Exponential Random Graph Modeling for Micro-blog Network Analysis

Dong-Hui Yang, Guang Yu

School of Management, Harbin Institute of Technology, China, 150001

Ydh95130@gmail.com

ABSTRACT

Social network analysis is used to study complex networks by analyzing static structure and dynamic changes. Nowadays micro-blog as a new social media is becoming the most popular communication platform. How to capture micro-blog network structure especially dynamic structure poses more scientific interest. In this paper, we choose Chinese micro-blog, Sina weibo, on topic of diabetes as our test bed. We calculate degree, average shortest path, betweenness and clustering coefficient to analyze its static structure. More important works, we introduce a general model for micro-blog with directed network data, Exponential-family Random Graph Models (ERGMs), and illustrate the utility for modeling, analyzing and simulating micro-blog network. We also provide the goodness-of-fit approach to capture and reproduce the structure of the fitted micro-blog network. We demonstrate the characteristic results of average degree, diameter and clustering coefficient of diabetes micro-blog static structure. Parameters estimation of model, similarity results of simulated networks and observed networks, and goodness of fit analysis for micro-blog network are all illustrated that ERGMs are excellent methods to deeply capture the complex network structure.

KEYWORDS

Exponential random graph models, Micro-blog, Directed network, Network analysis, Goodness-of-fit

1INTRODUCTION

Social networks describe the relationships between participating social actors. Facebook, Twitter, LinkedIn and MySpace are examples of social networks that two actors are linked if there are interactions between them [1]. Actors and links are essential elements of social networks whether they are directed or undirected networks. There are two types of social networks: static and dynamic networks. Static analysis is to discover the structural regularities of the actors and links at the time. Dynamic analysis is to find the patterns of changes in the network over time [2].

As usual static structural analysis, several network properties should be given such as degree centrality, betweenness, closeness and eigenvector centrality to reflect the importance of actors and links in the network [3]. Most previous works are focus on finding out the key player and average shortest path to show the most important actors and the network distance. However, dynamic analysis is aimed to find the evolutionary process of network structure. Therefore, how to describe, model and predict the dynamics is of vital importance. Previous descriptions of the changes in a network over time are relatively simple. Researchers use topological statistics to embody, such as the changes of average degree and clustering coefficient. In real networks, many of them are scale free topologies which show the power-law distribution in degree and preferential attachment mechanism. However, how to model and predict the structural dynamics of social network is still much more challenging.

Nowadays, micro-blog is increasingly becoming a critical platform for individuals and organizations to seek and share real-time news updates. On the platform of micro-blogs, users are more active and renew their messages very soon even in a short time. Twitter, with more than 140m active users, is the best-known micro-blog in the world [4]. Many works have shown how to use Twitter as a corpus for sentiment analysis [5-7]. But it is blocked in China. As a substitute, Sina Weibo is a local micro-blog that has over 250m users. Its users and the relationship between users are changed quickly. According to the statistic of Hitwise, utilization rate and user stickiness of Sina Weibo have surpassed Twitter by April, 2011. Therefore, how to model this social network and dynamic structure are much more valuable.

Exponential-family random graph models (ERGM), also known as p*class of models have been utilized to analyze complex network data. The importance of this modeling framework lies in its capacity to represent social structural effects commonly observed in many human social networks [8]. It is a statistical model to estimate the effects of covariates and simulate features common in social networks. ERGMs were used to address the complex dependencies within relational data structures and provide a flexible framework for representing them. It can produce useful models with distinctive interpretations. ERGM categories into three steps: model estimation, model evaluation and model-based network simulation. It not only proposes dependence assumption of the model but also estimates the parameters and finds a good fit to the model. Moreover, the last step can be used to predict the dynamic structure of social networks. So, ERGM is a good approach to analysis, simulate and visualize the Micro-blogging data. We could use ERGM to find out what kind of dependence assumption and parameters are good in micro-blog networks.

2 RELATED WORKS

2.1 Static and Dynamic Networks

Static structure analysis is to find critical nodes and links on the snapshot of a network. It tries to extract topological properties from networks. Therefore, the key users, relationships and communication links are more critical 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 [9]. 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 are 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 is to calculate how many links it has with other nodes. The degree for a node k is the number of its neighbors. 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 [10].

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

$$\sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

Where, g_{ij} is the number of geodesic paths from *i* to *j*; 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.

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

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

Where, k_i is the degree of node i; E_i is the total number of links among node *i*'s neighbors.

Dynamic structural analysis is to describe, model and predict how the new node and links are added into the network. Because many real world networks are empirically found with character of scale-free, the research about dynamic analysis are mostly focus on the evolution process of scale-free topology. That is why researchers to verify power-law distribution in degree, preferential attachment in a new real networks. However, the researches on network dynamics are still needed new mining method which with more general assumption and parameters.

2.2 Exponential Random Graph Models

Social behavior is complex, and the social processes produce network structures. To understand network evolution or structures, models can be of great value in achieving efficient representation. There are many models proposed that are useful tools for assumption and simulation. However, what exactly we need is to estimate model parameters from data and evaluate how adequately the model represents the network. The ERGM simultaneously allows for arbitrarily complex network structures to be modeled.

According to different dependence assumption, there are diverse models expressions. Bernoulli random graph distributions are generated when we assume that edges are independent. Dyadic models are for directed networks which assume dyads are independent of one another. A much more realistic assumption empirically is Markov random graphs, in which two possible network ties that have a common actor are conditionally dependent. Based on realization-dependence structures, Snijders et al. developed new specifications for exponential random graph models that include new higher order terms [11]. 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 [12, 13]. Hunter used ERGM to model high school friendship networks of varying size and found that an improved fit appeared when new parameters were included [14].

Marijtje compared the bias, standard errors, coverage rates and efficiency of maximum likelihood and maximum pseudo-likelihood estimators [15]. Meanwhile, they proposed an improved pseudo-likelihood estimation method to reduce bias. Zachary M. Saul and Vladimir Filkov used ERGM to explore biological network structure and found the model could best be achieved by using pseudo-likelihood maximization [16]. But the properties of pseudo-likelihood estimator are not well understood and the estimates are not accurate for many data sets. Later, Monte Carlo maximum likelihood estimation techniques for EGRM have been presented [17-19]. It was found that the preferred option was to use Monte Carlo estimation procedures from their research works. Martina Morris described means for controlling the Markov chain Monte Carlo (MCMC) algorithm that the package used for estimation [20].

In the recently years, ERGM has been widely used into many fields to predict the real-life networks. Goodreau applied ERGM to adolescent friendship networks in 59 U.S. schools from the National Longitudinal Survey of Adolescent Health by operating on individual, dyadic, and triadic levels [21]. Robins studied closure, connectivity and degree distributions of directed organizational network data using ERGM [22].

Cranmer used ERGM to gain unexplored parameters for prediction and found structural characters on political networks: Cosponsorship networks in the U.S. Congress and conflict networks in the international system [23]. In 2011, Simpson illustrated the utility of ERGMs for modeling, analyzing, and simulating complex whole-brain networks, and proposed a graphical goodness of fit approach to capture and reproduce the structure of fitted brain networks [24]. Ouzienko and Krivitsky expanded ERGM into temporal social networks and valued networks for model and simulation respectively [1, 25]. However, to our best knowledge, there is rare research using ERGM to study the structures of micro-blog. Therefore, how to analysis the dynamic micro-blog structure is an interesting and meaningful work by applying ERGM into this area.

3EXPERIMENT

3.1 Test Bed

Sina Weibo, with more than 250million 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 service through micro-blog. For example, doctors and hospitals open their accounts to serve the patients on micro-blog. Healthcare is a prospective and useful area to provide convenient service in social medial. Among those patients, relative majority are suffering diabetes. In China, there are currently 40 million people with diabetes that need great awareness and basic education to improve healthcare services [26]. Fortunately, more and more doctors and hospitals open accounts to help patients by posting new information and correct treatments. We can get those data from Sina API. Therefore, we search diabetes as our topic and choose 50 users whose followers are much more than others in this field, including diabetes hospital accounts, famous diabetes doctors validated accounts and diabetes magazines validated accounts, as our research seeds. The network that they and their followers have made is big enough for us to do static and dynamic structure analysis. Accounts data of 50 seed users until April 30, 2012 are listed in Table1. Its network structure plotted in R can be seen in Figure1.

Table 1. Collected diabetes Micro-blogging information

Information	Follower	Followee	Micro-blogs	Average followed
Number	27,872	444,358	48,594	8,887

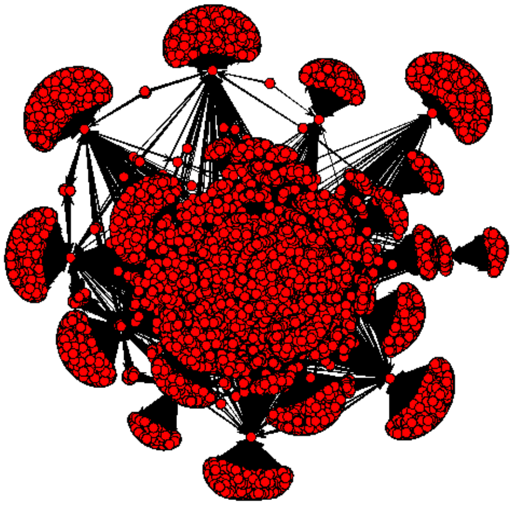


Figure1. The network structure of fifty micro-blog diabetes users

3.2 Static Analysis

In order to extract topological properties of the network that gained from micro-blog, we use four metrics to measure the static structure: degree, average shortest path, betweenness and clustering coefficient. As we discussed in section 2.1, those metrics are used to find the key node, relationship and communication links. In this paper, we use Pajek tool, which is professional software for performing network analysis, to calculate values of metrics as shown in Table 2.

Table 2. The metrics of diabetes network

Metrics	Values
Number of vertices	69547
Number of arcs	88537
Average degree	2.5461
Average shortest path	3.9262
Diameter	7
Betweenness centralization	0.1365
Clustering coefficient	0.0003

We abstract the vertices and arcs from our database that is crawled from micro-blog. It is consisted of 69,547 users and 88,537 relationships totally. We use partitions to produce a list of degree distribution. Because it is not easy to display, we use average degree as a substitution. The average degree of our network is 2.5461 that mean users averagely have 2.5 friends in the real world. However, the highest values is 5193, the lowest values is 0 whose frequency is 4. The highest frequency is 60764 and its cluster value is 1 that we can find in Figure1. In other words, there are many nodes linking to the same user who is a very famous doctor or hospital.

If we want to know the characters of links, the length of average shortest path and diameter of the network should be calculated. When using Pajek to calculate them, we use *Net>Paths between 2 vertices>All shortest/Diameter* commands to obtain the average geodesics between each two individuals and diameter in the network. The length of average shortest path is 3.9262, so users can have almost 4 lines or steps in the shortest path to connect other vertices. Among the shortest path, 7 is the largest length between two vertices which is called diameter in the network.

The more time a node is a go-between, the more central its position in the network. Betweenness centralization is one type of this metric to embody the centrality of network. In our experiment, network betweenness centralization is 0.1365, which means the proportion of go-between nodes is much small to the maximum variation in the whole network.

Clustering coefficient is often used to compute the egocentric density of all vertices in undirected network. If the directed network does not contain loops or bidirectional arcs, we can use clustering coefficient to measure the strength of sub-group formation and the density of the network. In the network we collected, there are no loops or bidirectional arcs. Therefore, we computer the clustering coefficient of our network by commanding *Net>Vector>Clustering coefficients> CC1*. We can find that the density of the network is very low with a clustering coefficient value of 0.0003. That is to say that there are many chances to link other users.

3.3 Dynamic Analysis by Using ERGMs

3.3.1 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:

$$\Pr(Y = y) = (\frac{1}{k})\exp\{\sum_{A} \eta_A g_A(y)\}$$

Where:

(i) k is a normalizing quantity to ensure the equation is a proper probability distribution;

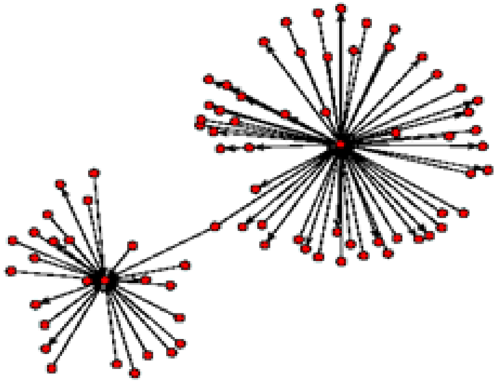
(ii) η_A is the parameter corresponding to configuration of type A;

(iii) $g_A(y)$ is the network statistic counting the frequency of sub-graph A in the graph $\mathbf{y}$;

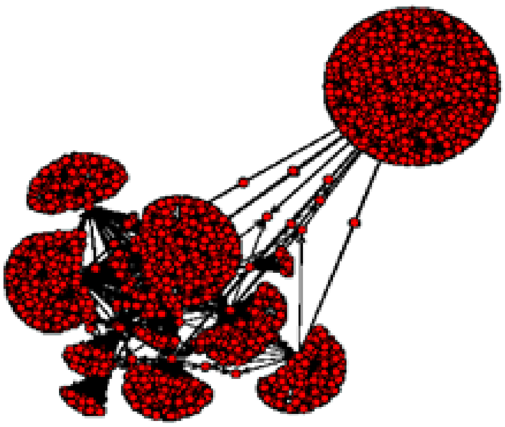
$g_A(y) = 1$ if the configuration is observed in the network **y**, and is 0 otherwise;

(iv) $\sum_A \eta_A g_A(y)$ is over all configurations types *A*.

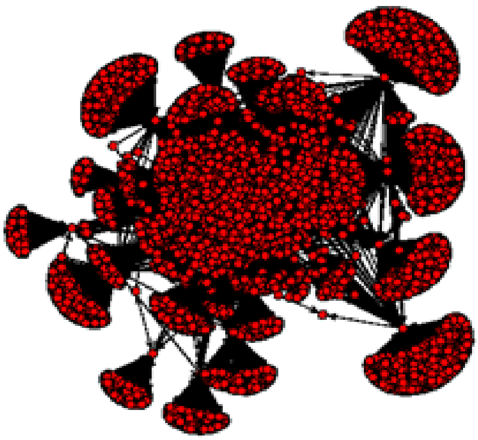
In our experiment, we collected the micro-blog data on topic of diabetes from August 2009, the time Sina Weibo was released to public using, to April 2012. To reveal the change of the network over time, we extract the data annually and accumulate new data into original dataset. Therefore, we gain four datasets at last: data set of 2009, data set of 2009 and 2010, data set from 2009 to 2011, data set from 2009 to 2012. We can use ERGM to model each network and compare their changes and development tendencies.



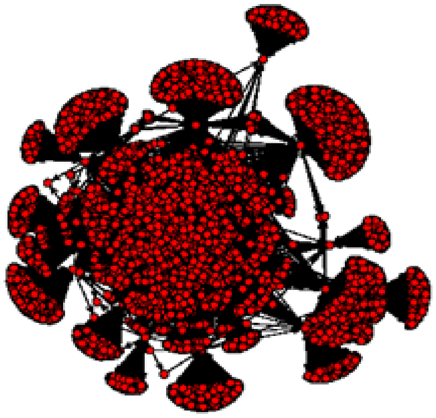
(a)



(b)



(c)



(d)

Figure2. Network structure of four data sets over time

Because of the memory limitation in R tool, we should adjust the data size of network by restraining the numbers of users' micro-bloggings they post, followers and followed before we using ERGM to model. Data set of 2009 (data set1) and appropriate data set of 2009 and 2010(data set2), data set from 2009 to 2011(data set3), data set from 2009 to 2012 (data set4) are plotted and shown in figure2 (a), (b), (c) and (d) respectively. The ***statnet*** suite of packages for R contains the ***ergm*** package (http://statnetproject.org/). More precisely, we use *"ergm"* to fit an ERGM; *"simulate"* to simulate networks from a fitted

ERGM; and "*gof*" to assess goodness of fit for an ERGM.

4 RESULTS OF ERGM

4.1 Estimation

In a certain network, the sub-graphs are the network metrics. For instance, an ERGM with just Edges metric is to show 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 use 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,

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

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

From Table 3, we can see the fit results and several phenomena. First, Edges metric with negative and large value plays an important role but is not prevalent in all four models. Mutual metric with positive and large values, 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, 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 model 1 and model 2, and the estimated parameter is significant. Although the estimated parameter of GWESP metric in data set 4 is significant, the value is small that means GWESP is not prevalent in this model. As considering the p-value of GWESP in model 3 which is not significant, it cannot be adopted in the model.

Third, there is no value of 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 Triangle metric are not significant both in model 2 and model 3. This metric is significant in model 4, but the estimated value is too small to play important role in the model.

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

Metrics		Estimate	Std. Error	MCMC s.e.	p-value
Model 1	Edges	-11.079	0.112	NA	<1e-04 ***
	Mutual	7.690	1.312	NA	<1e-04 ***
	GWESP	8.639	0.000	NA	<1e-04 ***
	Triangle	--Inf	NA	NA	NA
Model 2	Edges	-10.885	0.019	0	<1e-04 ***
	Mutual	14.388	6.904	0	0.03716 *
	GWESP	11.354	4.035	0	0.00489 **
	Triangle	2.378	1.2805	0	0.06330
Model 3	Edges	-9.618	0.008	1	<1e-04 ***
	Mutual	6.767	0.530	5	<1e-04 ***
	GWESP	0.804	1.003	4	0.423
	Triangle	0.169	0.590	4	0.775
Model 4	Edges	-9.669	0.122	0.017	<1e-04 ***
	Mutual	7.221	3.177	21.047	0.023*
	GWESP	0.818	0.026	0.054	<1e-04 ***
	Triangle	0.072	0.003	0.001	<1e-04 ***

4.2 Simulation

For a quantitative comparison of structural similarities in the generated network, we use "***simulate***" command to generate four simulated networks using ERGM based on Markov Chain Monte Carlo idea (MCMC). We compare original datasets with simulated networks referring to 10 statistics: edges, mutual ties, in-degree (0:3), out-degree (1:3), triangle. Because of the value limitation, we ignore much higher in- and out-degrees. We can find out which simulated network is more similar to original dataset from the statistics in Table4. If contrasting 10 statistics one by one, we would find that big gaps mostly exist in the degree statistics including in-degree 1, in-degree 2, out-degree1 and out-degree 2. However, it is not a good way to check the comparison results one by one. We may consider 10 statistics of dataset and simulated network as two vectors. Then we use the results of Cosine to measure their similarities. Therefore, we can see that the fourth simulated network is much more similar with its dataset than other three simulated networks in Table 4.

Since the distances of in-degree 1, in-degree 2, out-degree1 and out-degree 2 between dataset 2, data set 3 and their simulated networks are much big, the results of their Cosine 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% that is a high similarity level. According to the comparison of ten parameters and Cosine similarities, we can get the fact that the ERGM used to generate simulated networks has a good performance for dynamic micro-blog datasets.

4.3 Goodness of Fit

Comparing several parameters from the simulation to the original is of limited value. To compare the full distribution of our statistics of interest, we use "*gof*" command to visualize some common network distributions in the goodness-of-it automatically. Three metrics are 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 3-node sets having 0, 1, 2, or 3 edges among them. For a directed network, the triad census has 16 categories instead of 4.).

Because of model degeneracy issue, we cannot get last two distributions for model 1 but only get 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 [22].

Table 4. Structure comparison between simulated nets and data sets

	edges	mutual	in-0	in-1	in-2	in-3	out-1	out-2	out-3	triangle	Cosine
Dataset 1	81	1	2235	27	0	0	52	1	0	0	**99.95%**
Net 1	87	1	2178	85	1	0	83	2	0	0	
Dataset 2	4412	38	14610	277	18	2	3659	144	27	71	**98.68%**
Net 2	4437	38	12245	2283	264	13	3180	453	44	125	
Dataset3	18078	789	12636	1959	178	62	9839	1088	338	9031	**98.69%**
Net 3	18541	789	10156	3798	733	142	7219	2380	565	8601	
Dataset 4	18682	851	12845	2083	182	63	10063	1131	357	9397	**99.98%**
Net 4	18713	851	12564	2307	233	64	9676	1348	366	9319	

Therefore, we give goodness-of-fit diagnostics for model 2, model 3 and model 4 as followed Figure 4, Figure 5 and Figure 6. In these figures, the vertical axis is the logit of relative frequency, the boxplots summarize the statistics for the simulated networks resulting from the MLE, and the solid line in each plot represents the statistics of the observed networks.

Distance, is a global property of the network, can be used to measure that how well the observed and simulated distributions match. When facing Figure 3, Figure 5 and Figure 6, we can see that model 1, model 3 and model 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. That means the observed proportion of pairs of nodes with shortest connecting path length from 1 to 15 is much similar to simulated one for model 2.

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

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

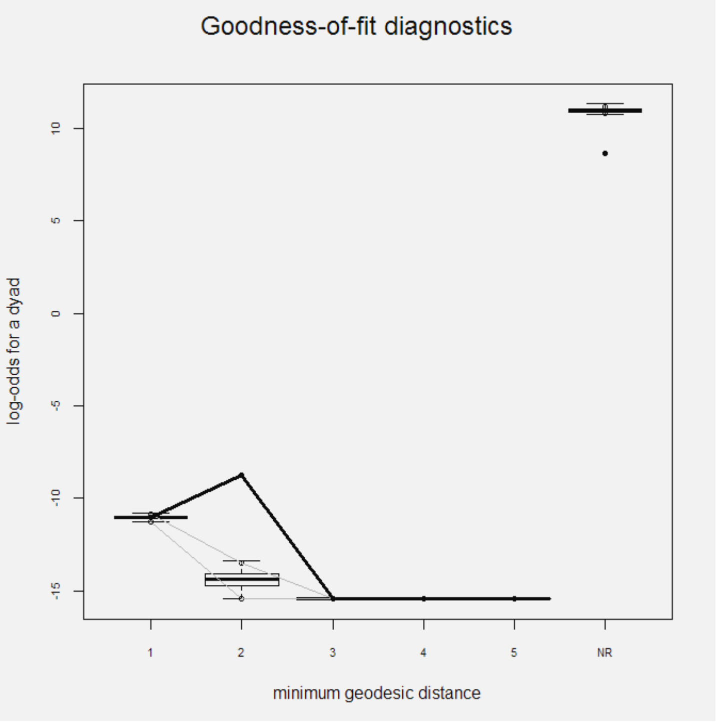


Figure 3. Model 1 goodness of fit for geodesic distance.

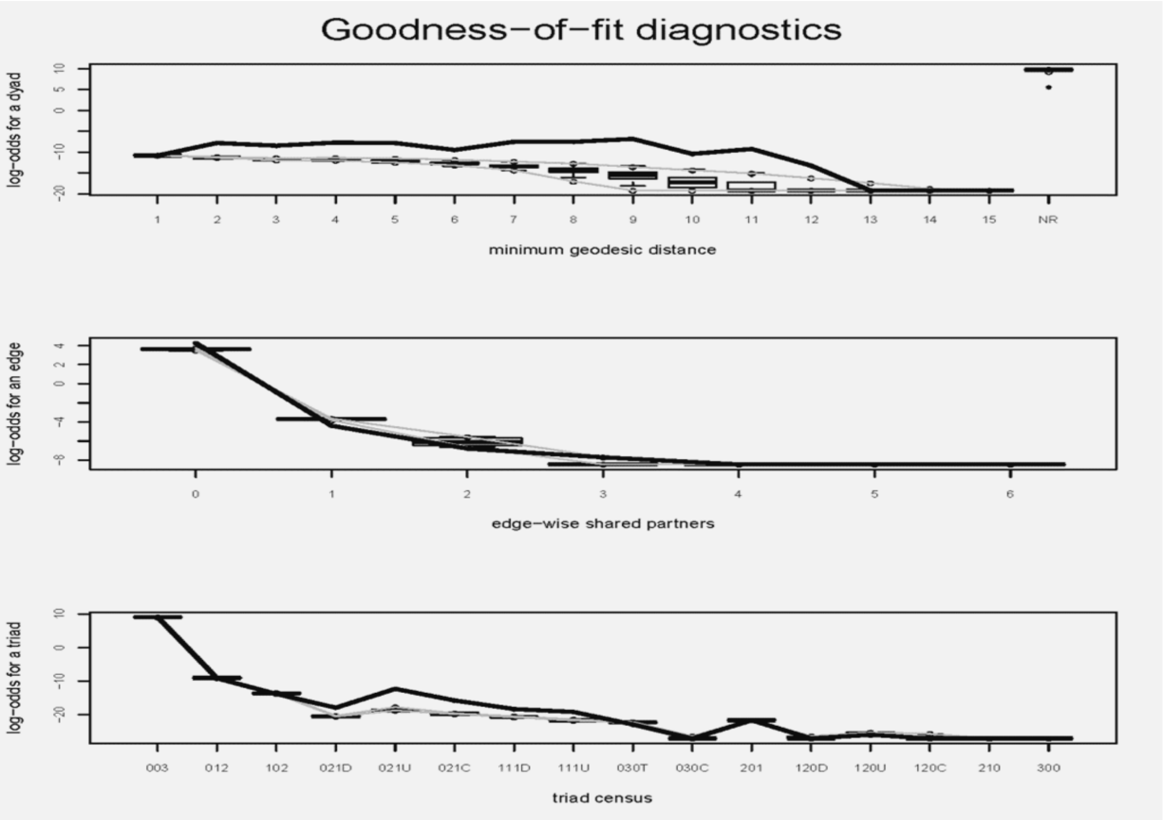


Figure 4. Model 2 goodness of fit for geodesic distance, edgewise shared partner and triad census.

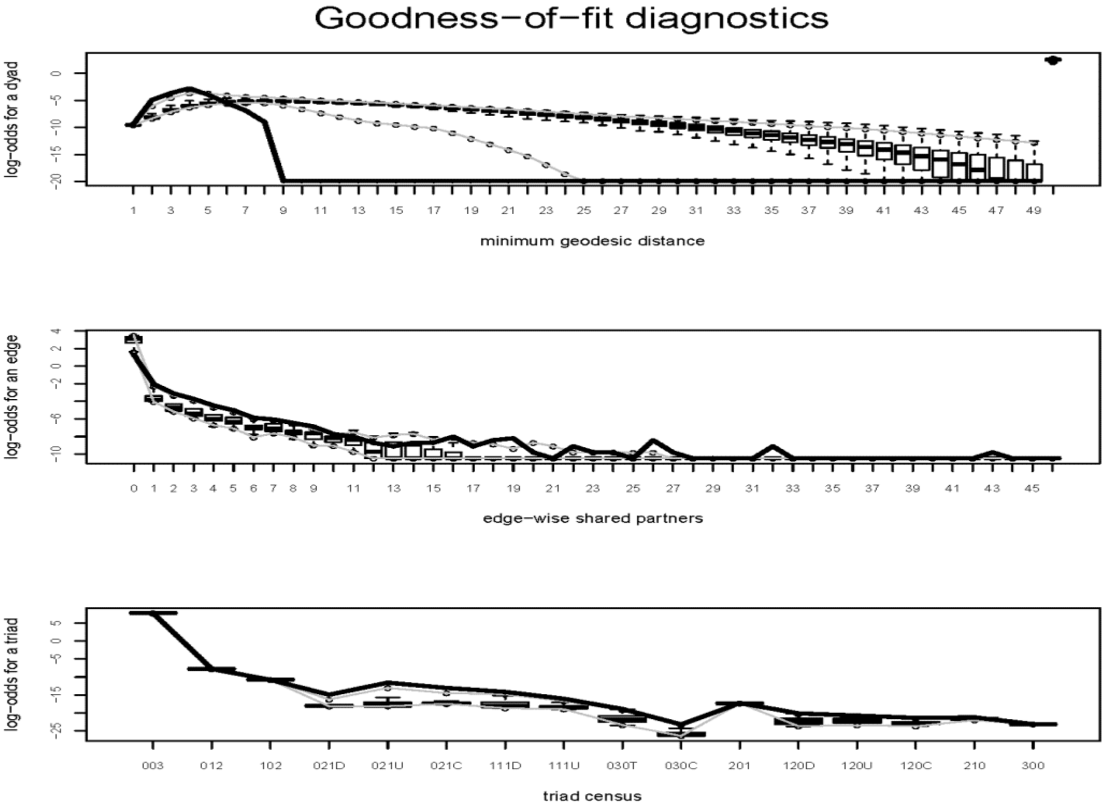


Figure 5. Model 3 goodness of fit for geodesic distance, edgewise shared partner and triad census.

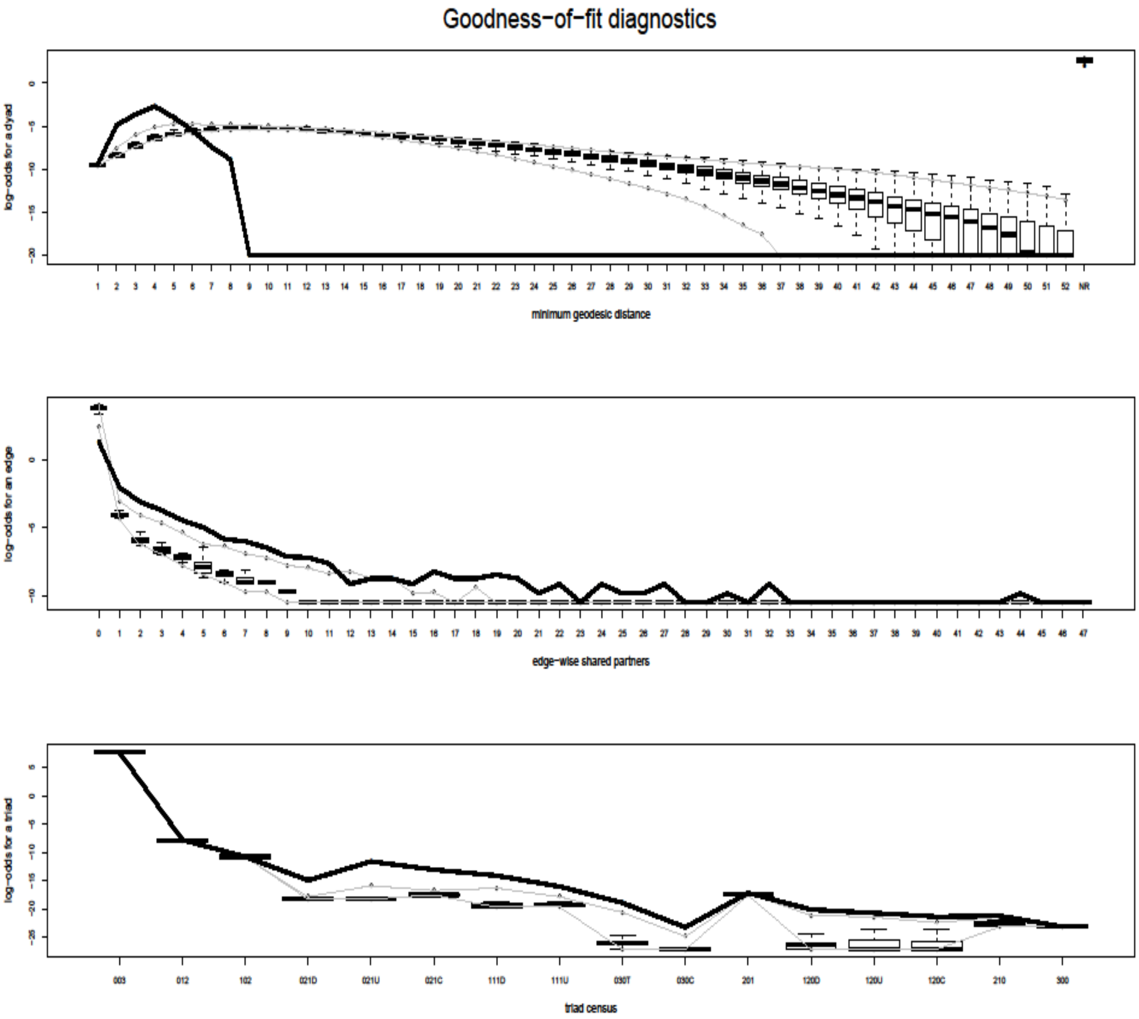


Figure 6. Model 4 goodness of fit for geodesic distance, edgewise shared partner and triad census.

5 CONCLUSION

In the previous sections, our analyses illustrate Chinese diabetes micro-blog structure both in static and dynamic way. Usually, four common metrics are chosen to analyze network static structure, including degree, average shortest path, betweenness and clustering coefficient. According to values of those metrics, we find 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 get more detail information and characteristics of diabetes micro-blog network if we just analyze the static structure. Therefore, we need to analyze the change of micro-blog network to know a network deeply based on scientific interest.

Most important contribution of this paper is focusing our analysis on dynamic structure analysis. Exponential random graph models (ERGMs) are adopted for modeling, estimating, and simulating micro-blog networks. To reveal the change of the networks over time, we extracted the data annually and accumulated new data into original dataset from August, 2009 to April, 2012. Then we use ERGM to model each network and compare their structure changes. It can be clearly seen that how the network expand from a small network to a complicated network in Figure 2. We use Edges, GWESP, Mutual and Triangle metrics to measure global efficiency, local clustering efficiency, mutual ties effect and transitivity respectively in our ERGMs. In order to check the model parameters, we choose Monte Carlo maximum likelihood estimation to find out which estimated parameter is significant in each model. In our experiment, mutual tie metric is significant in each model but triangle metric is only significant in the fourth dataset. We can get a conclusion that the transitivity evidently show up when the micro-blog network becoming big enough.

Also, four simulated networks using ERGM based on Markov Chain Monte Carlo are generated to compare with original datasets. We analyze the gaps of 10 statistics between simulated network and observed dataset. More meaningfully, we contrast their Cosine similarities and find that the last group has the highest similarity value (99.98%), and the second group has the lowest value (98.68%). All four Cosine similarities are much high. It means that simulated networks generated by ERGM have good performance for dynamic micro-blog datasets.

Moreover, the goodness-of-fit approach gives us the scientific interest to capture and reproduce the structure of the fitted network. We represent the complex network data using ERGM and exam the simulated network's distances and local structural components. For those models we exam the geodesic distribution, edgewise shared partners distribution, and the triad census distribution. Goodness-of-fit simulations suggest that model2 and model 3 are well behaved in reflecting edgewise shared partner of data set 2 and data set 3. And model 2 does much better than model 3 and model 4 when capturing the triad census distribution of micro-blog network.

ACKNOWLEDGEMENT

This work is also supported by the National Natural Science Foundation of China (Grant Nos. 71171068). The authors also acknowledge the support by the China Scholarships Council (File NO.2011612202).

6 REFERENCES

1. Ouzienko, V., Guo, Y., Obradovic, Z.: A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks. Statistical Analysis and Data Mining, pp.470--486 (2011).
2. Agarwal, N., Galan, M. Liu, H., Subramanya S.: Clustering of Blog Sites Using Collective Wisdom. Computational Social Network Analysis 1,107--134 (2010).
3. Memon, N., Xu, J., Hicks, D.L., Chen, H.: Social Network Data Mining: Research Questions, Techniques, and Applications. Data Mining for Social Network Data, pp.1--7 (2010).
4. Larsson, A.O., Moe, H.: Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society 14(5): 729-747(2012).
5. Pak, A. Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10 Valletta, 5 (2010).
6. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In Proceedings of the ACL 2011 Workshop on Languages in Social Media, pp. 30-38 (2011).
7. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in Twitter events. Journal of the American Society for Information Science and Technology 62(2):406-418 (2011).
8. Robins, G. Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p*) models for social networks. Social Networks 29:173-191 (2007).
9. Ahram, TZ., Karwowski, W.: Visual Social Network Analysis: Effective Approach to

Model Complex Human Social, Behaviour & Culture. Work-A Journal of Prevention Assessment & Rehabilitation 41: 3504-3510 (2012).

10. Hua, G., Sun, Y., Haughton, D.: Network Analysis of US Air Transportation Network. Data Mining for Social Network Data, pp.75-89 (2010).
11. Snijders, T., Pattison, P., Robins, GL., Handcock, M.: New Specifications for Exponential Random Graph Models. Sociological Methodology. pp. 36-99 (2006).
12. Steven M. Goodreau.: Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network. Social Networks 29: 231-248 (2007).
13. Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P.: Recent developments in exponential random graph (p*) models for social networks. Social Networks 29:192-215 (2007).
14. Hunter, DR., Handcock, MS., Butts, CT.: ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Journal of Statistical Software 24(3): 1-29 (2008).
15. Alfó, M., Nieddu, L., Vicari, D.: Finite Mixture Models for Mapping Spatially Dependent Disease Counts. Biometrical Journal 51: 84-97 (2009).
16. Saul, ZM., Vladimir Filkov, V.: Exploring biological network structure using exponential random graph models. Bioinformatics 23(19):2604-2611 (2007)
17. Snijders, T.A.B.: Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3, 2 (2002).
18. Handcock, MS., Hunter, DR., Butts, CT., Goodreau, SM., Morris, M.: statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. Journal of Statistical Software 24, 1 (2008).
19. Robins, G., Morris, M.: Advances in Exponential Random Graph (p*) Models. Social Networks 29(2): 169-172 (2007).
20. Morris, M., Handcock, MS., Hunter, DR.: Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. Journal of Statistical Software 24(4):1548-7660 (2008).
21. Goodreau, SM., Kitts, JA., Morris, M.: Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. Demography 46(1):103-25 (2009).
22. Robins, G., Pattison, P., Wang, P.: Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. Social Networks 31:105-117 (2009).
23. Cranmer, SJ., Desmarais, BA.: Inferential Network Analysis with Exponential Random Graph Models. Political Analysis 19: 66-86 (2011).
24. Simpson, SL., Hayasaka, S., Laurienti, PJ.: Exponential Random Graph Modeling for Complex Brain Networks. PLoS ONE 6, 5 (2011).
25. Krivitsky, PN.: Exponential-family random graph models for valued networks. Electronic Journal of Statistics 6:1100-1128 (2012).
26. Yang, W., Lu, J. et al.: Prevalence of Diabetes among Men and Women in China. The New England Journal of Medicine 362:1090-1101 (2010).