Modeling online social networks based on preferential linking*

Hu Hai-Bo(胡海波)[a)†], Guo Jin-Li(郭进利)[b)], and Chen Jun(陈 骏)[a)]

[a)] *Department of Management, East China University of Science and Technology, Shanghai* 200237, *China*
[b)] *School of Management, University of Shanghai for Science and Technology, Shanghai* 200093, *China*

(Received 5 April 2012; revised manuscript received 9 May 2012)

We study the phenomena of preferential linking in a large-scale evolving online social network and find that the linear preference holds for preferential creation, preferential acceptance, and preferential attachment. Based on the linear preference, we propose an analyzable model, which illustrates the mechanism of network growth and reproduces the process of network evolution. Our simulations demonstrate that the degree distribution of the network produced by the model is in good agreement with that of the real network. This work provides a possible bridge between the micro-mechanisms of network growth and the macrostructures of online social networks.

Keywords: online social network, preferential linking, model, power law

PACS: 89.65.–s, 89.75.Hc, 02.50.–r **DOI:** 10.1088/1674-1056/21/11/118902

1. Introduction

Web 2.0, which is characterized by social collaborative technologies, such as social networking sites (SNS), blogs, wiki, video or photo sharing, and folksonomy, has attracted much attention from researchers of diverse disciplines.[1] As a fast growing business, many SNSs of different scopes and purposes have emerged on the web, many of which, such as Facebook,[2,3] Renren,[4] MySpace, and Orkut,[5] are among the most popular sites on the web. Users of those sites, by establishing friendship relationships with other users, can form online social networks (OSNs). The popularity of those sites and the availability of network data sets offer an invaluable opportunity to study the structure and the evolution of OSNs at scale. It is believed that having a proper understanding of network evolution enables the prediction of the future growth of networks and the exploration of human behavior in virtual communities.[6–8]

Recently, the evolution of OSNs has been extensively investigated by physicists, computer scientists, and sociologists. Holme *et al.* studied the evolution of Pussokram and found that its degree correlation coefficient is always negative over time,[9] which is in stark contrast with the significant assortative mixing of the real-world social networks. Viswanath *et al.* studied the evolution of the activity network of Facebook and found that the mean degree, the clustering coefficient, and the average path length are all relatively stable over time.[10] Hu and Wang studied the evolution of Wealink[11] and found that many network properties show obvious non-monotone features, including a sigmoid growth of the network scale which was also observed by Chun *et al.* in Cyworld,[12] and a transition from the degree assortativity characteristic of real social networks to the degree disassortativity characteristic of many OSNs which was also observed by Szell and Thurner in Pardus.[13]

Excepting empirical research, many network models have also been proposed to reproduce the macrostructure of OSNs or social networks in real life. In the traditional field of social networks, exponential random graph models (ERGMs) have been proposed to model and fit the empirical network data.[14] For example, Traud *et al.* examined the relative importance of different user attributes to the social organization of large institutions at different institutions in Facebook by using ERGMs.[3] Recently, due to the availability of the data of large-scale social networks, some more specific models have also been brought forward to reproduce some characteristics of the so-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 61104139, 70871082, and 71101053) and the ECUST for Excellent Young Scientists, China.

†Corresponding author. E-mail: sdhuzi@163.com

© 2012 Chinese Physical Society and IOP Publishing Ltd

cial networks. Hu and Guo proposed an analyzable model, which captures the real evolution of Facebook networks and can well reproduce their homogeneous degree distributions.[15] Grindrod and Parsons introduced a class of models for the evolving social networks using evolving graphs with memory dependent edges, and the model parameters could be calibrated based on the limited samples of time dependent data.[16] Zou *et al.* proposed an evolving network model growing fast in units of module.[17] Each module was a small-world network containing several interconnected vertices, and the vertices between the modules were linked by preferential attachments. Motivated by the evolution of mobile users, Guo *et al.* proposed a directed network model, in which the vertices grew following a power-law acceleration.[18]

Uncovering how the micro-mechanisms of network growth lead to the macrostructures of OSNs is of paramount importance in understanding the evolution of OSNs. However it is very difficult to capture this process due to the fact that the detailed empirical data of network growth with time labels integrating the joining of new users and the establishment of new friend relationship are still scarce. To date, there are few analytical studies on the verifiable mechanisms dominating the evolution of OSNs. Most of the understanding comes from the theoretical models which lack validation utilizing empirical data. In this paper, based on the empirical network data, we will bring forward an analyzable model, which can reproduce the process of network growth and connect the users' behavior and the network structure.

2. Data sets

We first study Wealink, a large SNS in China, whose users are mostly professionals, typically, businessmen and office clerks.[11] The network data, logged from 0:00:00 on 11 May 2005 (the inception day for the Web 2.0 site) to 15:23:42 on 22 August 2007, include all friendship relationships and the time of formation of each tie.

The final data format is a time-ordered list of triples $\langle U_i, U_j, T_k\rangle$. Triple $\langle U_i, U_j, T_k\rangle$ indicates that user U_i sends a link request to user U_j or U_i accepts U_j's previous request and they become friends at time T_k. Like Facebook and YouTube, the friendship relationships are established only when the invitations are accepted. The online community is a dynamically evolving one with new users joining the network and new ties established between users.

3. Preferential linking

Like many other OSNs, the degree distribution of Wealink shows the power-law feature.[11] This kind of distribution can be produced through the linear preferential attachments, i.e., new users tend to attach to already popular old users. However in the real scenario of network growth, when new users establish friendship relationships with old ones, or new ties are established between old users, the old users with large degrees of distribution are all likely to be preferentially selected.

To test the preference features of different types of link establishments, we separate the preferential linkings into three aspects: preferential acceptance, preferential creation, and preferential attachment. The preferential acceptance implies that the larger an old user's degree is, the more likely he/she will accept link invitations from the other old users. The preferential creation implies that the larger an old user's degree is, the more likely his/her link invitations will be accepted by the other old users. And the preferential attachment implies that new users tend to attach to already popular old users with large degrees.

Let k_i be the degree of user i. The probability with which user i with degree k_i is chosen can be expressed as

$$\Pi(k_i) = \frac{k_i^\beta}{\sum_j k_j^\beta}. \tag{1}$$

We can compute the probability $\Pi(k)$ with which an old user of degree k is chosen, and it is normalized by the number of users of degree k that exist just before this step

$$\Pi(k) = \frac{\sum_t [e_t = v \wedge k_v(t-1) = k]}{\sum_t |\{u : k_u(t-1) = k\}|} \sim k^\beta, \tag{2}$$

where $e_t = v \wedge k_v(t-1) = k$ represents that at time t the old user whose degree is k at time $t-1$ is chosen. We use $[\cdot]$ to denote a predicate (which takes a value of 1 if the expression is true, else 0). Generally, $\Pi(k)$ has significant fluctuations, particularly for large k. To reduce the noise level, instead of $\Pi(k)$, we study the cumulative function

$$\kappa(k) = \int_0^k \Pi(k)\mathrm{d}k \sim k^{\beta+1} = k^\alpha. \tag{3}$$

Figure 1 shows the relationship between degree k of users and preference metric κ. The least-square linear regression gives $\alpha = 1.93 \pm 0.01$ ($R^2 = 0.99$) for the preferential creation, $\alpha = 1.97 \pm 0.01$ ($R^2 = 0.98$) for the preferential attachment, and $\alpha = 2.06 \pm 0.01$ ($R^2 = 0.99$) for the preferential acceptance. All are with significance level $p < 0.001$. Thus $\beta \approx 1$ indicates the linear preference.

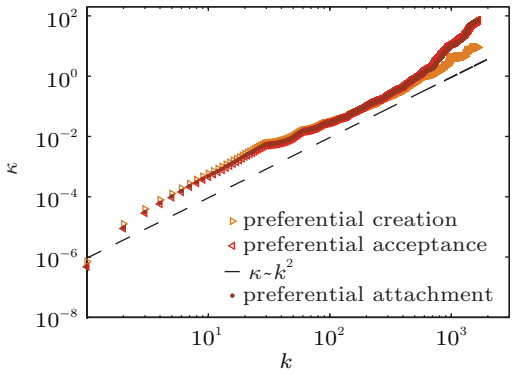


Fig. 1. (colour online) Preference characteristics in the evolution of OSN. The slopes for preferential creation, preferential attachment, and preferential acceptance are 1.93, 1.97, and 2.06, respectively.

4. Model

Like the other OSNs, the evolution of Wealink includes two processes. The first one is that a new user joins in the network and establishes a friend relationship with an old user. The second one is that a friend relationship is established between two old users. Certainly, there exists the case that a tie forms between two new users; however this situation is rare in the real world and can be neglected.

Based on the linear preference, we bring forward a realistic network model. Starting with a small connected network with m_0 vertices, at every time step, there are two alternatives. (i) With probability p, we add a new vertex with m_1 edges connected to the vertices already present in the network. The probability that the new vertex will be connected to old vertex i with degree k_i is $\Pi(k_i) = k_i / \sum_j k_j$, i.e., the linear preference. (ii) With probability $q = 1 - p$, we add m_2 new edges connecting the old vertices. The two endpoints of each edge are chosen according to the linear preference.

After t steps, the model leads to a network with a mean number of vertices $N(t) = m_0 + pt$. For a large t, $N \approx pt$, and the total degree of the network is $k_{\rm all}(t) \approx 2t(pm_1 + qm_2)$. Let $P(k)$ be the degree distribution of the model and $N_k(t) = ptP(k)$ be the mean number of vertices with degree k, then the rate equation for the time evolution of $N_k(t)$ is

$$\frac{{\rm d}N_k}{{\rm d}t} = pm_1 \frac{(k-1)N_{k-1} - kN_k}{\sum_j jN_j} + 2qm_2 \frac{(k-1)N_{k-1} - kN_k}{\sum_j jN_j} + p\delta_{km_1}. \quad (4)$$

Since $\sum_j jN_j \approx 2t(pm_1 + qm_2)$, thus

$$\begin{aligned} & pP(k) \\ = {} & p^2 m_1 \frac{(k-1)P(k-1) - kP(k)}{2(pm_1 + qm_2)} \\ & + pqm_2 \frac{(k-1)P(k-1) - kP(k)}{pm_1 + qm_2} + p\delta_{km_1}. \end{aligned} \quad (5)$$

When $k \gg 1$, we obtain

$$\begin{aligned} & P(k)\left[\frac{p^2 m_1 + 2pqm_2}{2(pm_1 + qm_2)}k + p\right] \\ = {} & P(k-1)\left[\frac{p^2 m_1 + 2pqm_2}{2(pm_1 + qm_2)}(k-1)\right], \end{aligned} \quad (6)$$

i.e.,

$$P(k)\left[k + \frac{2(pm_1 + qm_2)}{pm_1 + 2qm_2}\right] = P(k-1)\left[k - 1\right]. \quad (7)$$

Let

$$\begin{aligned} a &= -1, \\ b &= \frac{2(pm_1 + qm_2)}{pm_1 + 2qm_2}, \end{aligned} \quad (8)$$

accordingly equation (7) becomes

$$P(k) = \frac{k + a}{k + b} P(k-1). \quad (9)$$

Since b is not an integer, it is difficult to obtain $P(k)$ in a simple form. From Eq. (9), we obtain

$$\ln \frac{P(k)}{P(k-1)} = \ln \frac{k + a}{k + b}, \quad (10)$$

which can be well approximated by $P(k) \sim (k + c)^{-\gamma}$.[19] The exponent γ is

$$\gamma = \ln\left(\frac{k + a}{k + b}\right) \Big/ \ln\left(\frac{k + c - 1}{k + c}\right), \quad (11)$$

i.e.,

$$\begin{aligned} & \ln\left(1 + \frac{a}{k}\right) - \ln\left(1 + \frac{b}{k}\right) \\ = {} & \gamma\left[\ln\left(1 + \frac{c-1}{k}\right) - \ln\left(1 + \frac{c}{k}\right)\right]. \end{aligned} \quad (12)$$

For large k, implementing the Maclaurin series expansions of degree 2, we can obtain

$$\gamma = b - a = \frac{3pm_1 + 4qm_2}{pm_1 + 2qm_2},$$

$$c = \frac{a+b+1}{2} = \frac{pm_1 + qm_2}{pm_1 + 2qm_2}. \quad (13)$$

Thus, degree distribution $P(k)$ is

$$P(k) \sim (k+c)^{-\gamma} = \left(k + \frac{pm_1 + qm_2}{pm_1 + 2qm_2}\right)^{-\frac{3pm_1+4qm_2}{pm_1+2qm_2}}. \quad (14)$$

The tail of the distribution is $P(k) \sim k^{-\frac{3pm_1+4qm_2}{pm_1+2qm_2}}$. When $p = 1$, the model is reduced to the Barabási–Albert model. The model can produce networks with a power-law degree distribution and reveal the mechanism governing the evolution of Wealink.

The linear preference can be generalized to describe the evolution of general OSNs. Mislove *et al.* studied the evolution of Flickr and found that the users tend to create and receive links in proportion to their outdegrees and indegrees, respectively.[20] Leskovec *et al.* studied the evolutions of Flickr, del.icio.us, Yahoo!Answers, and LinkedIn, and examined whether the preferential attachment holds for those networks.[21] They found that Flickr and del.icio.us show the linear preference, and Yahoo!Answers show the slightly sublinear preference $\Pi(k) \sim k^{0.9}$. In LinkedIn, for low degrees, $\Pi(k) \sim k^{0.6}$; however, for large degrees, $\Pi(k) \sim k^{1.2}$, indicating the superlinear preference. Although there are minor differences in the exponents β for different networks, we can see that $\beta \approx 1$, implying that the linear preference may be universal for OSNs. Thus, the proposed model provides a theoretical foundation for the comprehension of the evolution of OSNs and reveals how the micro-mechanisms of the preferential linking result in the networks with power-law degree distribution.

Apparently, a larger β leads to a greater heterogeneity of the network structure. Figure 2 shows the numerical result obtained by averaging over 10 independent realizations. If rule (ii) of the model is unchanged, and in rule (i), we suppose that the probability with which the new vertex will be connected to old vertex i with degree k_i is $\Pi(k_i) \sim k_i^{\beta}$, the corresponding numerical results for different β are shown in Fig. 2(a). If rule (i) is unchanged, and in rule (ii), we suppose that an endpoint is chosen according to $\Pi(k_i) \sim k_i^{\beta}$, the corresponding numerical results are shown in Fig. 2(b). We find that when β increases, $P_c(k)$ gets more and more decentralized, and hubs emerge for larger β.

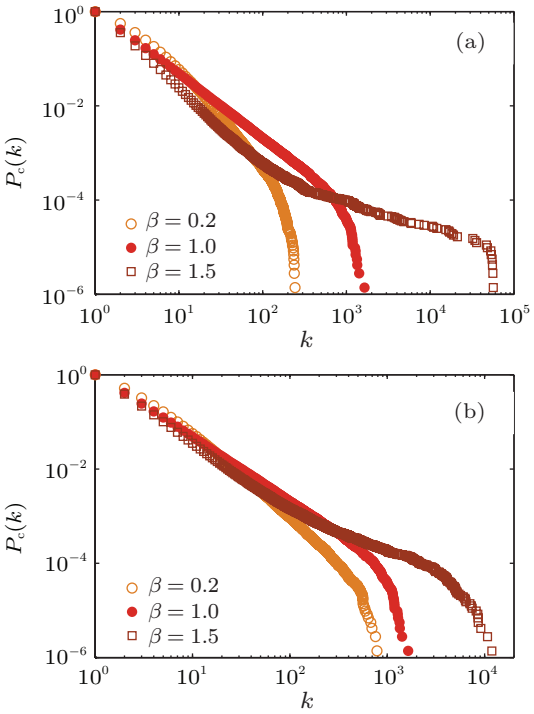


Fig. 2. (colour online) Complementary cumulative degree distributions of networks obtained by numerical simulations with $p = 0.6$, $N = 10^5$, and $m_1 = m_2 = 1$. (a) Rule (ii) is unchanged, and in rule (i), the probability with which the new vertex will be connected to the old vertex i with degree k_i is $\Pi(k_i) \sim k_i^{\beta}$. (b) Rule (i) is unchanged, and in rule (ii), an endpoint is chosen according to $\Pi(k_i) \sim k_i^{\beta}$.

According to the empirical data, we obtain $p = 0.7941$ and $q = 0.1939$. The links created between two new users are few and thus can be negligible. In addition, $m_1 = m_2 = 1$ for the real growth of the network. Based on the parameters and Eq. (14), we obtain $P(k) \sim k^{-2.67}$. Figure 3 shows the numerical result obtained by averaging over 10 independent realizations with $p = 0.7941$ and the same number of vertices as that in Wealink. Its degree exponent 2.62 agrees well with the predicted value of 2.67. Figure 3 also presents the complementary cumulative degree distribution of Wealink. We fit the network data with the power-law model utilizing the maximum likelihood estimate method and obtain $\gamma = 2.91$.[22] The predicted value of the degree exponent 2.67 achieves proper agreement with the real value 2.91. We also compute p for the estimated power-law fit to the network implementing the Kolmogorov–Smirnov test and obtain $p = 0.704$. We choose a threshold of 0.1, and thus the power-law fit is a good match to the degree distribution of Wealink.

In the real world, different from the ideal model, the probability p cannot be stationary during the evo-

lution of OSNs. In one stage, p can be very large, while in another stage, p can be very small, which can lead to the difference between the real exponent and the predicted one. Figure 4 shows the evolutions of p and q and demonstrates the above fact. As a guide, we also label the positions of $p = 0.7941$ and $q = 0.1939$.

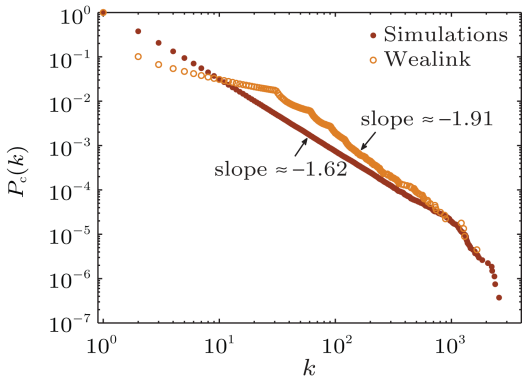


Fig. 3. (colour online) Complementary cumulative degree distributions of Wealink and the networks obtained by numerical simulations. The degree exponents obtained by the maximum likelihood estimate for Wealink and the numerical results are 2.91 and 2.62, respectively.

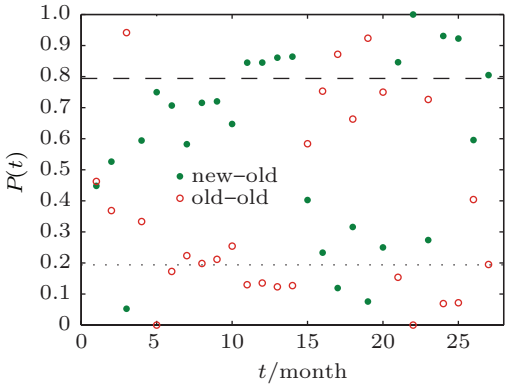


Fig. 4. (colour online) Evolutions of the fractions of two kinds of edges. Dashed and dotted lines indicate $p = 0.7941$ and $q = 0.1939$, respectively.

5. Discussion and conclusion

In summary, we empirically study the preferential linking in an evolving OSN, find and validate that the linear preference holds for preferential creation, preferential acceptance, and preferential attachment, i.e., the probability that an old user is chosen is directly proportional to his/her degree. We propose an analyzable model, which integrates the linear preference and the introduction of new vertices and edges. The model captures the real growth process of OSNs and reveals the underlying mechanism dominating their evolution. The simulation results, the analytical results, and the empirical observations agree very well with each other. This work not only provides a verifiable origin for the emergence of scaling law in OSNs, but also contributes to the understanding of the correlation between users' behavior and the organization of virtual communities.

The preferential linking indicates that a user's ability to acquire new ties depends on his/her current degree. Users looking for new connections use the other user's degree as a proxy for his/her fitness. The number of friends of a user usually is shown on his/her personal profile, which can be browsed by the other users. A user with many friends will be judged as more popular than another one with few friends.[23] However, it is noteworthy that from the perspective of sociology, there are diverse mechanisms which can lead to the formation of dyadic ties, such as assortative, relational, or proximity mechanisms.[24] Preferential linking is just one of the mechanisms. It can account for the emergence of power-law in OSNs; however it cannot explain the other structural or sociological characteristics of the networks. Different mechanisms usually are intimately interwoven in the evolution of OSNs. Thus, to understand the formation of dyadic ties, we need to incorporate sociological viewpoints and approaches into the study of OSNs.

References

[1] O'Reilly T 2007 *Comm. Strat.* **1** 17
[2] Lewis K, Kaufman J, Gonzalez M, Wimmer A and Christakis N 2008 *Social Networks* **30** 330
[3] Traud A L, Mucha P J and Porter M A 2012 *Physica* A 391 4165
[4] Jiang J, Wilson C, Wang X, Huang P, Sha W, Dai Y and Zhao B Y 2010 *Proceedings of the 10th Annual Conference on Internet Measurement*, November 1–3, 2010 Melbourne, Australia, p. 369
[5] Ahn Y Y, Han S, Kwak H, Moon S and Jeong H 2007 *Proceedings of the 16th International Conference on World Wide Web*, May 8–12, 2007 Banff, Canada, p. 835
[6] Bainbridge W S 2007 *Science* **317** 472
[7] Lazer D, Pentland A, Adamic L, Aral S, Barabási A L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D and Van Alstyne M 2009 *Science* **323** 721
[8] Bai M, Hu K and Tang Y 2011 *Chin. Phys.* B **20** 128902
[9] Holme P, Edling C R and Liljeros F 2004 *Social Networks* **26** 155
[10] Viswanath B, Mislove A, Cha M and Gummadi K P 2009 *Proceedings of the 2nd ACM Workshop on Online Social Networks*, August 16–21, 2009 Barcelona, Spain, p. 37
[11] Hu H and Wang X 2009 *Phys. Lett.* A **373** 1105
[12] Chun H, Kwak H, Eom Y H, Ahn Y Y, Moon S and Jeong H 2008 *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, October 20–22, 2008 Vouliagmeni, Greece, p. 57

[13] Szell M and Thurner S 2010 *Social Networks* **32** 313

[14] Robins G, Snijders T, Wang P, Handcock M and Pattison P 2007 *Social Networks* **29** 192

[15] Hu H B and Guo J L 2012 *Advances in Complex Systems* **15** 1250030

[16] Grindrod P and Parsons M 2011 *Physica* A **390** 3970

[17] Zou Z Y, Liu P, Lei L and Gao J Z 2012 *Chin. Phys.* B **21** 028904

[18] Guo J L, Guo Z H and Liu X J 2011 *Chin. Phys.* B **20** 118902

[19] Ren X Z, Yang Z M, Wang B H and Zhou T 2012 *Chin. Phys. Lett.* **29** 038904

[20] Mislove A, Koppula H S, Gummadi K P, Druschel P and Bhattacharjee B 2008 *Proceedings of the First Workshop on Online Social Networks*, August 17–22, 2008 Seattle, USA, p. 25

[21] Leskovec J, Backstrom L, Kumar R and Tomkins A 2008 *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 24–27, 2008 Las Vegas, USA, p. 462

[22] Clauset A, Shalizi C R and Newman M E J 2009 *SIAM Rev.* **51** 661

[23] Utz S 2010 *J. Comput. Mediat. Commun.* **15** 314

[24] Rivera M T, Soderstrom S B and Uzzi B 2010 *Annu. Rev. Sociol.* **36** 91