Published in "Physics Letters A 381(11): 970–975, 2017"
which should be cited to refer to this work.

Evolution properties of the community members for dynamic networks

Kai Yang [a], Qiang Guo [a], Sheng-Nan Li [a], Jing-Ti Han [b], Jian-Guo Liu [b,c,*]

[a] Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China
[b] Data Science and Cloud Service Research Centre, Shanghai University of Finance and Economics, Shanghai 200433, PR China
[c] Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland

The collective behaviors of community members for dynamic social networks are significant for understanding evolution features of communities. In this Letter, we empirically investigate the evolution properties of the new community members for dynamic networks. Firstly, we separate data sets into different slices, and analyze the statistical properties of new members as well as communities they joined in for these data sets. Then we introduce a parameter φ to describe community evolution between different slices and investigate the dynamic community properties of the new community members. The empirical analyses for the Facebook, APS, Enron and Wiki data sets indicate that both the number of new members and joint communities increase, the ratio declines rapidly and then becomes stable over time, and most of the new members will join in the small size communities that is $s \leq 10$. Furthermore, the proportion of new members in existed communities decreases firstly and then becomes stable and relatively small for these data sets. Our work may be helpful for deeply understanding the evolution properties of community members for social networks.

1. Introduction

Community structures are the natural properties of social networks [1–5], which evolves with the network structures changing [6–11]. Detecting community structures in dynamic networks [12–16] has attracted much attention. Palla et al. [17] and Greene et al. [18] developed models for tracking the evolution process of communities for dynamic networks, where each community is characterized by a series of significant evolutionary events, including growth, contraction, merging, splitting, birth and death. Asur et al. [19] developed a framework for capturing and identifying community events which are used to characterize complex behavioral patterns of individuals and communities over time. Gauvin et al. [20] used the non-negative tensor factorization method to extract the community activities of dynamic networks. Wang [21] found that community merging depends largely on the clustering coefficient of the nodes connecting two communities directly, while community splitting depends on the clustering coefficient of the nodes in the community for social networks. What's more, the human collective behavior dynamics during community evolution for the networks is important for understanding the evolution of networks. Backstrom et al. [22] analyzed the role of common friends in community formation and focused on what are the structural features that influence whether individuals will join in communities, and which communities will grow rapidly. Most previous studies have concentrated on determining community events based on the community features extracted at different timestamps. Understanding the structure and dynamics of communities is a natural goal for network analysis, since such communities tend to be embedded within larger social network structures and many new members join in the network. Therefore, it should be noticed that the collective behaviors of new community members play an important role for the network evolution.

In this Letter, we empirically analyze the behavior characteristics of new community members. Firstly, we separate data sets into different slices, and investigate the number of new members and the number of the communities they joined in at each timestamp. By using the Blondel method [23], we investigate the size of communities which the new members join in and evolution properties of new community members. We introduce a parameter φ for describing the evolution characteristics of the new commu-

* Corresponding author.
E-mail address: liujg004@ustc.edu.cn (J.-G. Liu).

nity members at different timestamps. The empirical results for the Facebook, APS, Enron and Wiki data sets indicate that the number of new members for these data sets increases over time, the ratio of the new members to all community members decreases firstly and then becomes stable or decreases. As the number of joint communities by the new members increases, the ratio between the number of communities new members joined in and the total number of communities at each timestamp decreases firstly and then keeps stable. In addition, most of the new members tend to join in the small size communities ($s \leq 10$) by analyzing the number of new members joining in the communities of different sizes for all timestamps.

2. The methods and the theoretical hypothesis

2.1. Detection algorithm

In this section, we introduce the Blondel method to detect communities at each timestamp. Blondel et al. [23] proposed a fast greedy approach based on modularity optimization [24], which could be divided into two steps repeated iteratively. Initially, each node in network is formed a community. Then, for each node i, one considers the neighbor j of node i and calculate the modularity increment by removing i from its community and by placing it in the community of j. The node i is placed in the community for which this increment is maximum, but only if this increment is positive. If no positive increment is possible, the node i stays in its original community. This process is applied repeatedly for all nodes until no further increment can be achieved and the first step is then complete. The second step of the algorithm consists in building a new network whose nodes are the communities found in the first step. The weights of the links between the new nodes which are communities in the first step are given by the sum of the number of the links between nodes in the corresponding two communities. Once this second step is completed, it is then possible to reapply the first step of the algorithm. The whole process is described in Algorithm 1.

Algorithm 1 Pseudo-code of Blondel method.

```
1: G is the initial network
2: repeat
3:    Put each node of G in its own community
4:    while some nodes are moved do
5:     for all node n of G do
6:       place n in its neighboring community including its
         own which maximizes the modularity improvement
7:     end for
8:    end while
9:    if the new modularity is higher than the initial
      then
10:      G = the network between communities of G
11:   else
12:      Terminate
13:   end if
14: until
```

2.2. Evolving communities

We model the dynamic networks as a sequence of networks $\{G_1, G_2, \cdots, G_n\}$, where $G_t = (V_t, E_t)$ denotes a network at timestamp t. And $|V_t|$ is the number of nodes at timestamp t. The communities at each timestamp can be detected by the Blondel method. Suppose there are $N_c(t)$ communities detected at the tth timestamp, denoted by $\{C_t^1, C_t^2, \cdots, C_t^{N_c(t)}\}$, where the ith community could be denoted as $C_t^i = (V_t^i, E_t^i)$ and $V_t^i \subseteq V_t$, $E_t^i \subseteq E_t$.

In this Letter, we introduce a parameter $\varphi_{i,j}(t, t+1)$ to describe community evolution, that is Jaccard similarity coefficient to qualify the similarity of communities at consecutive times. As the number of the existed communities is influenced by the threshold φ, we set $\varphi \geq 0.3$ for these data sets.

$$\varphi_{i,j}(t, t+1) = \frac{|V_t^i \cap V_{t+1}^j|}{|V_t^i \cup V_{t+1}^j|}, (i = 1, 2, \ldots, N_c(t), \quad j = 1, 2, \ldots, N_c(t+1)), \tag{1}$$

$\varphi_{i,j}(t, t+1) \in [0, 1]$. The parameter $\varphi_{i,j}(t, t+1)$ measures the similarity between community C_t^i and community C_{t+1}^j. The larger the value of $\varphi_{i,j}(t, t+1)$ is, the more similar these communities are. For a community C_t^i in G_t, if there is a community C_{t+1}^j ($j = 1, 2, \cdots, N_c(t+1)$) such that $\varphi_{i,j}(t, t+1) \geq \varepsilon$, the community C_t^i exists in the next timestamp $t+1$, where ε is the threshold to exist at timestamp $t+1$ for a community. In this Letter, the general parameter φ represents the similarity of two arbitrarily communities given at different timestamps.

As the number of existed communities at each timestamp is influenced by the threshold ε. The results are shown in Fig. 1, from which one can find that the number of existed communities N_c^e with $\varphi \geq \varepsilon$ is very small when $\varepsilon > 0.3$ for the data sets we used in this Letter. Therefore, we set $\varphi \geq 0.3$ to investigate the property of the new members in the existed communities for the data sets.

2.3. Theoretical hypothesis

In order to have an insight into justification of the observation on changing new members and communities, we provide three theoretical hypothesizes: Firstly, we assume that the community structures of networks are not influenced by detecting community algorithms. Secondly, as some users may leave system at a period of time and come back for reality system, in this Letter we only research the new members that have not existed in previous timestamps. Finally, there may be more than one similarity communities for a community at next timestamp. We suppose that the existed community is the community that has maximum similarity threshold.

3. The empirical analysis

3.1. Data sets

We introduce four data sets. The first one is the Facebook data of New Orleans network [25], which spans from 1 Sept., 2006 to 22 Jan., 2009. The timestamp of each link indicates the time when a pair of users become friends. We treat these interactions as undirected links and the period is set from 1 Sept., 2006 to 12 Dec., 2008. A node represents a user and a link represents a friendship between two users.

The second one is the author collaboration network, namely the American Physical Society (APS) [26] consisting of all papers published by journals of American Physical Society between 1893 and 2009. A node denotes a author, and the link between two nodes represents a common publication for two authors in the networks. Timestamp denotes the date of a paper publication. In this Letter, we only consider the papers published from 1960 to 2009.

The third one is Enron email network [27], which consists of 270,451 emails from 2000 to 2001. Nodes in the network are employees and links are emails. It is possible to send an email to oneself, and thus this network contains self-loops.

The last one is Wikipedia conflict network (Wiki) [28], which has 9,044 nodes and 38,059 links, spanning from 2004 to 2008. A node represents a user and a link represents a conflict between two users, with the link sign representing a positive or negative interaction.

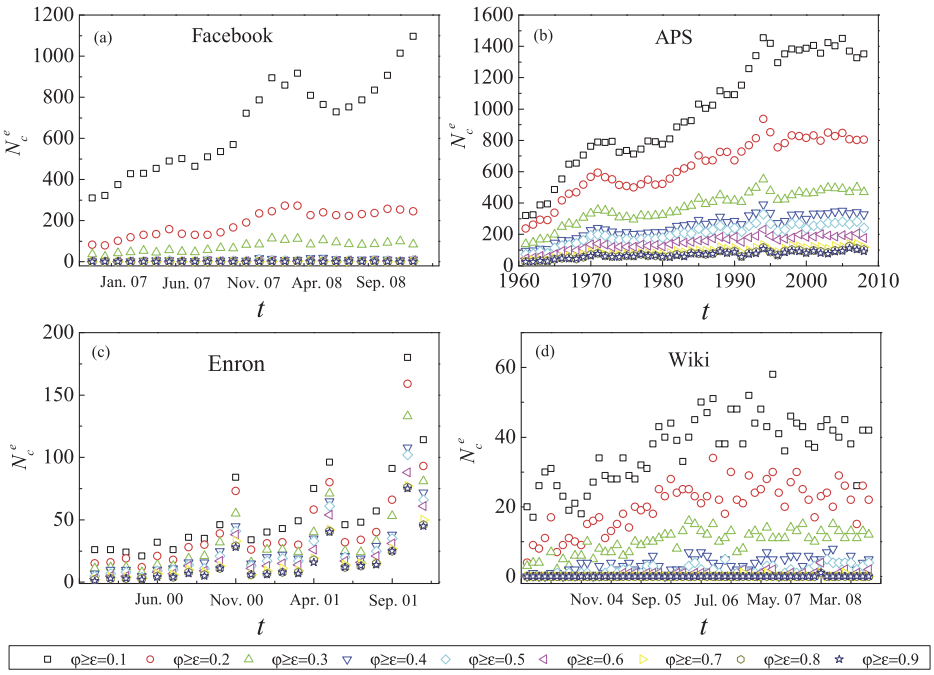

Fig. 1. (Color online.) The number of existed communities N_c^e with $\varphi \geq \varepsilon$ for the data sets. N_c^e represents the number of the existed communities with $\varphi \geq \varepsilon$, from which one can find that the number of the existed communities N_c^e is too small when $\varepsilon > 0.3$ for the data sets.

Table 1
The basic statistical properties of the Facebook, APS, Enron and Wiki data sets, where n and m denote the total number of nodes and links in the networks respectively.

Data sets	n	m	Time span
Facebook	63,731	817,190	2006.09–2008.12
APS	236,049	19,873,879	1960.01–2009.12
Enron	78,311	270,451	2000.01–2001.12
Wiki	9,044	38,059	2004.01–2008.12

We clean four networks by removing self-loops and duplicate links. Then we separate the Facebook network into slices with the interval of one month. The first slice is set from 1 Sept., 2006 to 30 Sept., 2006. The second one is set from 1 Oct., 2006 to 31 Oct., 2006, and the last one is set from 1 Dec., 2008 to 31 Dec., 2008. Similarly, the Enron and Wiki data sets are divided into 24 slices and 60 slices, respectively. And the APS data is divided into 50 slices in terms of one year from 1960 to 2009. The basic statistical properties of the data sets are shown in Table 1.

3.2. Measurements

To investigate the evolution properties of the community members, we present following measurements. The $n_{new}(t)$ is defined as the number of new members joining in the network at timestamp t. And we define $\rho(t)$ as the ratio between the number of new nodes and the number of nodes $|V_t|$ at timestamp t.

$$\rho(t) = \frac{n_{new}(t)}{|V_t|}. \tag{2}$$

In order to investigate the properties of communities new members joined in, we define the $\tau(t)$ as follows,

$$\tau(t) = \frac{N_c'(t)}{N_c(t)}, \tag{3}$$

where $N_c'(t)$ is the number of communities new members joint in, $N_c(t)$ represents the total number of communities at timestamp t.

3.3. The empirical results

3.3.1. The basic statistical features for new members

We apply the local community detecting algorithm, Blondel method [23], to identify disjoint communities for each timestamp. Firstly, the basic evolution features of new members are presented, including the number of new members and the communities they joined in.

From Fig. 2(a), we can find that the number of new members $n_{new}(t)$ increases from 1,008 to 5,075 for the Facebook data set. The value of $\rho(t)$ in the inset decreases rapidly from 0.51 to 0.14 and then keeps stable, which means that a small proportion of community members are new at the later timestamps. Besides, from Fig. 2(b), the number of new members $n_{new}(t)$ increases from 1356 to 9335. The ratio $\rho(t)$ approximately decreases slowly from 0.68 to 0.18 for the APS data set. From the Fig. 2(c) we can find that the number of the new members increases from 1319 to 8420 in Enron data set, and the ratio $\rho(t)$ decreases from 0.6337 to 0.3283. The number of new members grows from 372 to 4272, then reduces to 1216 for Wiki data set. The ratio $\rho(t)$ decreases from 0.6596 to 0.1345 from Fig. 2(d). As the number of the new members increases over time, the ratio between the number of new member and total members decreases.

Fig. 3 shows the number of communities $N_c'(t)$ the new members joined in, from which we can find that as the number of new members increases, they join in more communities. The ratio $\tau(t)$ decreases from 0.94 to 0.52 and then keeps stable for the Facebook data set, and the fraction $\tau(t)$ has the same tendency from 0.92 to 0.67 for the APS data set. In Fig. 3(c), one can find that the ratio $\tau(t)$ decreases from 0.97 to 0.8 for the Enron data set. While the ratio $\tau(t)$ keeps 0.92 for the Wiki data set.

3.3.2. The properties of new members with community evolution

In this section, we investigate the properties of new members with the community evolution in which they joined. Firstly, we investigate the size s of communities the new members joined in. We define the n' is the total number of the new members in specific size of communities for all timestamps. The results are shown

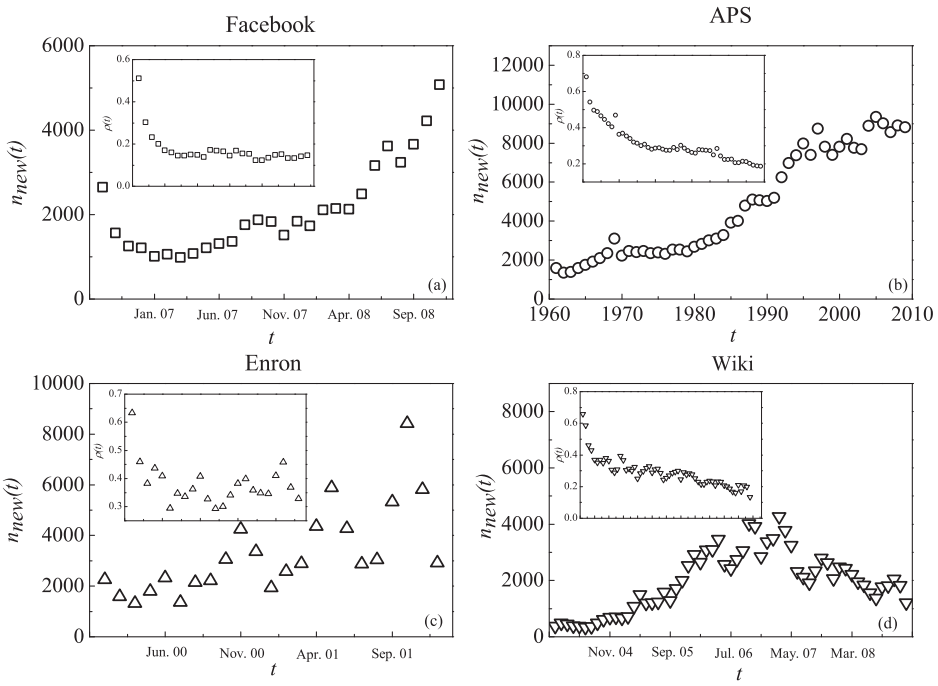

Fig. 2. The evolution property of the number of new members $n_{new}(t)$ and ratio $\rho(t)$ between the number of new members and the total members in the inset for the Facebook, APS, Enron and Wiki data sets, from which one can find that the number of new members $n_{new}(t)$ has the increasing tendency for the data sets, and the ratio $\rho(t)$ in the inset decreases for the APS and Wiki data sets and then keeps stable for the Facebook and Enron data sets.

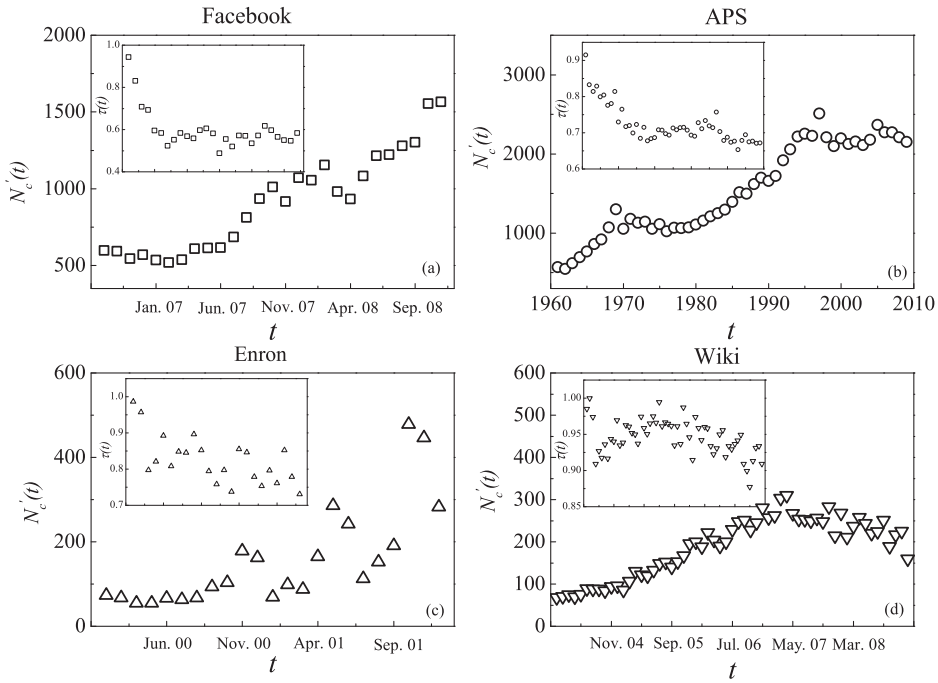

Fig. 3. The number of communities $N_c'(t)$ the new members joined in and the fraction $\tau(t)$ that the number of communities $N_c'(t)$ in which the new members joined accounts for the total number of communities $N_c(t)$ in the inset for time t on the Facebook, APS, Enron and Wiki data sets. The number of communities $N_c'(t)$ in which the new members joined has the increasing tendency and the value of $\tau(t)$ in the inset decreases firstly and then becomes stable for these data sets.

in Fig. 4, one can find that initially the tendency has decrease gradually for the four data sets, and then increase slowly for the Facebook, APS and Enron data sets. In order to analyze characteristics of these new members, we calculate the fraction of the new members who join in the different sizes of communities. We find that 51.54% of all the new members join in the small size community $s \leq 10$ for Facebook and 42.8% for APS, 32.86% for Enron and 27.28% for Wiki data set, suggesting that when a new member joins in the network he or she prefers to join in a small size community. At the same time, the fraction of the members that join in the large community size $s \geq 1000$ is 9.49%, 5.92%, 8.08%, 5.96% for Facebook, APS, Enron and Wiki data sets respectively, which indicates there are less new members joining in large-size communities.

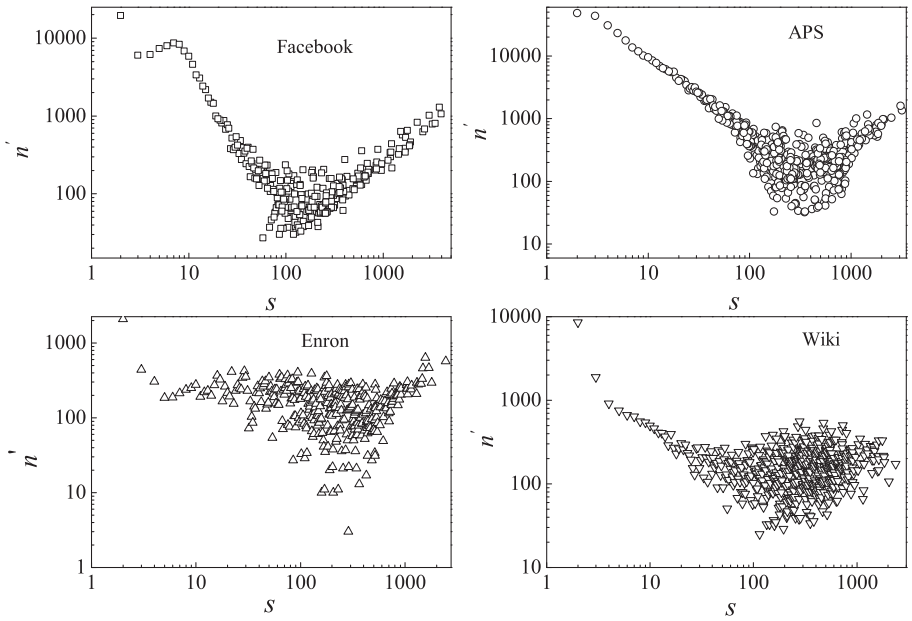

Fig. 4. Correlation between the community size s and the number of new members, denoted by n'. From which, we can find that although the tendency has decrease gradually and then increase slowly, the most of the new members join the small-size communities.

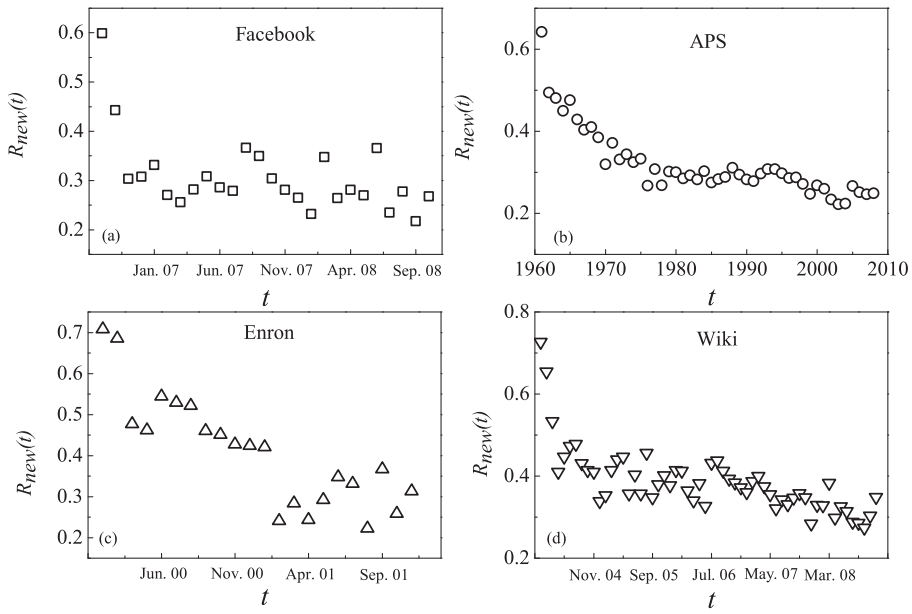

Fig. 5. The evolution property of the fraction of new members $R_{new}(t)$ to total members in the existed communities for the Facebook, APS, Enron and Wiki data sets, from which one can find that the proportion $R_{new}(t)$ decreases firstly and then keeps stable when $\varphi \geq \varepsilon = 0.3$ for these data sets.

To investigate the properties of the community evolution, we introduce the average ratio $R_{new}(t)$ between the new nodes and total nodes in the communities at each timestamp t when $\varphi \geq 0.3$ for the Facebook data set and the APS data set.

From Fig. 5, one can find that the average ratio $R_{new}(t)$ decreases firstly and then becomes stable for the data sets. The fraction $R_{new}(t)$ in the existed communities decreases from 60% to 25%, and keeps small for the Facebook data set when $\varphi \geq 0.3$ from the Fig. 5(a). For the APS data set, the average ratio $R_{new}(t)$ decreases from 64% to 30% in the initial stage and keeps stable later period from the Fig. 5(b). From Fig. 5(c)–(d), we can find that the average ratio $R_{new}(t)$ decreases from 70% to 30%, from 73% to 35% for Enron and Wiki respectively.

4. Conclusion and discussions

In this Letter, we investigated the evolution properties of community members for dynamic networks. We firstly separated data sets into different slices with same time interval, then calculated the number or the ratio of the new members. By using the Blondel method, we investigated the basic properties of community structure joined by the new members. Experimental results for the Facebook, APS, Enron and Wiki data sets showed that the number of new members for these data sets increases, but the ratio of the new members to all community members decreases from 0.51 to 0.14 and then keeps stable over time for the Facebook data set while it decreases from 0.68 to 0.18 for the APS data set, and the ratio decreases from 0.6337 to 0.3283 for Enron data set,

from 0.6596 to 0.1345 for Wiki data set. Furthermore, the number of communities in which the new members joint increases with the expansion of the networks. Besides, the ratio $\tau(t)$ between the number of communities in which the new members joined and the total number of communities at timestamp t is relatively stable and keeps 0.52, 0.67, 0.80 and 0.92 for the Facebook, APS, Enron and Wiki data set at later timestamp respectively. Then, we introduced a parameter φ to describe the community evolution. We investigated the ratio of the new community members when $\varphi \geq 0.3$ for the these data sets. The results indicated that the rate of the new members in the existed communities for these data sets has decrease trend and then keeps 25%, 30%, 30% and 35% respectively.

We analyze the evolution properties of the new community members, however, there are following problems to be resolved. Firstly, how different values of φ affecting the empirical results should be further investigated. In addition, tracing the community evolution is a challenge work, using the Markov process to describe the evolution of community structure [29] is an important method for this problem. Then, during the community evolution, how to explore the importance of nodes for dynamic networks [30–33] is also important problem to understand deeply the structure of networks. Finally, considering the characteristics of real networks, many of the networks are directed and the number of communities is unknown. How to effectively and quickly detect the community structure of the directed networks [34] and how to determine exactly how many communities in the networks should be addressed [35].

Acknowledgements

We thank Xiao-Lu Liu, Jian-Hong Lin and Xin-Yu Guo for useful comments and suggestions. This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 61374177, 71371125), JGL supported by The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shuguang Program Project of Shanghai Educational Committee (Grant No. 14SG42), and the Sino-Swiss Science and Technology Cooperation (No. 09-032016).

References

[1] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA 99 (2002) 7821.
[2] M.E.J. Newman, M. Girvan, Phys. Rev. E 69 (2004) 026113.
[3] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Nature 435 (2005) 814.
[4] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 103 (2006) 8577.
[5] Y. Pan, D.H. Li, J.G. Liu, J.Z. Liang, Physica A 14 (2010) 2849.
[6] A. Cuzzocrea, F. Folino, in: Proceedings of the Joint EDBT/ICDT 2013 Workshops, ACM, 2013, p. 93.
[7] M. Van Nguyen, M. Kirley, R. García-Flores, in: 2012 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2012, p. 1.
[8] J.G. Liu, Z.M. Ren, Q. Guo, D.B. Chen, PLoS ONE 9 (2014) e104028.
[9] L.Y. Tang, S.N. Li, J.H. Lin, Q. Guo, J.G. Liu, Int. J. Mod. Phys. C 27 (2016) 1650046.
[10] Z.L. Hu, Z.M. Ren, G.Y. Yang, J.G. Liu, Int. J. Mod. Phys. C 25 (2014) 1440013.
[11] P.J. Mucha, T. Richardson, K. Macon, M. Porter, J. Onnela, Science 328 (2010) 5980.
[12] L. Tang, H. Liu, J. Zhang, Z. Nazeri, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, p. 677.
[13] A. Cuzzocrea, F. Folino, in: Proceedings of the Joint EDBT/ICDT 2013 Workshops, ACM, 2013, p. 93.
[14] X. Lu, C. Brelsford, Sci. Rep. 4 (2014) 6773.
[15] A. Clementi, M. Di Ianni, G. Gambosi, E. Natale, R. Silvestri, Theor. Comput. Sci. 584 (2015) 19.
[16] M. Cordeiro, R.P. Sarmento, J. Gama, Soc. Netw. Anal. Min. 6 (2016) 1.
[17] G. Palla, A.L. Barabási, T. Vicsek, Nature 446 (2007) 664.
[18] D. Greene, D. Doyle, P. Cunningham, in: 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, p. 176.
[19] S. Asur, S. Parthasarathy, D. Ucar, ACM Trans. Knowl. Discov. Data 3 (2009) 16.
[20] L. Gauvin, A. Panisson, C. Cattuto, PLoS ONE 9 (2014) 1.
[21] X.G. Wang, Neurocomputing 168 (2015) 1037.
[22] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, p. 44.
[23] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, J. Stat. Mech. Theory Exp. 2008 (2008) P10008.
[24] M.E.J. Newman, Phys. Rev. E 69 (2004) 066133.
[25] B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, in: Proceedings of the 2nd ACM Workshop on Online Social Networks, ACM, 2009, p. 37.
[26] H.W. Shen, A.L. Barabási, Proc. Natl. Acad. Sci. USA 111 (2014) 12325.
[27] B. Klimt, Y. Yang, The Enron corpus: a new dataset for email classification research, in: Machine Learning: ECML 2004, Springer, 2004, p. 217.
[28] U. Brandes, J. Lerner, Structural similarity: spectral methods for relaxed blockmodeling, J. Classif. 27 (3) (2010) 279.
[29] L. Hou, X. Pan, Q. Guo, J.G. Liu, Sci. Rep. 4 (2014) 6560.
[30] J.G. Liu, J.H. Lin, Q. Guo, T. Zhou, Sci. Rep. 6 (2016) 21380.
[31] J.G. Liu, Z.M. Ren, Q. Guo, Physica A 392 (2013) 4154.
[32] Z.M. Ren, A. Zeng, D.B. Chen, H. Liao, J.G. Liu, Europhys. Lett. 106 (2014) 48005.
[33] J.G. Liu, Z.M. Ren, Q. Guo, B.H. Wang, Acta Phys. Sin. 62 (2013) 178901.
[34] K. Yang, X.L. Liu, Q. Guo, J.G. Liu, J. Univ. Electron. Sci. Tech. China 45 (2016) 1014.
[35] M.E.J. Newman, G. Reinert, preprint, arXiv:1605.02753, 2016.