International Journal of Modern Physics C
Vol. 25, No. 5 (2014) 1440007 (11 pages)

DOI: 10.1142/S0129183114400075

An improved sampling method of complex network

Qi Gao*, Xintong Ding†, Feng Pan‡ and Weixing Li§

School of Automation, Beijing Institute of Technology
5 South Zhongguancun Street, Haidian District
Beijing 100081, P. R. China
**gaoqi@bit.edu.cn*
†*dingxintong@126.com*
‡*andropan@gmail.com*
§*liweixing@bit.edu.cn*

Received 30 June 2013
Accepted 30 September 2013
Published 4 December 2013

Sampling subnet is an important topic of complex network research. Sampling methods influence the structure and characteristics of subnet. Random multiple snowball with Cohen (RMSC) process sampling which combines the advantages of random sampling and snowball sampling is proposed in this paper. It has the ability to explore global information and discover the local structure at the same time. The experiments indicate that this novel sampling method could keep the similarity between sampling subnet and original network on degree distribution, connectivity rate and average shortest path. This method is applicable to the situation where the prior knowledge about degree distribution of original network is not sufficient.

Keywords: Complex network; sampling method; subnet.

PACS Nos.: 11.25.Hf, 123.1K.

1. Introduction

Complex network has been a hot field since Watts and Strogatz published the paper about small-world network. With the rapid development of technology, complex network model is widely used in mathematics, biology, social sciences, Internet, etc.[1,2] Various models to explain the observed characteristics of those real networks have been introduced and studied by both numerical and analytic approaches.[3,4]

Some networks, such as social network and Internet, are quite huge. It is impossible to analyze the characteristics of such networks. So people focus on sampled subnets. For example, Y2H-derived partial sampling method was used in protein research.[5] In other case, some networks treated as original networks are also sampling subnets in fact. In scientists collaborations networks, two scientists are

‡Corresponding author.

considered to be connected if they have coauthored a paper, which neglects the other forms of collaborations in real scientific research. In a sense, multidimensional properties of complex network would cause this kind of result.[6]

It is known that the characteristics of subnets are different from the characteristics of original networks.[7] The methods applied in sampling process influence the similarity between subnets and original networks.[8] The regular way to improve the quality of sampling subnets is to increase sampling rate with increasing of cost. In this paper, an efficient sampling method is designed in order to improve the effect of sampling process with low cost.

2. Sampling Methods

2.1. *Common sampling methods*

Random sampling method, hub sampling method and snowball sampling method are three common sampling methods used in complex networks.

In random sampling method, every node in the network is picked out with the same probability. This method ignores the structural property of original network, but sampling through out the entire network.

Hub sampling method outperforms other methods when the degree distribution $P(k)$ of the original network is known. The probability that a node with degree k is picked out as a sampled node $\rho(k)$ depends on k.

$$\rho(k) = \frac{k^{\alpha}}{\sum_k k^{\alpha} p(k)} \rho_0, \quad 0 \leq \alpha \leq \infty. \tag{1}$$

ρ_0 is sampling rate. If $\alpha = 0$, it turns into random sampling method. The greater a node's degree is, the more likely it is picked out.

In snowball sampling,[9] first, choose a single node and all the nodes directly connected to it are picked out. Second, all the nodes connected to those picked in the last step are selected, and this process is continued until the desired number of nodes are sampled. Snow ball sampling method could find out the local structure of networks. In Cohen's paper,[10] snowball sampling is slightly changed: in the second step, all the nodes connected to those picked nodes in the last step are selected with a probability P, the rest process is the same. It turns out that this change can seek out nodes with high degree.

2.2. *RMSC sampling method*

Consider the abilities and shortcoming of different sampling methods, Random Multiple Snowball with Cohen (RMSC) process sampling method is proposed:

First, pick certain nodes out randomly as seeds. For each seed, look for the nodes connected to it and these nodes are picked out with a probability P_c.

Second, all the nodes connected to those picked nodes in the last step are found and picked out with the same probability P_c and this process continues until the desired number of nodes are sampled. The growth rate of each seed is constrained in the sampling process.

RMSC sampling method gives considerations to both global exploring and local discovering.[7,10] It would be helpful to acquire quality sampling subnets with low cost.

2.3. *Evaluation indexes*

Single evaluation index could not be used to evaluate the similarity. Degree distribution, connectivity rate average shortest path (ASP) and average clustering coefficient are normally used in evaluating.

The degree of a node is the number of links connecting to itself. Degree distribution denoted by $P(k)$ describes the frequency of degree k of all the nodes in the network.

Its inevitable that isolated nodes appears in the subnets. These isolated nodes reduce the valid nodes of subnets and have an influence on the characteristics of subnets. Connectivity rate is about how many reachable nodes pairs there are in the subnet. Define connectivity rate

$$\eta = \frac{P_p}{P_n} \times 100\%, \tag{2}$$

P_p is reachable nodes pairs,

P_n is the maximum number of pairs can be created in N nodes.

$$P_n = \frac{1}{2} N(N-1). \tag{3}$$

Shortest path d_{ij} is the minimum number of links between two different reachable nodes i and j. The ASP L of a network is the average of all pairs of nodes.

$$L = \frac{1}{P_p} \sum_{i>j} d_{ij}. \tag{4}$$

Clustering coefficient stands for the community characteristic of the network. Node i connects k_i nodes. These k_i nodes can have $\frac{k_i(k_i-1)}{2}$ links at most. While the real links are E_i. The clustering coefficient C_i of node i is

$$C_i = \frac{2E_i}{k_i(k_i-1)}. \tag{5}$$

The average clustering coefficient of the network C is

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{6}$$

3. Data and Experiment

3.1. *Networks*

Connecting nearest neighbors (CNN) network, random network, rule network and small-world network are used as original networks in the experiment.

CNN model comes from social networks and it is the representative model for social network evolution. CNN network is generated by iteratively performing two rules:

(1) With probability $1 - \mu$ introduce a new node in the graph, create a link from the new node to a node j selected at random (implying the creation of a potential link between the new node and all the neighbors of node j).
(2) With probability μ convert one potential link selected at random into a link.[11–13]

CNN network is generated as experiment data with 10 000 nodes and the parameter $\mu = 0.7$.

Random network is one of the most common networks. It has short ASP and its degree distribution is Poission distribution. Rule network is the opposite of random network. In rule network, the degree of each node is same. It has big clustering coefficient and long ASP. Small-world network is a network between rule network and random network. Its ASP is short and clustering coefficient is big at the same time.

Watts and Strogatz built WS small-world model with the following rules:

It starts from rule network. Consider a circle of N nodes, each node connects k (even number) nearest nodes around its both sides. k is the degree of each node. Then reconnect all the links with probability P, namely fix one node of one link and change the other node into another one comes from the N nodes randomly. Only one link at most exist between any two different nodes and every node cannot connect itself. If $P = 0$, it is rule network. If $P = 1$, it is random network. If $0 < P < 1$, it is WS small-world network.[14]

In the experiment, a rule network including 10 000 nodes has been used. The degree of each node is 20. Small-world network is generated with parameter $P = 0.5$ and random network is generated with parameter $P = 1$ from the same rule network above.

3.2. *Experiment*

In the experiment, CNN network, random network, rule network and small-world network are used as original networks. Each network is sampled by RMSC sampling method, random sampling method and hub sampling method.[15] It starts with 20 seeds in every networks. For RMSC sampling method, P_c is 0.5. Degree distribution, connectivity rate, ASP and average clustering coefficient of subnets are regarded as the evaluation indexes.

4. Result

4.1. *Degree distribution*

Figures 1–4 show the degree distribution of networks.

In Fig. 1, CNN network, when sampling rate $\rho_0 = 0.2$, RMSC sampling method follows the topology structure of original network. RMSC sampling method and hub

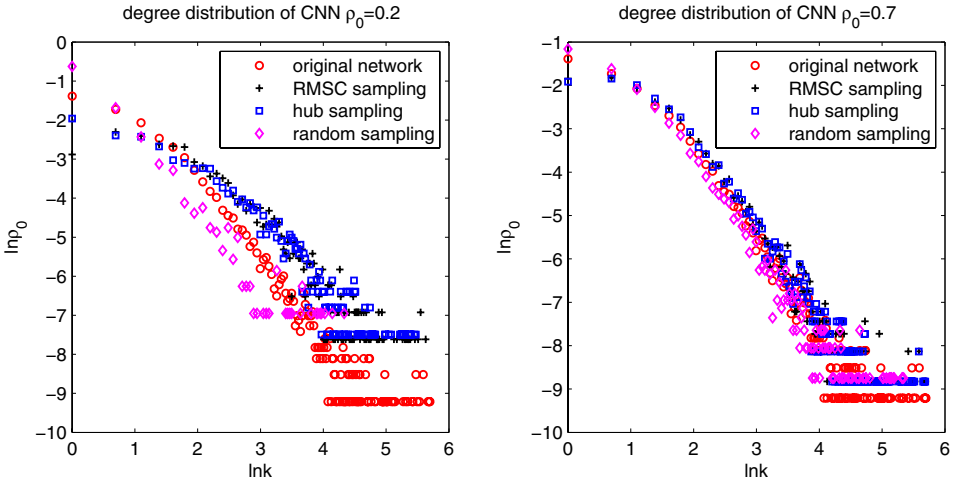

Fig. 1. (Color online) Degree distribution of CNN network with sampling rate $\rho_0 = 0.2$ and $\rho_0 = 0.7$.

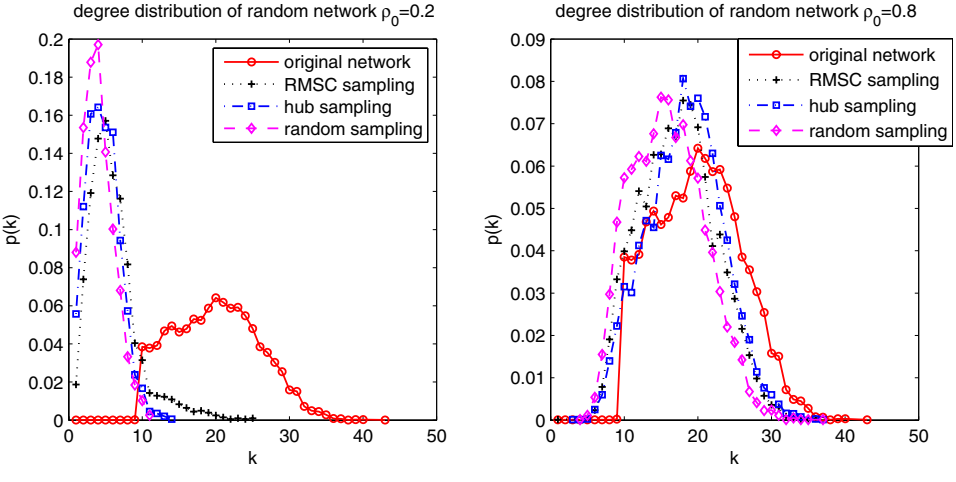

Fig. 2. (Color online) Degree distribution of random network with sampling rate $\rho_0 = 0.2$ and $\rho_0 = 0.8$.

sampling method perform quite similarly, but random sampling deviates from the original network. In rule network (Fig. 3), RMSC sampling reflect the degree distribution of original network very well. In the rest two networks, all the sampling methods does not keep the degree distribution of original networks while the subnet sampled by RMSC is the closest one.

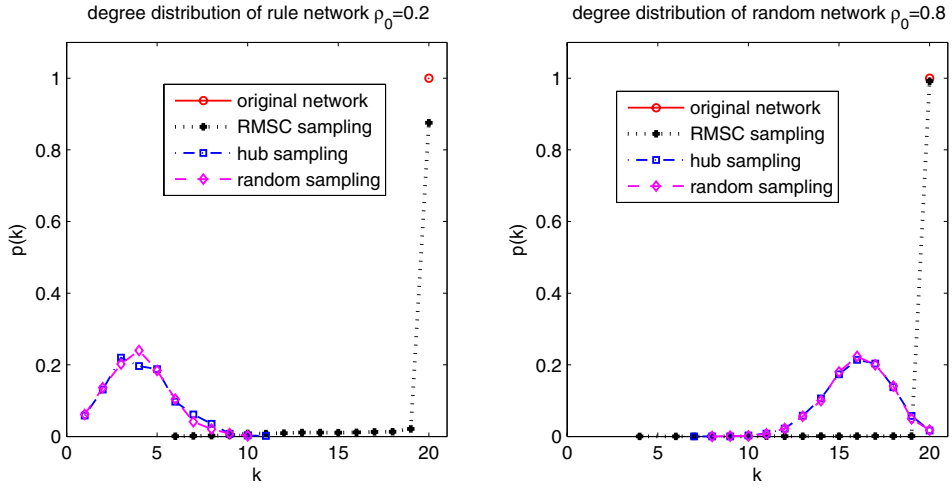

Fig. 3. (Color online) Degree distribution of rule network with sampling rate $\rho_0 = 0.2$ and $\rho_0 = 0.8$.

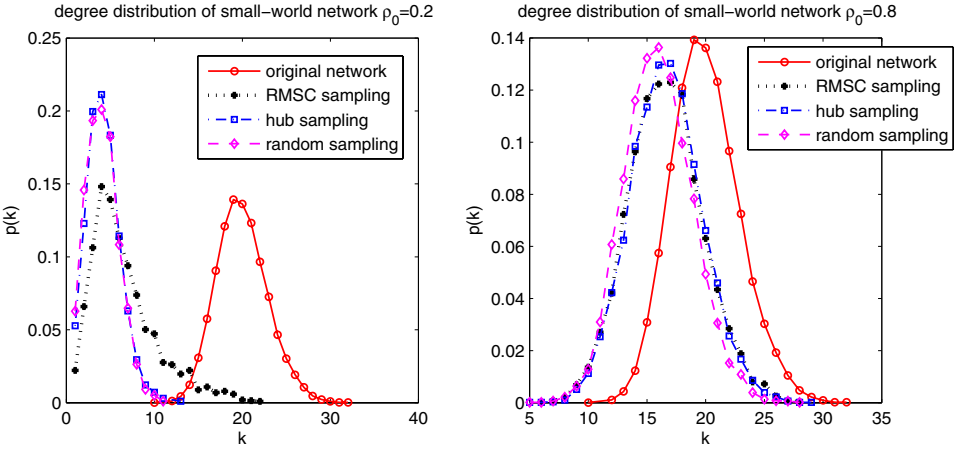

Fig. 4. (Color online) Degree distribution of small-world network with sampling rate $\rho_0 = 0.2$ and $\rho_0 = 0.8$.

4.2. *Connectivity rate*

Tables 1–4. show the connectivity rate of subnets. In CNN network, random network and small-world network, RMSC sampling method keeps high connectivity rate at low sampling rate. In rule network, the result is not good.

4.3. *Average shortest path*

Figures 5–8 show the ASP of networks.

In Fig. 5, the effect of RMSC sampling method is similar to that of hub sampling. Figures 6 and 8 show the satisfactory results of RMSC sampling method. In rule

Table 1. Connectivity rate of CNN.

	Random sampling		Hub sampling		RMSC sampling	
Sampling rate	ASP	η	ASP	η	ASP	η
0.2	4.41537	6.78025%	4.18630	71.42%	4.48865	98.89%
0.4	5.44713	23.49%	4.96034	82.00%	4.84327	99.97%
0.6	6.10441	40.28%	5.71193	88.71%	5.72207	100%
0.7	6.5602	57.75%	6.01989	92.55%	5.9157	100%
1.0	6.87258	100%	6.87258	100%	6.87258	100%

Table 2. Connectivity rate of random network.

	Random sampling		Hub sampling		RMSC sampling	
Sampling rate	ASP	η	ASP	η	ASP	η
0.2	5.54269	95.45%	5.06704	97.12%	4.27727	100%
0.4	4.18566	99.95%	3.91856	99.95%	4.27727	100%
0.6	3.76696	100%	3.58038	100%	3.59756	100%
0.8	3.54891	100%	3.41075	100%	3.46528	100%
1.0	3.39231	100%	3.39231	100%	3.39231	100%

Table 3. Connectivity rate of rule network.

	Random sampling		Hub sampling		RMSC sampling	
Sampling rate	ASP	η	ASP	η	ASP	η
0.2	4.40492	0.84%	3.96475	0.83%	10.44381	10.75%
0.4	34.34222	7.15%	42.28821	9.33%	26.2216	13.51%
0.6	356.55738	100%	347.06856	95.57%	44.15877	16.37%
0.8	256.67554	100%	256.25618	100%	99.53316	34.87%
1.0	250.47505	100%	250.47505	100%	250.47505	100%

Table 4. Connectivity rate of small-world network.

	Random sampling		Hub sampling		RMSC sampling	
Sampling rate	ASP	η	ASP	η	ASP	η
0.2	6.01751	96.92%	5.81225	99.96%	4.47254	100%
0.4	4.42147	100%	4.3603	100%	4.06150	100%
0.6	3.92529	100%	3.84803	100%	3.79222	100%
0.8	3.68133	100%	3.6371	100%	3.64387	100%
1.0	3.54031	100%	3.54031	100%	3.54031	100%

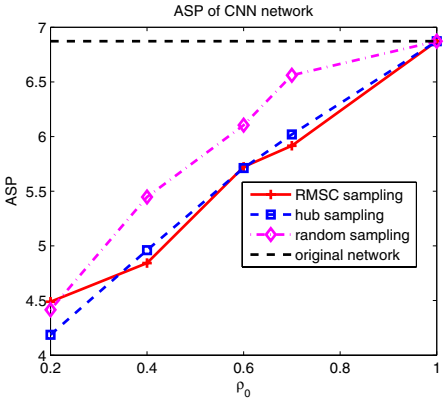

Fig. 5. (Color online) ASP of CNN.

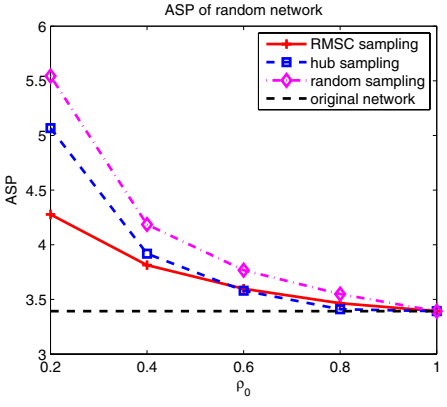

Fig. 6. (Color online) ASP of random.

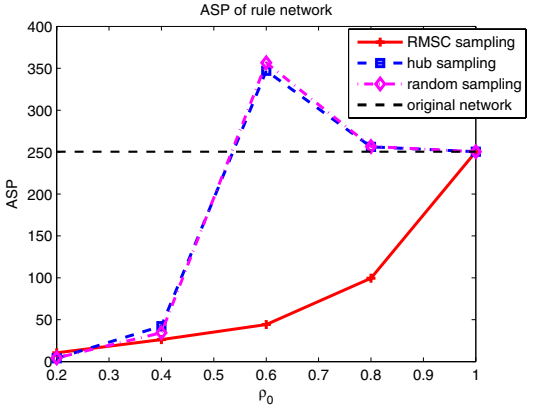

Fig. 7. (Color online) ASP of rule network.

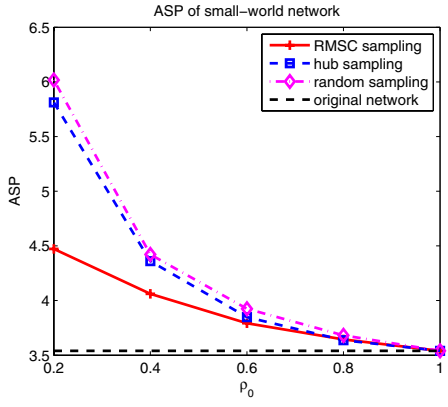

Fig. 8. (Color online) ASP of small-world network.

network, the ASP of subnet sampled by RMSC sampling method is short. It is because that RMSC sampling method is breadth first search sampling method, the paths between the picked out nodes are short. As a result, the ASP is quite short.

4.4. *Average clustering coefficient*

Figures 9–12 show the average clustering coefficient of networks.

The average clustering coefficient of subnets sampled by RMSC sampling method is bigger than that of original networks. The reason is that the subnet sampled by RMSC sampling method grows depends on the connection between nodes, the picked nodes starts from the same seed are all reachable from same snowball. When connectivity rate is high, the average clustering coefficient of subnet is big.

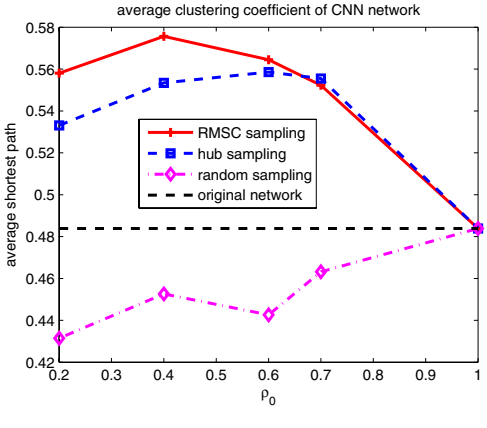

Fig. 9. (Color online) Average clustering coefficient of CNN.

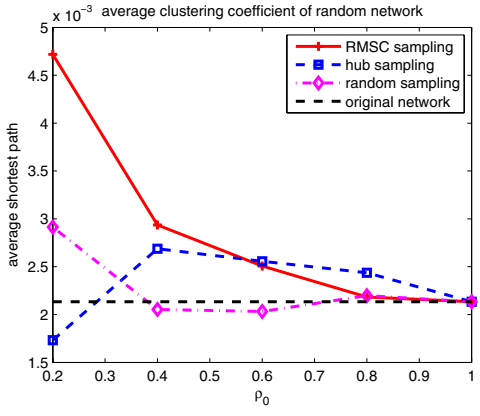

Fig. 10. (Color online) Average clustering coefficient of random network.

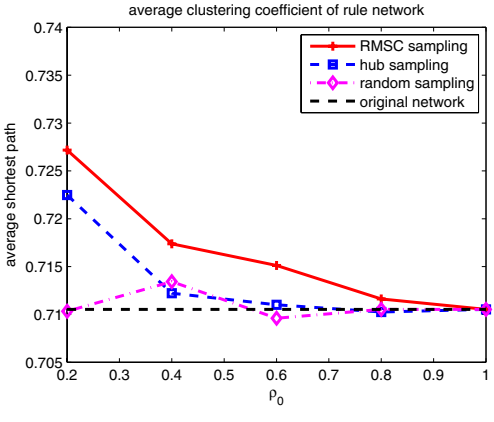

Fig. 11. (Color online) Average clustering coefficient of rule network.

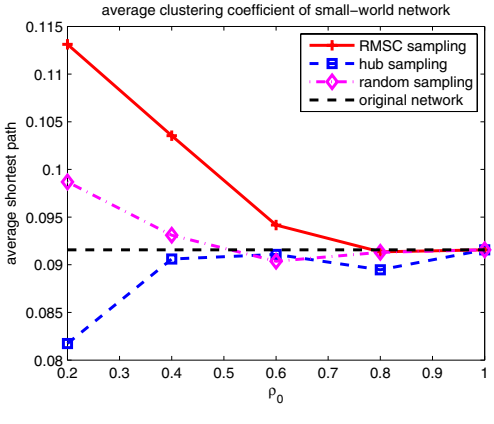

Fig. 12. (Color online) Average clustering coefficient of small-world network.

5. Conclusion

RMSC sampling method is an improved sampling method which integrates random sampling and snowball sampling. It is able to exploring global information and local structure of networks at the same time. Experiment results demonstrate that RMSC sampling method can reflect the characteristics of original networks very well. The effect is similar to that of hub sampling method. The obvious advantage is that RMSC sampling method does not rely on the prior knowledge about degree distribution of original network. In most cases, we do not have enough information of original network. We use sampling subnets to find out the basic structure properties of it. So RMSC sampling method is a good choice under this situation. It should be a useful and quality sampling method of complex network.

References

1. D. J. Watts, *Annu. Rev. Sociol.* **30**, 243 (2004).
2. S. N. Dorogovtsev and J. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
3. S. Yong *et al.*, Using Complex Network Theory in the Internet Engineering, in *Int. Conf. on Computer Science & Education (ICCSE), 7th International Conference on IEEE* (2012), p. 390.
4. K. Wehmuth and A. Ziviani, Distributed assessment of network centralities in complex social networks, in *ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM) IEEE/ACM International Conference on IEEE* (2012), p. 1046.
5. D. Bernheim and M. D. Whinston, *Rand J. Econ.* **21** (1990).
6. G. Qi and L. Xiaoting, The multidimensional properties of complex network, in *Proc. Int. Conf. on Information Technology, Computer Engineering and Management Sciences (ICM), International Conference on IEEE*, Vol. 4 (2011), p. 299.
7. M. P. H. Stumpf, C. Wiuf and R. M. May, *Proc. Nat. Acad. Sci. USA*, 102 (2005) 4221.
8. S. H. Lee, P.-J. Kim and H. Jeong, *Phys. Rev. E* **73** (2006).
9. L. A. Goodman, *Ann. Math. Statist.* **32**, 148 (1961).
10. R. Cohen, S. Havlin and D. Ben-Avraham, *Phys. Rev. Lett.* **91** (2003).
11. A. Vzquez, *Phys. Rev. E* **67** (2003).
12. B. Rudolf *et al.*, *Phys. Rev. E* **85** (2012).
13. D. Juher, J. Saldana and J. Soler, *Physical D* **214**, 132 (2006).
14. J. Illenberger and G. Flotterod, *Social Networks* **34** (2012).
15. Y. Bo, G. Hai-xia, C. Zhong, Efficient sampling strategies for large-scale complex networks, in *Int. Conf. on Management Science and Engineering (ICMSE), 15th Annual Conference Proceedings, International Conference on IEEE* (2008), p. 334.