higher-order terms [14]. They introduced the constraints on *k*-star parameters, *k*-triangle configurations and higher-order star and triangle effects. Goodreau and Robins continued this idea and obtained improved model performance on both convergent parameter estimates and goodness of fit [15, 16]. Hunter used ERGM to model high school friendship networks of varying sizes and found that an improved fit appeared when new parameters were included [17].

Moreover, Marijtje et al. compared the bias, standard errors, coverage rates and efficiency of maximum likelihood and maximum pseudo-likelihood estimators [18]. Meanwhile, they proposed an improved pseudo-likelihood estimation method to reduce bias. Saul and Filkov used ERGM to explore biological network structure and found the model could best be achieved by using pseudo-likelihood maximization [19]. However, the properties of a pseudo-likelihood estimator were not well understood and the estimates were not accurate for many datasets. Later, Monte Carlo maximum likelihood estimation techniques for EGRM were presented [20–22]. It was found that the preferred option was to use Monte Carlo estimation procedures from their research works. Martina Morris described means to control the Markov chain Monte Carlo (MCMC) algorithm that the package was used for estimation [23].

In recent years, ERGM has been widely used in many fields to predict real-life networks. Goodreau applied ERGM to adolescent friendship networks in 59 US schools from the National Longitudinal Survey of Adolescent Health by operating on individual, dyadic and triadic levels [24]. Robins studied the closure, connectivity and degree distributions of directed organizational network data using ERGM [25]. Cranmer used ERGM to discover unexplored parameters for prediction and found structural characters on political networks: cosponsorship networks in the US Congress and conflict networks in the international system [26]. In 2011, Simpson et al. illustrated the utility of ERGMs for modelling,

analysing and simulating complex whole-brain networks. They also proposed a graphical goodness of fit approach to capture and reproduce the structure of fitted brain networks [27]. Traud used ERGM to study the dynamic processes on the 'friendship' network of Facebook at 100 American colleges and universities [3]. Ouzienko and Krivitsky expanded ERGM into temporal social networks and valued networks for modelling and simulating respectively [28, 29]. Pallotti et al. carried out their research with ERGM on inter-organizational networks, communities and fields [30]. Shalizi and Rinaldo focused on the popular class of ERGMs and showed their consistency under sampling [31]. However, there is little research using ERGM to study the structures of Chinese micro-blog. Because of that research gap, we have studied this problem by applying ERGM to the network of the Chinese micro-blog.

3. Methods

3.1. Static and dynamic network analysis

Static structure analysis is performed to find critical nodes and links on a snapshot of a network. It extracts topological properties from social networks. The key users, relationships and communication links are much more important in a network. There are four metrics to measure the network properties of nodes and linked paths: degree, average shortest path, betweenness and clustering coefficient [32]. As graph theory is the mathematical foundation for network analysis, we will introduce definitions of graph theory first. A network is denoted as $G = (V, E)$ in graph theory, where V is the set of vertices (or nodes) of the graph G and E comprises two-element subsets of V, referred to as edges (or links or connections).

Degree centrality is a method for measuring the importance of a node by calculating how many links it has with other nodes. The degree for a node k is its number of neighbours. In a directed network, it can be classified into input degree and output degree. The average shortest path is the average of the smallest distance between pairs of nodes, while the distance between two nodes is defined as the length of a geodesic between them [33].

Betweenness is a measurement of how many shortest paths are going through a given node. It is a node influence on the spread of information through the network. The higher the betweenness of a node between many nodes through their shortest paths is, the greater its influence as it flows in the network. For example, the betweenness of a node k ($k \neq i \neq j$) is calculated as:

$$B = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}} \tag{1}$$

where g_{ij} is the number of geodesic paths from i to j and g_{ikj} is the number of these geodesics that pass through k. Betweenness centrality is the proportion of all geodesics between pairs of other nodes that include this node.

The clustering coefficient is defined as the probability that a node's neighbours are all connected with each other. It is used to measure the strength of sub-group formation and the density of the network. For an undirected network, it can be expressed as:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{2}$$

where k_i is the degree of node i and E_i is the total number of links among node i's neighbours.

3.2. Exponential random graph models

For better understanding the network structure, we need to know the dynamic changes of network. ERGM helps to reveal the underlying factors or variables that explain the dynamic of network formation over time [22]. The general form of exponential random graph models is as follows:

$$\Pr(Y = y) = \left(\frac{1}{k}\right) \exp\left\{\sum\nolimits_A \eta_A g_A(y)\right\} \tag{3}$$

where

(1) k is a normalizing quantity to ensure the equation is a proper probability distribution;
(2) η_A is the parameter corresponding to configuration of type A;

(3) $g_A(y)$ is the network statistic counting the frequency of sub-graph A in the graph $\mathbf{y}$; $g_A(y) = 1$ if the configuration is observed in the network $\mathbf{y}$, and is 0 otherwise;
(4) $\Sigma_A \eta_A g_A(y)$ is over all configurations types A.

As mentioned above, ERGM has been used to study many real-life networks. It is shown that ERGM is a useful tool to discover network structures. Among previous works, ERGM is operated through three steps: model estimation, model-based network simulation and model evaluation. It not only proposes a dependence assumption of the model but also estimates the parameters and finds a good fit for the observed model. Moreover, the last step can be used to predict the dynamic structure of social networks. In the following section, we will illustrate those three steps in order.

4. Experiments and results

4.1. Data bed

Sina Weibo, with more than 250 million users, is the first and the biggest micro-blog website in China. On this platform, users share their information and opinions on diverse topics. Meanwhile, personnel in specific fields or companies open accounts to provide services through the micro-blog. For example, doctors open accounts to serve patients. Among those patients, a large proportion are suffering from diabetes. In China, there are currently 40 million people with diabetes that need great awareness and basic education to improve healthcare services. Healthcare is a prospective and useful area to provide convenient service through social media. Fortunately, more and more doctors and hospitals are opening accounts to help patients by posting new information and correct treatments. We can obtain this data from Sina API. Therefore, we searched diabetes as our topic and chose 50 users whose followers were more numerous than others in this field, including diabetes hospital accounts, validated accounts of well-known diabetes doctors and validated accounts of diabetes magazines, as our research seeds. The network among them and their followers is big enough for us to perform the static and dynamic structure analysis. Account information for the 50 seed users until 30 April 2012 is listed in Table 1. Their network structure plotted in R can be seen in Figure 1.

Table 1. Accounts information for 50 users.

Information	Number of followers	Number of fans	Number of micro-blogs	Average number of fans
Number	27,872	444,358	48,594	8887

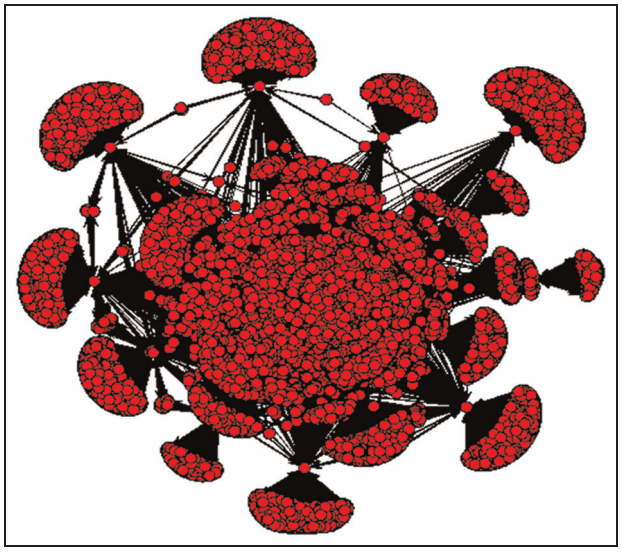

Figure 1. The network structure of 50 micro-blog diabetes users.

Table 2. The metrics of the diabetes network structure.

Metrics	Values
Number of vertices	69,547
Number of arcs	88,537
Average degree	2.546
Average shortest path	3.926
Diameter	7
Betweenness centralization	0.137
Clustering coefficient	0.03

4.2. Statical analysis

In order to extract topological properties of the network from the micro-blogs, we use four metrics to measure the static structure: degree, average shortest path, betweenness and clustering coefficient. As we discussed in Section 3.1, these metrics are used to find the key nodes, relationships and communication links. We used the PAJEK tool, which is professional software for performing network analysis, to calculate the values of metrics as shown in Table 2.

We abstracted the vertices and arcs from our database culled from micro-blogs. It consisted of 69,547 users and 88,537 relationships in total. We used partitions (*Net* > *Partitions* > *Degree* > *All*) to produce a list of degree distribution. Because it is not convenient for display, we used average degree as a substitution. The average degree of our network was 2.55, which means users on average had 2.5 friends in the network of the micro-blog. The highest value of degree was 5193, and the lowest value was 0, whose frequency was 4. The highest frequency was 60,764 and its cluster value was 1 (Figure 1). In other words, there were many nodes linking to the same user who was a very famous doctor or hospital.

To determine the characters of links, the length of average shortest path and diameter of the network were calculated. When using PAJEK to calculate them, we used *Net* > *Paths between 2 vertices* > *All shortest/Diameter* commands to obtain the average geodesics between each two individuals and the diameter in the network. The average length of the shortest path was 3.93, so users had almost 4 links or steps in the shortest path to connect other vertices. Among the shortest paths, 7 was the largest length between two vertices, which is called diameter in the network.

The more times a node is a go-between, the more central is its position in the network. Betweenness centralization is one type of metric to embody the centrality of the network. In our experiment, network betweenness centralization (*Net* > *Vector* > *Centrality* > *Betweenness*) was 0.137, which means the proportion of go-between nodes was much smaller compared with the maximum variation in the whole network.

The clustering coefficient is often used to compute the egocentric density of all vertices in undirected networks. If the directed network does not contain loops or bidirectional arcs, we can use the clustering coefficient to measure the strength of sub-group formation and the density of the network. In the network we collected, there were no loops or bidirectional arcs. Therefore, we computed the clustering coefficient of our network by commanding *Net* > *Vector* > *Clustering coefficients* > *CC1*. We found that the density of the network was very low, with a clustering coefficient value of 0.03, that is to say, there were many chances to link to other users.

4.3. Dynamic analysis using ERGMs

In our experiment, we collected the micro-blog data on the topic of diabetes from August 2009, the time that Sina Weibo was released to public use, to April 2012. To discover the change in the network over time, we extracted the data annually and accumulated new data into the original dataset. Therefore, we gained four datasets: those for 2009, 2009 and 2010, from 2009 to 2011, and from 2009 to 2012. We used ERGM to model each network and compared their changes and development tendencies.

Because of the memory limitation in R tool, we adjusted the data size of network by restricting the number of users' posted micro-blogs, followers and fans before we used ERGM for modelling. The dataset from 2009 (dataset 1) and the appropriate datasets from 2009 and 2010 (dataset 2), from 2009 to 2011 (dataset 3), and from 2009 to 2012 (dataset 4) are plotted in Figure 2(a–d), respectively. The *statnet* suite of packages for R contains the *ergm* package (http://statnet project.org/). In detail, we used '*ergm*' to fit an ERGM, '*simulate*' to simulate networks from a fitted ERGM and '*gof*' to assess goodness of fit for an ERGM.

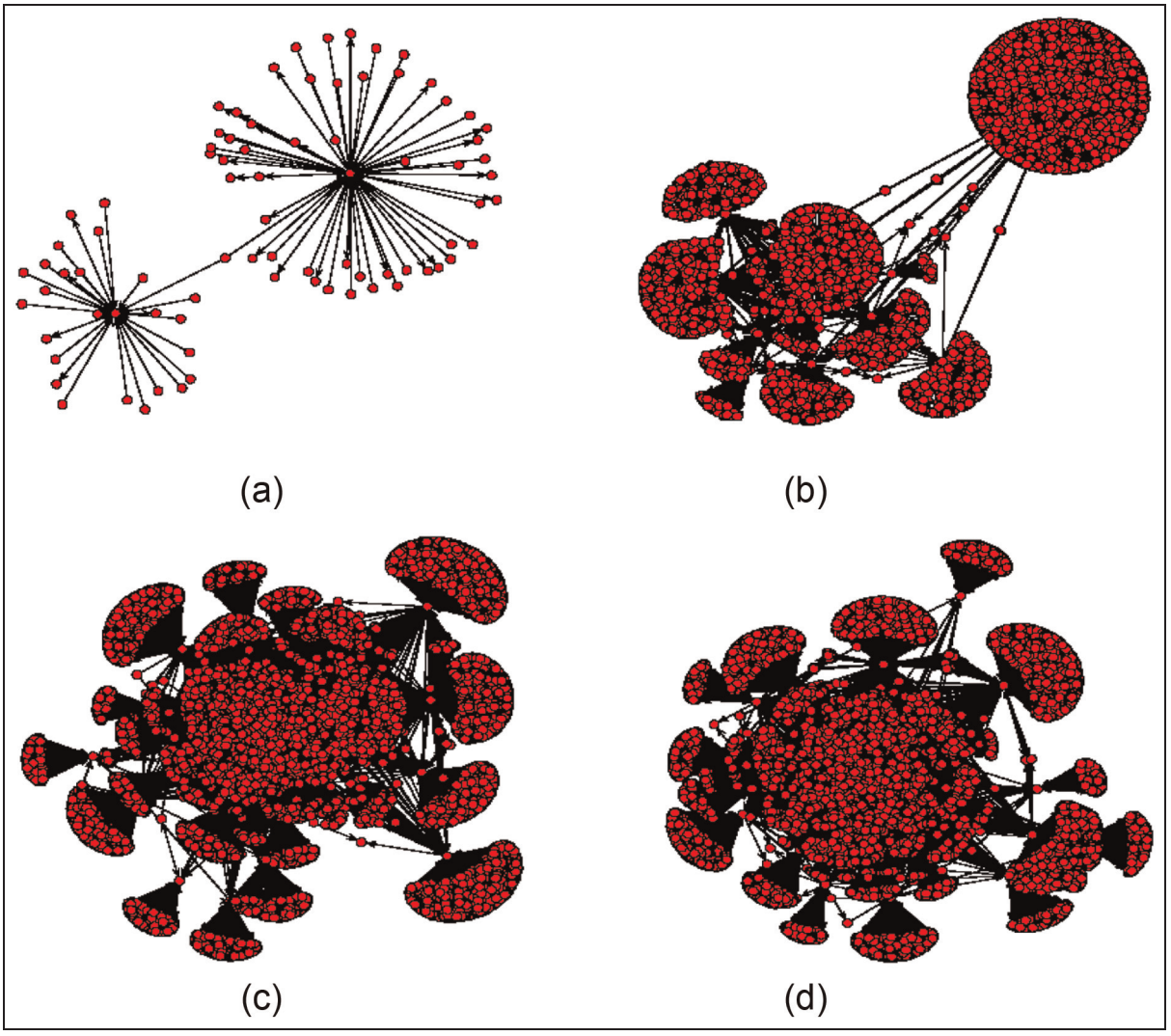

Figure 2. The network structures of four datasets.

4.3.1. Estimation. In a given network, the sub-graphs are the network configurations. For instance, an ERGM with just the Edges metric shows the connectedness of the network. Geometrically weighted edge-wise shared partner (GWESP), Mutual and Triangle metrics are used to measure local clustering efficiency, mutual ties effect and transitivity respectively. Therefore, we used those metrics to uncover the sub-graph of above four networks in our experiment. Those metrics lead to an overall ERGM for micro-blog diabetes network. That is,

$$P(Y=y)=\frac{1}{k(\theta)}\exp\{\theta_1 Edges(y)+\theta_2 Mutual(y)+\theta_3 GWESP(y)+\theta_4 Triangle(y)\} \quad (4)$$

We used this model to fit an ERGM and obtain estimations for θ_1,θ_2,θ_3 and θ_4. If its value for a given metric is positive and large, then this metric is more prevalent than in the null model and plays a considerable role in explaining the network structure. Conversely, if the estimated value for a given metric is negative and large, this metric also plays a considerable role in explaining the network structure but is less prevalent than in the null model, while the standard error, Markov chain Monte Carlo standard error and p-value are used for metric selection. The significant p-value (*) illustrates that the estimated value can be adopted as an appropriate parameter for this model. The Monte Carlo maximum likelihood estimation (Monte Carlo MLE) results of four micro-blog datasets are shown in Table 3.

From Table 3, we can see the fit results and several phenomena. First, the Edges metric with a negative and large value plays an important role but is not prevalent in all four models. The Mutual metric with positive and large values,

Table 3. Monte Carlo MLE results of four micro-blog datasets.

Metrics		Estimate	Standard error	MCMC standard error	*p*-Value
Model 1	Edges	−11.0790	0.1124	NA	$< 1 \times 10^{-4}$***
	Mutual	7.6904	1.3116	NA	$< 1 \times 10^{-4}$***
	GWESP	8.6387	0.0000	NA	$< 1 \times 10^{-4}$***
	Triangle	NA	NA	NA	NA
Model 2	Edges	−10.8847	0.0188	0	$< 1 \times 10^{-4}$***
	Mutual	14.3876	6.9038	0	0.03716 *
	GWESP	11.3540	4.0349	0	0.00489 **
	Triangle	2.3780	1.28050	0	0.06330
Model 3	Edges	−9.6180	0.0080	1	$< 1 \times 10^{-4}$***
	Mutual	6.7665	0.5303	5	$< 1 \times 10^{-4}$***
	GWESP	0.8035	1.0027	4	0.423
	Triangle	0.1689	0.5899	4	0.775
Model 4	Edges	−9.6691	0.1215	0.017	$< 1 \times 10^{-4}$***
	Mutual	7.2210	3.1768	21.047	0.023*
	GWESP	0.8178	0.0259	0.054	$< 1 \times 10^{-4}$***
	Triangle	0.0720	0.0028	0.001	$< 1 \times 10^{-4}$***

Significance codes: *** $p < 0.001$; ** $0.001 < p < 0.01$; * $0.01 < p < 0.05$.

Table 4. Structure comparison between simulated nets and datasets.

Network	Edges	Mutual	In-0	In-1	In-2	In-3	Out-1	Out-2	Out-3	Triangle	Cosine
Dataset 1	81	1	2,235	27	0	0	52	1	0	0	99.95%
Net 1	87	1	2,178	85	1	0	83	2	0	0	
Dataset 2	4,412	38	14,610	277	18	2	3,659	144	27	71	98.68%
Net 2	4,437	38	12,245	2,283	264	13	3,180	453	44	125	
Dataset3	18,078	789	12,636	1,959	178	62	9,839	1,088	338	9,031	98.69%
Net 3	18,541	789	10,156	3,798	733	142	7,219	2,380	565	8,601	
Dataset 4	18,682	851	12,845	2,083	182	63	10,063	1,131	357	9,397	99.98%
Net 4	18,713	851	12,564	2,307	233	64	9,676	1,348	366	9,319	

especially in model 2, plays a considerable role and is prevalent in the network structure. Meanwhile, its estimated value is significant in every model. Second, the geometrically weighted edge-wise shared partner metric (the τ parameter associated with GWESP is set to 0.5 as this value generally led to better fitting model) is positive and large only in models 1 and 2, and the estimated parameter is significant. Although the estimated parameter of GWESP metric in dataset 4 is significant, the value is small, which means that GWESP is not prevalent in this model. Since the *p*-value of GWESP in model 3 is not significant, it cannot be adopted in the model. Third, there is no value of the Triangle metric in model 1 because it has no triangle structure in this network, which can be clearly seen in Figure 2(a). The *p*-values of the Triangle metric are not significant in models 2 and 3. This metric is significant in model 4, but the estimated value is too small to play an important role in the model.

4.3.2. Simulation. For a quantitative comparison of structural similarities in the generated network, we used the 'simulate' command to generate four simulated networks using ERGM based on MCMC idea. We compared the original datasets with simulated networks referring to 10 statistics: edges, mutual ties, in-degree (0–3), out-degree (1–3) and triangle. Because much higher values were all equal to 0, we ignored much higher in- and out-degrees. The statistics in Table 4 show which simulated network is more similar to the original dataset. If we contrast the 10 statistics one by one, we find that big gaps mostly exist in the degree statistics, including in-degree 1, in-degree 2, out-degree 1 and out-degree 2. However, this is not a good way to check the comparison results.

We may consider the 10 statistics of dataset and simulated network as two vectors. Then we can use the cosine distance to measure their similarities. We can see that the fourth simulated network is much more similar to its dataset than the other three simulated networks in Table 4. Since the distances of in-degree 1, in-degree 2, out-degree 1 and

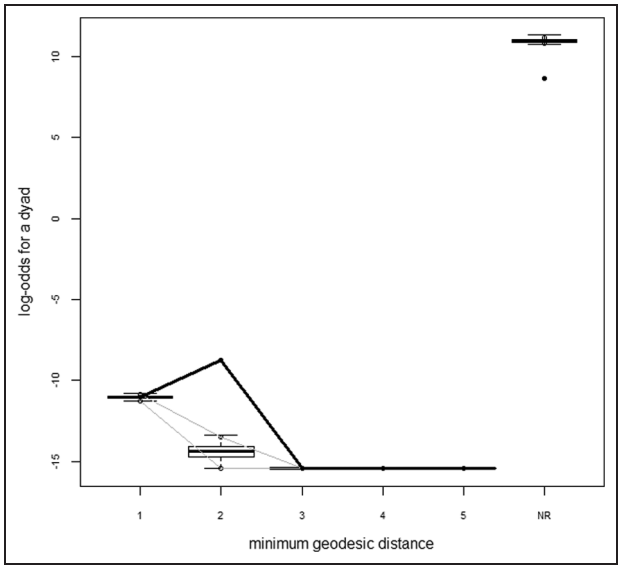

Figure 3. Goodness of fit for geodesic distance of model 1.

out-degree 2 between datasets 2 and 3 and their simulated networks are large, the cosine distances are just 98.68 and 98.69%, respectively. Although the second and third cosine similarities are lower than the first and fourth ones, their values are greater than 98%, which is a high similarity level. According to the comparison of 10 parameters and cosine similarities, we find that the ERGM used to generate simulated networks has a good performance for dynamic micro-blog datasets.

4.3.3. Goodness of fit. Comparing just 10 parameters of the simulated network with the original one is of limited value. To compare the full distribution of our statistics of interest, we used the '*gof*' command to visualize some common network distributions in the goodness-of-it automatically. Three metrics were adopted to plot their distributions in our work: the geodesic distribution (the number of actor pairs for which the shortest path between them is of length k, for each value of k), the distribution of edgewise shared partners (the number of edges in which two friends have exactly k friends in common, for each value of k), and the triad census distribution (the proportion of three-node sets having no, one, two or three edges among them. For a directed network, the triad census has 16 categories instead of four).

Because of the model degeneracy issue, we cannot obtain the last two distributions for model 1 but only a geodesic distribution as shown in Figure 3. Problems with model degeneracy are common when parameter values imply that only one or two graphs have substantial non-zero probabilities [20]. Therefore, we give goodness-of-fit diagnostics for models 2–4 in Figures 4–6. In these figures, the vertical axis is the logit of relative frequency, the box-plots summarize the statistics for the simulated networks resulting from the Monte Carlo maximum likelihood estimation, and the solid line in each plot represents the statistics of the observed networks.

Distance, as a global property of the network, can be used to measure how well the observed and simulated distributions match. When examining Figures 3, 5 and 6, we can see that models 1, 3 and 4 do a poor job of capturing geodesic distance distribution. The upper plot of Figure 4 reveals that ERGM does better than the others of producing network to reflect geodesic distance distribution. This means that the observed proportion of pairs of nodes with shortest connecting path length from 1 to 15 is much more similar to the simulated one for model 2.

For local efficiency, both models 2 and 3 do a good job of producing networks that reflect the edgewise shared partner of datasets 2 and 3, respectively. Therefore, we learn that edges between two nodes that share exactly i neighbours are common in models 2 and 3 which we cannot clearly see from Figure 2(b) and (c). Additionally, model 4 does not capture very well the edgewise shared partner distribution. We can see that the observed curve is very close to the simulated one in the middle plot of Figure 6.

That situation also occurs for triad census distribution analysis when comparing motifs distribution to the observed ones of models 3 and 4. However, model 2 performs much better than models 3 and 4 in capturing the triad census distribution of the micro-blog network. That is to say, the observed proportions of three-node sets, which actually have 16 categories in the directed network among models 3 and 4, are much higher than in the simulated ones. However, the simulated proportion of three-node sets is close to the observed one for model 2.

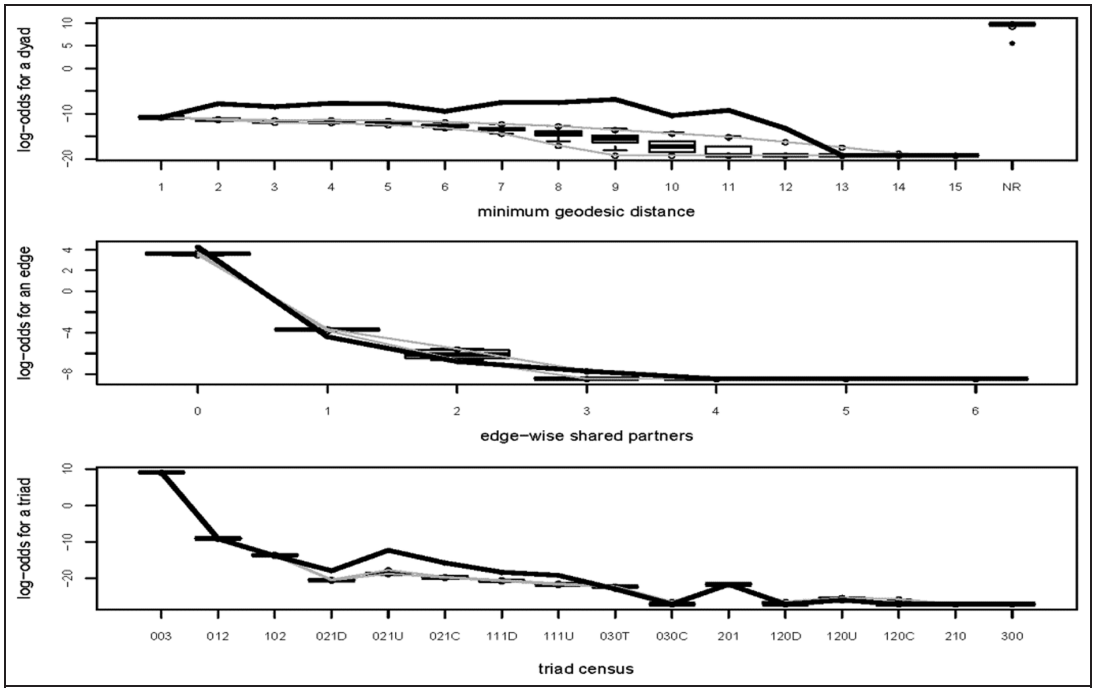

Figure 4. Goodness of fit for geodesic distance, edgewise shared partner and triad census of model 2.

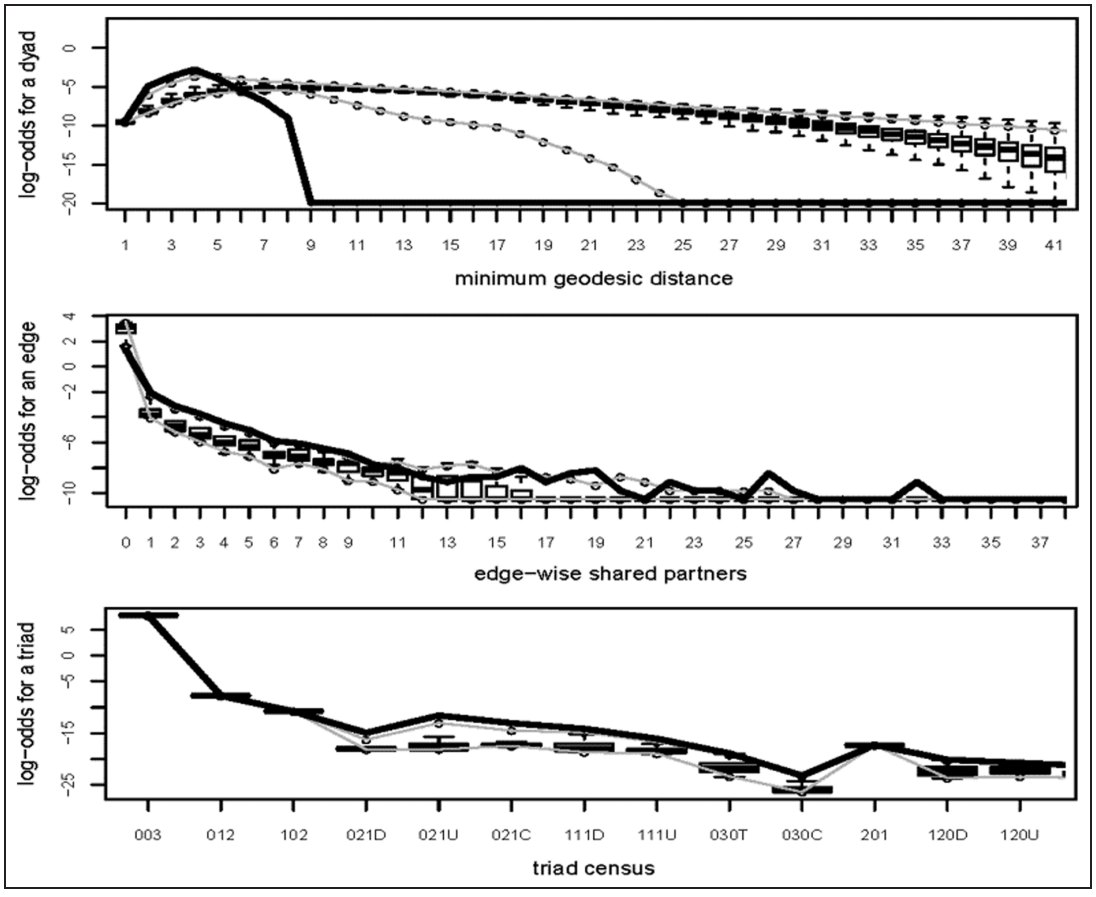

Figure 5. Goodness of fit for geodesic distance, edgewise shared partner and triad census of model 3.

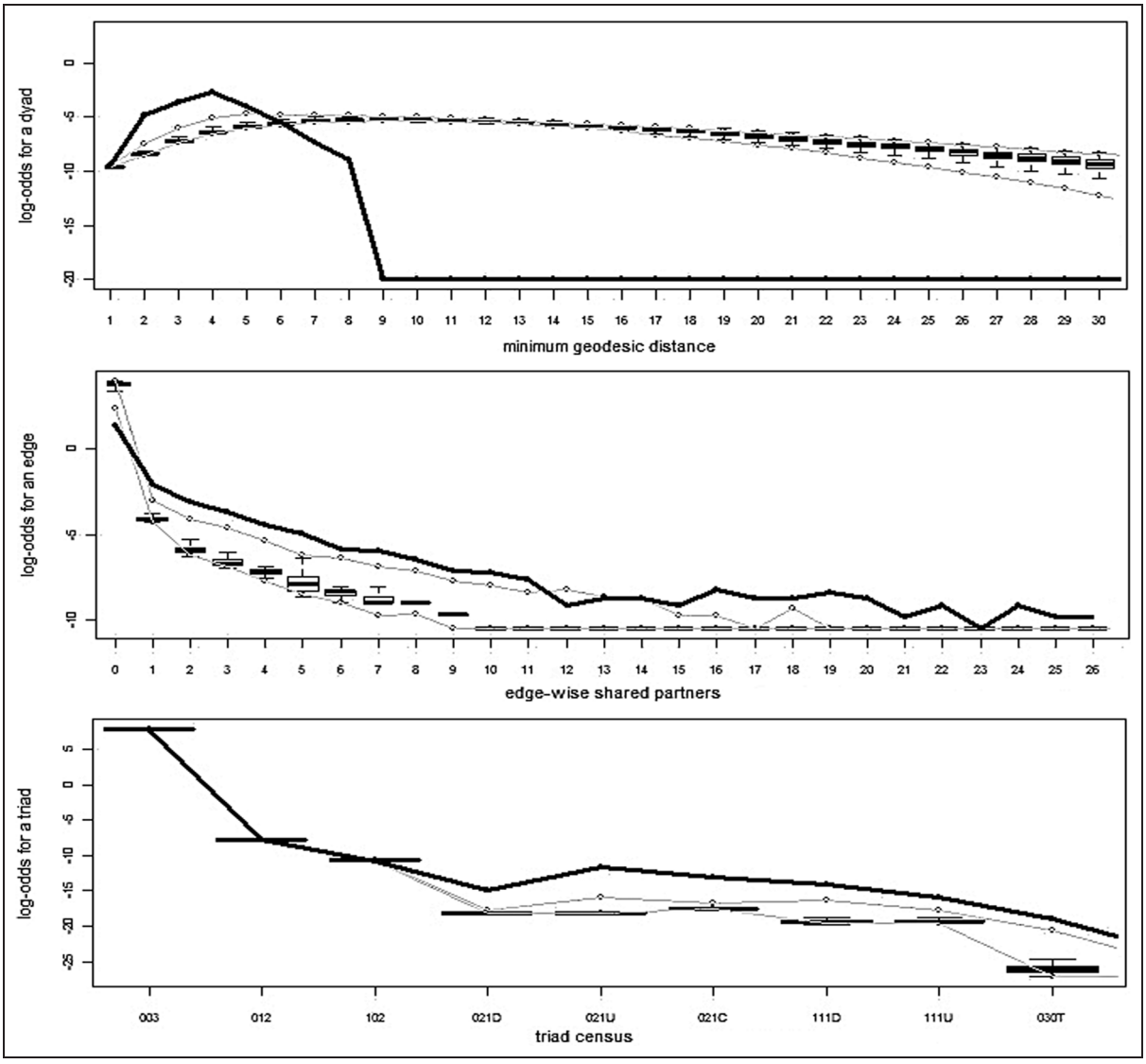

Figure 6. Goodness of fit for geodesic distance, edgewise shared partner and triad census of model 4.

5. Conclusions

Our analyses illustrate the diabetes network structure of Chinese micro-blog both in a static and a dynamic way. Usually, four common metrics are chosen to study network static structure: degree, average shortest path, betweenness and clustering coefficient. According to the values of those metrics, we found the key node (famous doctor or hospital), relationship (go-between users) and communication links (density of the network) of diabetes micro-blog users in our work. However, we cannot obtain more detailed information and characteristics of diabetes micro-blog network if we just analyse the static structure. Therefore, we need to analyse the changes in micro-blog networks to discover a network deeply based on scientific interest.

The most important contribution of this paper is focusing on dynamic structure analysis. Exponential-family random graph models are adopted for modelling, estimating and simulating micro-blog networks. To discover the changes in the networks over time, we extracted the data annually and accumulated new data into original dataset from August 2009 to April 2012. Then we used ERGM to model each network and compared their structure changes. It can be clearly seen how the network expanded from a small network to a complicated network in Figure 2. We used edges, GWESP, mutual and triangle metrics to measure global efficiency, local clustering/efficiency, mutual ties effect and transitivity in our ERGMs. In order to check the model parameters, we chose Monte Carlo maximum likelihood estimation to find out

which estimated parameter is significant in each model. In our experiment, the mutual tie metric is significant in each model but the triangle metric is only significant in the fourth dataset. We drew the conclusion that the transitivity evidently shows up when the micro-blog network becomes big enough.

Also, four simulated networks using ERGM based on Markov Chain Monte Carlo were generated for comparison with the original datasets. We analysed the gaps of 10 statistics between simulated network and the observed dataset. More meaningfully, we contrasted their cosine similarities and found that the last group had the highest similarity value (99.98%), and the second group the lowest value (98.68%). All four cosine similarities were very high. This means that simulated networks generated by ERGM have good performance for dynamic micro-blog datasets.

Moreover, the goodness-of-fit approach gave us the scientific interest to capture and reproduce the structure of the fitted network. We represented the complex network data using ERGM and examined the simulated network's distances and local structural components. For those models we examined the geodesic distribution, edgewise shared partners distribution and triad census distribution. Goodness-of-fit simulations suggest that models 2 and 3 are well behaved in reflecting the edgewise shared partner of datasets 2 and 3. Model 2 performs much better than models 3 and 4 when capturing the triad census distribution of the micro-blog network.

Acknowledgements

Many thanks are due to Fred Niegocki, Joshua Chuang and Cathy Larson for their assistance in language revision. Also, thanks to the editor and all anonymous reviewers for their constructive comments.

Funding

This work is supported by the National Natural Science Foundation of China (grant no. 71171068 and no. 71003020). This research is also supported by the China Scholarships Council (file no. 2011612202).

References

[1] Borgatti SP, Mehra A, et al. Network analysis in the social sciences. *Science* 2009; 13: 892–895.
[2] Ouzienko V, Guo Y and Obradovic Z. A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks. *Statistical Analysis and Data Mining* 2011; 4(5): 470–486.
[3] Traud AL, Mucha PJ and Porter MA. Social structure of Facebook networks. *Physica A* 2012; 391: 4165–4180.
[4] Agarwal N, Galan M, Liu H and Subramanya S. Clustering of blog sites using collective wisdom. *Computational Social Network Analysis* 2010; 1: 107–134.
[5] Alguliev R, Aliguliyev R and Ganjaliyev F. Investigation of the role of similarity measure and ranking algorithm in mining social networks. *Journal of Information Science* 2011; 37(3): 229–234.
[6] Brot H, Muchnik L, Goldenberg J and Louzoun Y. Feedback between node and network dynamics can produce real-world network properties. *Physica A* 2012; 391: 6645–6654.
[7] Larsson AO and Moe H. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society* 2012; 14(5): 729–747.
[8] Pak A and Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation, LREC'10*, Valletta, 2010, pp. 1320–1326.
[9] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data. In *Proceedings of the ACL 2011 workshop on languages in social media*, 2011, pp. 30–38.
[10] Thelwall M, Buckley K and Paltoglou G. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 2011; 62(2): 406–418.
[11] Read P, Shah C, S-O'Brien L and Woolcott J. 'Story of one's life and a tree of friends' – understanding millennials' information behaviour in social networks. *Journal of Information Science* 2012; 38(5): 489–497.
[12] Robins G, Pattison P, Kalish Y and Lusher D. An introduction to exponential random graph (p*) models for social networks. *Social Networks* 2007; 29: 173–191.
[13] Desmarais BA and Cranmer SJ. Statistical mechanics of networks: Estimation and uncertainty. *Physica A* 2012; 391: 1865–1876.
[14] Tom AB, Snijders PE, et al. New specifications for exponential random graph models. *Sociological Methodology* 2006; 36(1): 36–99.
[15] Goodreau SM. Advances in exponential random graph (p*) models applied to a large social network. *Social Networks* 2007; 29: 231–248.
[16] Robins G, Snijders T, Wang P, Handcock M and Pattison P. Recent developments in exponential random graph (p*) models for social networks. *Social Networks* 2007; 29: 192–215.

[17] Hunter DR, Handcock MS and Butts CT. ERGM: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software* 2008; 24(3): 1–29.
[18] Marijtje AJ, Gile KJ and Handcock MS. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks* 2009; 31(1): 52–62.
[19] Saul ZM and Filkov V. Exploring biological network structure using exponential random graph models. *Bioinformatics* 2007; 23(19): 2604–2611.
[20] Snijders TB. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* 2002; 3(2): 1–40.
[21] Handcock MS, Hunter DR, Butts CT, Goodreau SM and Morris M. Statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software* 2008; 24(1): 1548.
[22] Robins G and Morris M. Advances in exponential random graph (p*) models. *Social Networks* 2007; 29(2): 169–172.
[23] Morris M, Handcock MS and Hunter DR. Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software* 2008; 24(4): 1548–7660.
[24] Goodreau SM, Kitts JA and Morris M. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* 2009; 46(1): 103–125.
[25] Robins G, Pattison P and Wang P. Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks* 2009; 31: 105–117.
[26] Cranmer SJ and Desmarais BA. Inferential network analysis with exponential random graph models. *Political Analysis* 2011; 19: 66–86.
[27] Simpson SL, Hayasaka S and Laurienti PJ. Exponential random graph modeling for complex brain networks. *PLoS One* 2011; 6(5): e20039.
[28] Ouzienko V, Guo Y and Obradovic Z. A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks. *Statistical Analysis and Data Mining* 2011; 4(5): 470–486.
[29] Krivitsky PN. Exponential-family random graph models for valued networks. *Electronic Journal of Statistics* 2012; 6: 1100–1128.
[30] Pallotti F, Lomi A and Mascia D. From network ties to network structures: Exponential random graph models of interorganizational relations. *Quality & Quantity* 2013; 47(3): 1665–1685.
[31] Shalizi CR and Rinaldo A. Consistency under sampling of exponential random graph models. *The Annals of Statistics* 2013; 41(2): 508–535.
[32] Ahram TZ and Karwowski W. Visual social network analysis: Effective approach to model complex human social, behaviour & culture. *Work – A Journal of Prevention Assessment and Rehabilitation* 2012; 41: 3504–3510.
[33] Hua G, Sun Y and Haughton D. Network analysis of US air transportation network. *Data Mining for Social Network Data* 2010; 12: 75–89.