Physica A 389 (2010) 5871–5877

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

The structural evolution of an online discussion network

Yang Yang [a,b,*], Qiang Chen [a,b], Wenjie Liu [a]

[a] School of Economics and Management, Tongji University, Shanghai 200092, China

[b] China Academy of Science and Technology Management, Tongji University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 27 October 2009
Received in revised form 13 July 2010
Available online 17 September 2010

Keywords:
Discussion network
Online social network
Complex network
Structural evolution

ABSTRACT

A discussion network is an important kind of social network, and it has been researched by many scholars in recent years. In this paper, we mainly studied its structural evolution based on an empirical study of a famous online discussion that happened in China in 2008. We found that the scale growth of the network shows an S shape, the degree distribution represents the power law in the first halfway, and the network shows a degree of disassortativity characteristic. We also classified the most active participants into different groups by their opinions and studied the structural evolution of opposite groups. It was observed that the evolution of nodes in each group were very similar, but the evolution of densities were obviously different. Specifically, we found that the most active participants preferred to converse with those belonging to the opposite groups at the beginning and tended to converse with anybody regardless of group as the discussion network grew. In the paper, these evolution patterns are revealed, and future lines of research are also considered.

1. Introduction

Discussion networks are very important for the exchange of diverse and often controversial viewpoints among citizens, and the exchange can motivate individuals to become more knowledgeable and participatory citizens [1]. Many scholars have studied the structures of various discussion networks, such as work-based or church-based discussion networks [1–3]. Recent research has focused on the discussion networks on the Internet, as they are becoming more and more important in modern democratic life [4,5]. Kelly et al. analyzed the structure and norms of political discussion networks in newsgroups, and they revealed the boundary between core authors and fringe authors [6]. Andreas et al. studied the emergence of different types of online discussion networks, and they found that the topics of the discussions are related to the time that users invest in those discussions, which may lead to these differences [7]. Himelboim et al. studied the structure of a political discussion group, which was found to be centralized. They found both that the structure of the discussion was hierarchal and that a few dominant participants (hubs) attracted a disproportionate number of responses [8].

Although many scholars have researched the structures of online discussion networks, detailed studies on their whole evolution processes are lacking. In this paper, we analyzed the structural evolution of a large-scale, online discussion about a teacher's behavior during the Sichuan earthquake in China. The discussion took place from May 25, 2008, to July 2, 2008, on *Tianya* (www.tianya.cn) (it was one of the largest online political forums in China), and it involved 9865 participants. We downloaded all 36,838 messages (including posts and replies) posted in this discussion, and we extracted 14,553 "peer–peer" messages, meaning messages containing an author's comment on another participant's previous messages. By regarding peer–peer messages as discussion relationships between two participants, a discussion network was constructed.

* Corresponding author at: School of Economics and Management, Tongji University, Shanghai 200092, China. Tel.: +86 15901828766; fax: +86 21 65984954.
E-mail address: yang_yang@tongji.edu.cn (Y. Yang).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.08.041

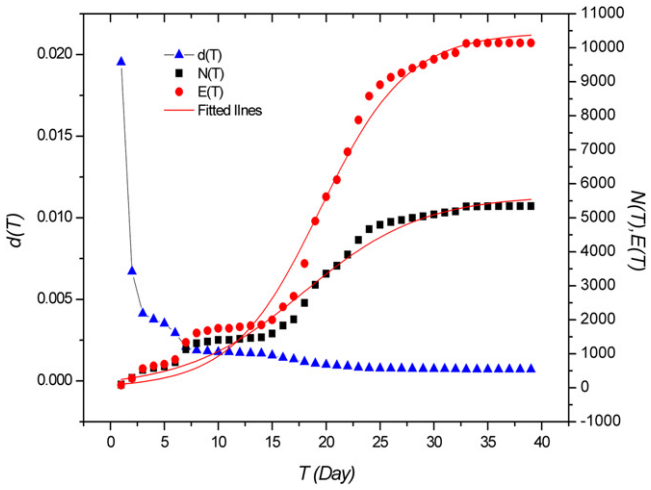

Fig. 1. Evolution of the numbers of nodes $N(T)$ and edges $E(T)$ and the network density $d(T)$ for the whole network from May 25, 2008, to July 2, 2008, with intervals of one day. Solid lines are the fitted curves by the logistic function. Least squares fitting gives: $m = 5659$, $a = 17.988$ and $s = 0.182$ with the adjusted R square value of 0.9798 for $N(T)$; and $m = 10451.84$, $a = 19.42$ and $s = 0.247$ with the adjusted R square value 0.9848 for $E(T)$.

The whole discussion network contained 5548 participants, and its structure features and evolution patterns were studied. It was found both that the scale growth of the network shows an S shape and that the network shows a degree of disassortativity characteristic, with the degree of assortativity coefficient increasing from -0.5 to -0.12. To reveal the behavior patterns of participants with different opinions, we extracted a sub-graph of the whole network that consists of the most active 181 participants classified into three groups by their opinions. We studied the characters of the two main oppositional groups in this sub-network; we found both that the evolution of inner-group discussions in each group were different and that the proportion of inter-group discussions between opposite groups converged from a high value to that of inner-group discussions as the scale grew.

2. Data sets

The *Tianya* forum is a big bulletin board with about 26,000,000 registered users and 500,000 messages (including posts and replies) submitted per day. After the Sichuan earthquake, Mr. Fan, a teacher at a middle school, posted an article in a *Tianya* forum that defended his choice to flee from his classroom during the disaster while leaving students behind. This article exploded into a large-scale discussion in the *Tianya* forum. At least 36,838 messages focusing on the topic had been posted by 9865 users from May 25 to July 2, 2008. Some participants thought that his behavior was the general reaction of an ordinary person and should not be condemned, while many others disagreed with this opinion. During the discussion, they opposed or supported each other by submitting messages commenting on one another's messages. So, it can be said that a discussion relationship (either supporting or opposing) exists between two participants when one of them indicates another's name and writes comments in his messages. If we view the participants as nodes V and the discussion relationships as edges E, an undirected and un-weighted discussion network $G(V, E)$ can be constructed from these messages.

We did not construct a directed network, like some similar studies on email networks or reply networks [9–11], because discussion just involves talking or writing about something from different viewpoints. Suppose that participant A posts a message with a comment on participant B's previous massages: they both have written from different views about one thing, regardless of whether or not B replies to A, and a peer–peer discussion has happened between them.

3. Structural evolution

We extracted 39 snapshots of the discussion using intervals of one day from May 25 to July 2 and thus investigated the evolution of the network. Fig. 1 shows the growth of the numbers of nodes and edges and the variation of network density over time. The density of a network is defined as the ratio of the number of real edges E to the number of total possible edges $N(N-1)/2$ (N is the number of nodes).

In Fig. 1, the growth pattern for the network scale, including the numbers of nodes and edges, shows an S shape that can be simulated by the logistic function $S(T) = \frac{m}{1+e^{-(T-a)s}}$. This growth pattern was also reported in a recent work by Hu and Wang [12] about the evolution of the large OSN (online society network) [11]. However, in this discussion network the evolution curves are not as smooth as those in the OSN, especially in the initial stage from the 1st day to the 15th day. The density of the network decreased with the growth of scale, as the average edges per node remains constant during the evolution. This is different from either the increasing pattern found in an email network [13] or the non-monotone pattern

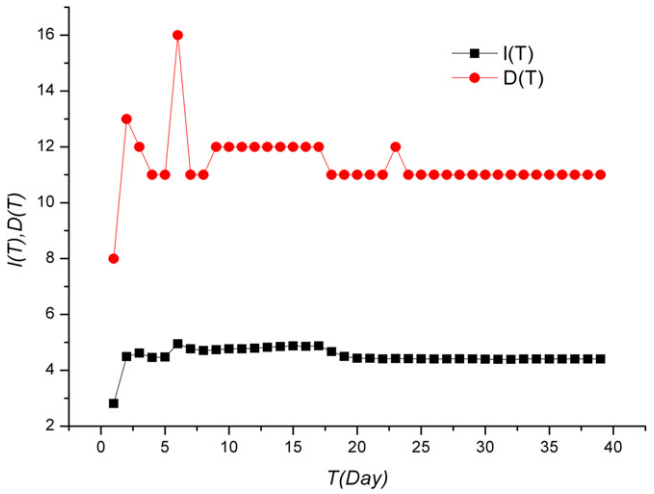

Fig. 2. Evolution of the average length $l(T)$ and diameter $D(T)$ of the largest connected sub-graph.

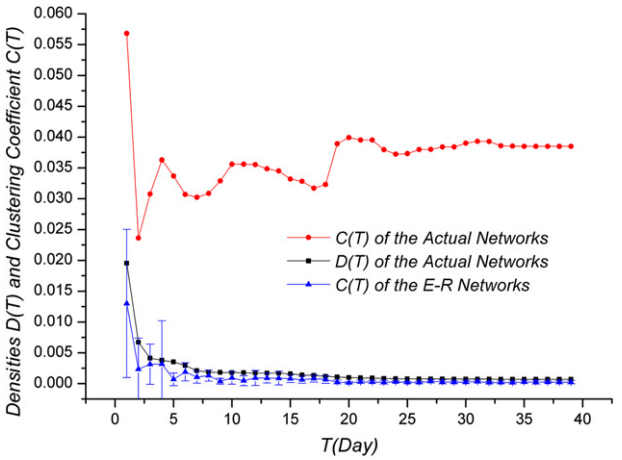

Fig. 3. Evolution of the clustering coefficient $C(T)$ and the density $d(T)$. The mean clustering coefficient $C(T)$ of ten simulated E–R networks with the same size of the actual network each day is also plotted for comparison.

found in OSNs [12–14]. One possible reason for this is that, for most participants, the interest in a topic is limited. Therefore, one participant only converses with a few people, and the density declines quickly with more people joining in.

Fig. 2 shows the average path length and the diameter of the largest connected sub-graph of the network. As opposed to that observed in OSNs [12–14], the evolution of these two indicators show three different stages: in the first stage, l and D grow fast; in the next stage, they are almost constants; in the third stage, they decrease quickly to final values and remain stable.

Generally, there are many connected triads in a social network representing a high degree of transitivity, which can be quantified by the clustering coefficient [15]. However, in this discussion network, we found the average clustering coefficient (0.036) to be much lower than that of OSN networks (0.16–0.31) observed by Chun et al. [16], but higher than that of randomized networks (near 0 with scale growth). This weak transitivity means that if there is a discussion relationship between participant *A* and *B*, participant *A* cares little about others' comments on *B*'s opinion. Fig. 3 shows the evolution of the clustering coefficient compared with that of the density. The evolution of the clustering coefficient in E–R random networks with the same size of the empirical networks is also plotted.

In the research on the OSN *Wealink*, a strong positive correlation between the clustering coefficient and the density in the network was observed. However, in Fig. 3, there is no obvious evidence that these two metrics are analytically connected to each other.

Fig. 4 shows the evolution of the degree distribution of the discussion network by each day. We used the maximum likelihood estimators (MLR) to estimate the degree exponents [16] and found that the degree exponent kept decreasing from 2.7 to 1.8 during the whole period. To verify the power law distribution, we used the Kolmogorov–Smirnov test, and

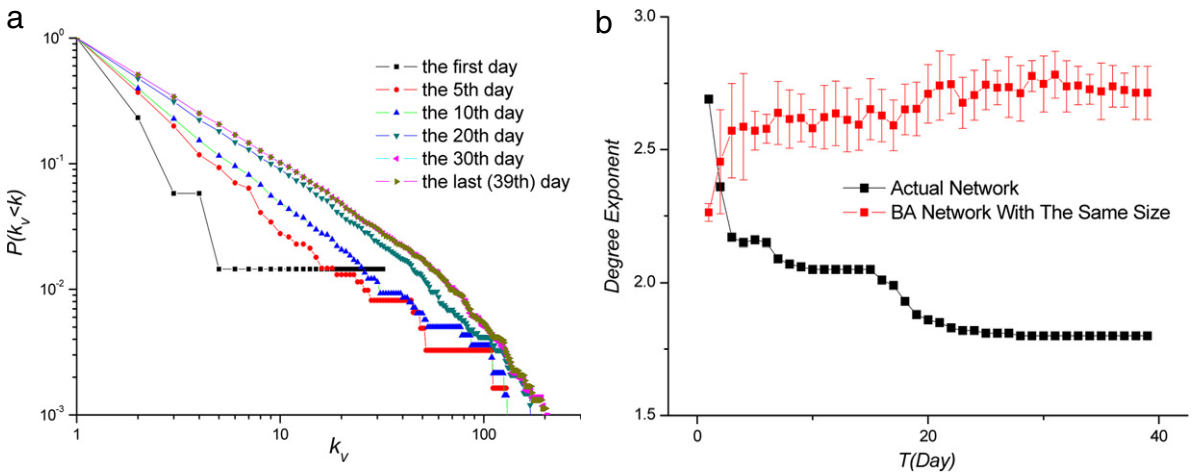

Fig. 4. Evolution of the degree distribution. (a) The cumulative degree distribution $P(k)$ of six days from the first day to the last day in the semi-logarithmic coordinate. The distribution of the 30th day and that of the 39th day are so close that we can hardly differentiate between them in the figure. (b) The change of the degree exponent during the observed period, and the mean value of ten simulated BA model networks of the same size as the actual networks for comparison.

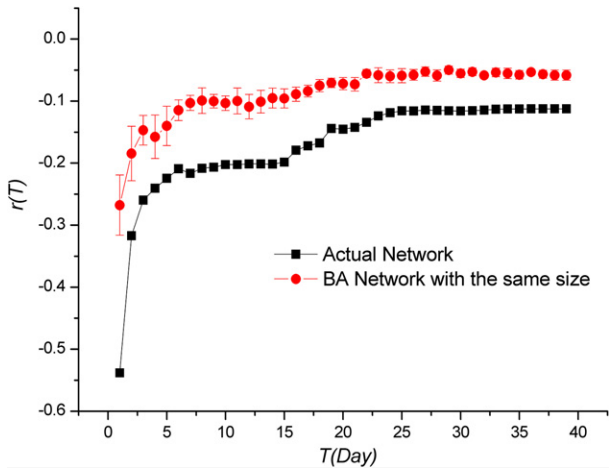

Fig. 5. Evolution of the assortativity r in the discussion network, compared with the mean value of ten simulated *BA* model networks of the same size.

the result shows that a power law distribution can be supported to a significance level of 0.05 in the first 18 days. However, in the following days, the actual distribution looks like a power law but cannot be supported by the K–S test with the accepted significance level. The cumulative degree distributions of six dates are shown in Fig. 4(a). The change of the degree exponents of the power law distribution is plotted in Fig. 4(b), as compared with that of the simulated BA model network of the same size [17]. The decreasing of the degree exponents of the actual network is obvious when compared to the BA model network.

The correlation existing between the degrees of adjacent nodes is often characterized by the assortativity r and defined as the Pearson correlation coefficient of the degrees of either node connected by a link [15]. In recent works, many online social networks are reported to be disassortative. In our research, the online discussion network also showed a disassortative character that is weakened with the scale of network growth. Fig. 5 shows the evolution of the assortativity r in this network. For comparison, the mean assortativity values of ten simulated BA networks (generally tends to be zero as the network becomes large [18]) of the same size is also plotted.

In Fig. 5, the assortativity r keeps increasing from -0.53 to -0.12 and then remains stable. The final value of r (-0.12) is very similar to that of other online communities, such as Cyworld (-0.13), nioki (-0.13) and Gnutella P2P (-0.10) [19–21]. This means that the wider scope and lower cost of online communication gives more chances for ordinary people to talk with opinion leaders in public discussion. This is an important reason why online discussions can contribute significantly to democratic life. Furthermore, the monotone increasing assortativity reveals that the interactivity between the ordinary people and the opinion leader keeps shrinking during the discussion. That means the function and behavior of the opinion leader may vary in different phases of the discussion.

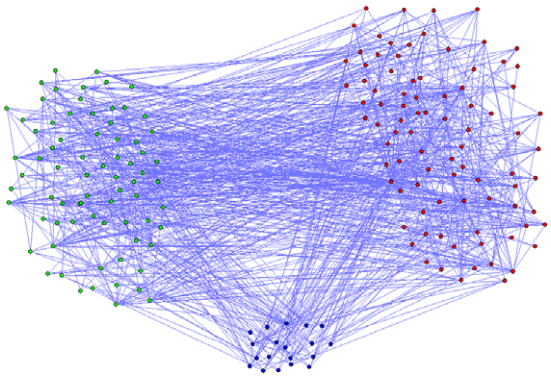

Fig. 6. The discussion network of the most active participants. Green nodes represent the participants who tolerated Mr. Fan, red nodes represent the participants who blamed him, and blue nodes represent the neutral participants.

4. Characters of different groups

Generally, participants in a discussion can be divided into different groups by their opinions. To disclose the characters of each group, we determined the 181 most active participants who had written at least 30 messages (including posts and replies) during the discussion, and we identified their opinions by manually reading all of their messages. We found that these 181 participants could be classified into three groups, including tolerating, blaming and neutral attitudes. There were 75 participants who tolerated Mr. Fan, 87 who blamed Mr. Fan, and 19 who retained neutrality. Fig. 6 shows the topology of the sub-network of these active participants. It was also found that nobody changed his or her opinion during the discussion.

Since most of these 181 participants belonged to the tolerating and blaming groups, we focused our research mainly on these two opposite groups and compared their evolutions in Fig. 7. It seems that the growth patterns of both groups also showed S shapes, similar to those shapes in Fig. 2, and the numbers of nodes in both groups are very close at all times, as shown in Fig. 7(a).

We classified the edges into two types: the inter-group edges, whose two end-points belong to different groups, and the inner-group edges, whose end-points come from the same group. In Fig. 7(c) and Fig. 7(e), it is shown that the number of inner-group edges in the blaming group is obviously higher than that in the tolerating group for most times, as is likewise true for the densities of the inner-group edges $d = \frac{2E_{ii}}{N_i(N_i-1)}$ (E_{ii} is the number of inner-edges in group i, and N_i is the number of nodes in group i). By the analyses on the contents of all messages, we found that most of the blaming participants demonstrated their opinions from the same point of view, that is, the professional morality of a teacher. In contrast, the tolerating participants' arguments varied much more, including aspects of culture, human rights, private freedom, etc. We think this scatter of the arguments makes it difficult for participants in the tolerating group to cite each other, which causes a lower density of inner-group edges.

As shown in Fig. 7(f), the final assortativity of this sub-network is very close to that of the whole network shown in Fig. 5. However, the final clustering coefficient $C \approx 0.197$ is obviously higher than that of the whole network shown in Fig. 3 and is close to the values found in many other OSNs.

We also studied the proportions of inter-group edges between opposite groups in all edges to find out whether participants preferred to discuss with those in the opposite group. If participants selected their discussion partners regardless of which groups they belonged to, the proportion of inter-group edges should be close to the random selection probability, where m and n are the counts of participants in each group. In Fig. 8, we compare the actual proportions of each kind of edge with the random selection probability.

In Fig. 8, the actual proportion of inter-group edges is obviously higher than the random selection probability at the beginning and converges with the latter as the scale of the discussion explodes. This means that people are more interested in opposite opinions and prefer to debate with others when a new discussion topic emerges. However, as the discussion became wider and deeper, they had to cite more words from people with the same opinion as supporting evidence, which caused the proportion of discussions between different groups to decrease. It is interesting that the actual proportion of inner-group edges in the tolerant group is always lower than the random selection probability. The reason for this is discussed in Fig. 7(c).

Generally, we believe that the power of consumers to filter information on the Internet will limit their exposure to topics and points of view of their own choosing [22], which causes group polarization on the Internet. While Fig. 8 shows that even the interactions between groups are very frequent, no one has changed his opinion during the discussion either. Further researches on the phenomena may disclose more details of group polarization.

5. Summary and future work

In this paper, we studied the structural evolution of an online discussion network. We found both that the scale growth of the network shows an S shape and that the network density is descending. We also showed the degree of disassortativity

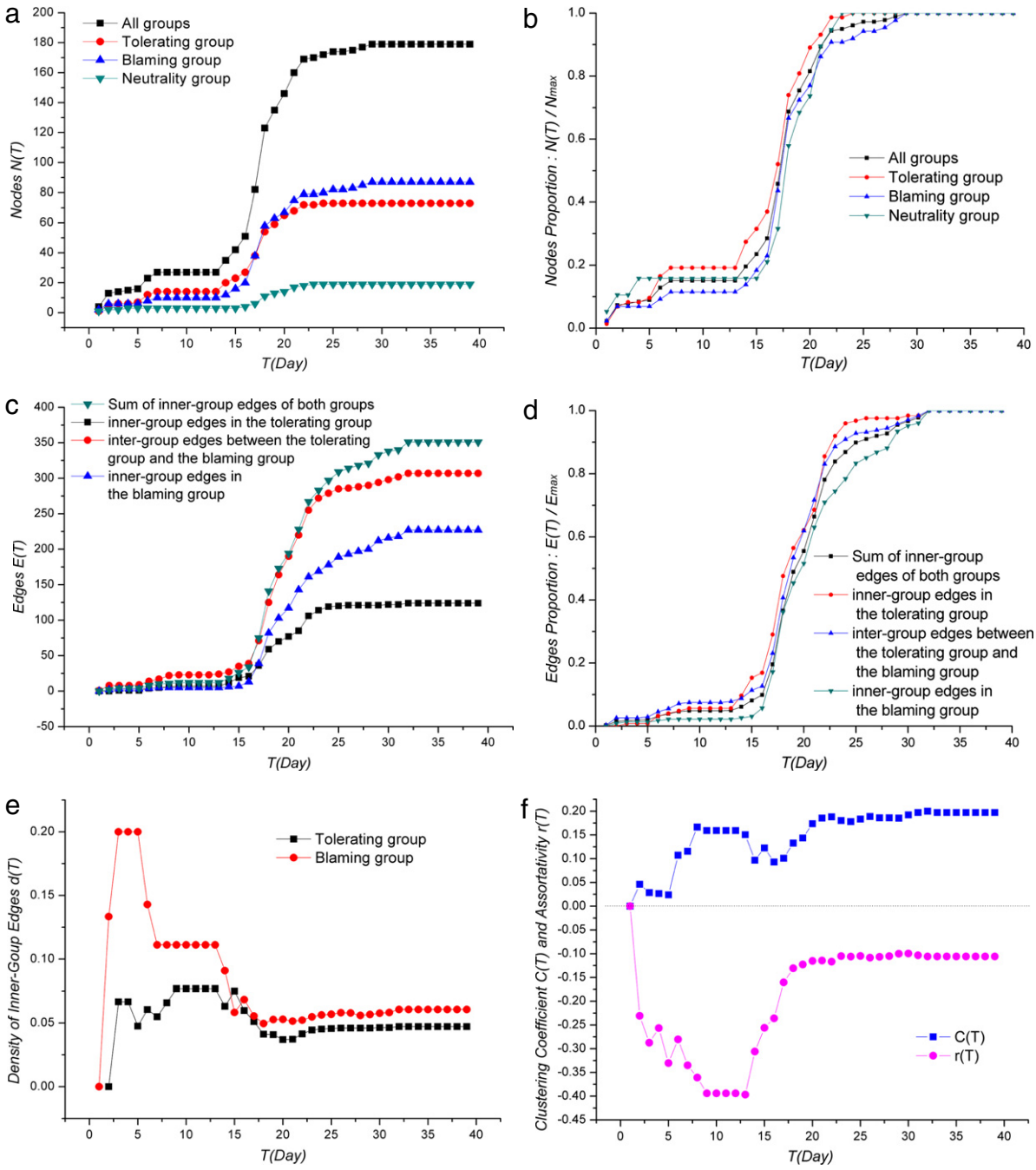

Fig. 7. (a) and (b) are the growths of the node numbers $N(T)$, and the node proportions $N(T)/Nmax$ of each group. (c) and (d) are the growths of each kind of edge, including inter-group edges between the tolerating group and the blaming group and inner-group edges in both groups. (e) is the densities of inner-group edges in both groups. (f) is the evolution of the clustering coefficient $C(T)$ and the assortativity r of the sub-network.

of the network. To disclose the behavior patterns of participants with different opinions, we classified the 181 most active participants into three groups by their opinions. The counts of nodes in both tolerating and blaming groups were found to be very close, while the counts and densities of inner-group edges in each group were obviously different. Specifically, we found that the proportion of discussions between opposite groups decreased from a high value to the random selection probability.

There are still many detailed features about online discussion networks that need to be revealed in the future. Based on the findings reported in this paper, we are now studying the preferential connectivity of new participants selecting their discussion partners. Our preliminary research shows that this may be different than some existing models on the probability

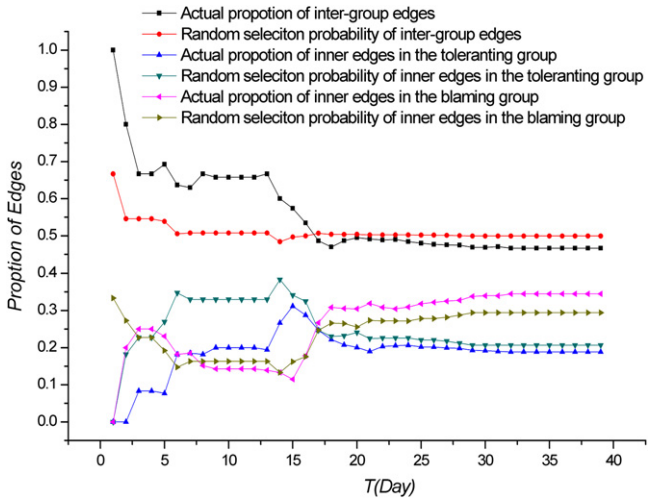

Fig. 8. The evolution of the actual proportions of inter-group edges compared with the random selection probability.

distribution [17]. We will model the special preferential connectivity of this online discussion network, and we will report it in future papers.

Acknowledgements

This work was supported by the Innovation Program of the Shanghai Municipal Education Commission.

References

[1] D.A. Scheufele, M.C. Nisbet, D. Brossard, Annual Meeting of the International Communication Association, New Orleans, 2004.
[2] Robert Huckfeldt, Paul E. Johnson, John Sprague, Journal of Politics 64 (2002) 1.
[3] Scott D. McClurg, American Journal of Political Science 50 (2006) 737.
[4] Brundidge Jennifer, Annual Meeting of the International Communication Association, Dresden, 2006.
[5] Wang Xiuli, Proceedings of the NCA 93rd Annual Convention, Chicago, 2007.
[6] John W. Kelly, Danyel Fisher, Marc Smith, ACM International Conference Proceeding Series, vol. 151, pp. 412–417.
[7] Andreas Kaltenbrunner, Sandra Gonzalez-Bailon, Rafael E. Banchs, Proceedings of the WebSci 09: Society On-Line, Athens, 2009.
[8] Himelboim Itai, M. Lavelle Shawn, Kafri Ran, Proceedings of the International Communication Association 55th Annual Conference, New York, 2005.
[9] Wang Ke, Haibo Hu, Wang Xiaofan, Complex Systems And Complexity Science 5 (2008) 66 (in Chinese).
[10] Zhongbao Kou, Changshui Zhang, Physical Review E 67 (2003) 036117.
[11] Sangman Han, Jun Kim Beom, Physica A 387 (2008) 5946.
[12] Haibo Hu, Xiaofan Wang, Physics Letters A 373 (2009) 1105.
[13] J. Leskovec, J. Kleinberg, C. Faloutsos, ACM Trans. Knowl. Discov. Data 1 (2007) 1.
[14] R. Kumar, J. Novak, A. Tomkins, In the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2006.
[15] M.E.J. Newman, Physical Review Letters 89 (2002) 208701.
[16] Aaron Clauset, Cosma R. Shalizi, M.E.J. Newman, SIAM Review 51 (4) 661.
[17] Albert-Laszlo Barabasi, Reka Albert, Science 286 (1999) 509.
[18] M.E.J. Newman, SIAM Review 45 (2) (2003) 161.
[19] H. Chun, H. Kwak, Y.H. Eom, Y.Y. Ahn, S. Moon, H. Jeong, 8th ACM SIGCOMM Conference on Internet Measurement, ACM Press, New York, 2008.
[20] P. Holme, C.R. Edling, F. Liljeros, Social Netw. 26 (2004) 155.
[21] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, 7th ACM SIGCOMM Conference on Internet Measurement, ACM Press, New York, 2007.
[22] Cass Sunstein, The Boston Review, 2001.