Optimizing Online Social Networks for Information Propagation

Duan-Bing Chen[1,2], Guan-Nan Wang[1], An Zeng[2]*, Yan Fu[1], Yi-Cheng Zhang[1,2]

1 Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, **2** Department of Physics, University of Fribourg, Fribourg, Switzerland

Abstract

Online users nowadays are facing serious information overload problem. In recent years, recommender systems have been widely studied to help people find relevant information. Adaptive social recommendation is one of these systems in which the connections in the online social networks are optimized for the information propagation so that users can receive interesting news or stories from their leaders. Validation of such adaptive social recommendation methods in the literature assumes uniform distribution of users' activity frequency. In this paper, our empirical analysis shows that the distribution of online users' activity is actually heterogenous. Accordingly, we propose a more realistic multi-agent model in which users' activity frequency are drawn from a power-law distribution. We find that previous social recommendation methods lead to serious delay of information propagation since many users are connected to inactive leaders. To solve this problem, we design a new similarity measure which takes into account users' activity frequencies. With this similarity measure, the average delay is significantly shortened and the recommendation accuracy is largely improved.

Citation: Chen D-B, Wang G-N, Zeng A, Fu Y, Zhang Y-C (2014) Optimizing Online Social Networks for Information Propagation. PLoS ONE 9(5): e96614. doi:10.1371/journal.pone.0096614

Editor: Rodrigo Huerta-Quintanilla, Cinvestav-Merida, Mexico

Received December 9, 2013; **Accepted** April 9, 2014; **Published** May 9, 2014

Funding: This work was partially supported by the National Natural Science Foundation of China with Grant Nos. 61103109, 61003231 and 60973120, and by the Swiss National Science Foundation (Grant No. 200020-143272). DBC acknowledges the Huawei University-Enterprise Cooperation Project YBCB2011057 and the Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: an.zeng@unifr.ch

Introduction

The information and communication technologies lead us to an information-rich era where recommender systems are widely used to filter out irrelevant information [1–3]. Recommendation algorithms include correlation-based collaborative filtering [4–6], Bayesian clustering [7], probabilistic latent semantic analysis [8], matrix decomposition [9,10]. Many issues related to recommender systems have been studied, such as the diversity of recommendations [11,12], the effect of network topology [13], and ground user [14]. Recent researches show that social influence [15] is more powerful than the purely mathematical analysis based recommendation, as people are more likely to accept the recommendations coming from their friends or peers [16]. Hence, a new technology named social recommendation has emerged [17–19] in which users (followers) can select some other users as information sources (leaders) and the information will automatically flow from leaders to followers. This framework has been successfully applied in many real online websites, such as *delicious.com*, *twitter.com* and *digg.com*. The information can refer to news, movies, books, bookmarks, and so on. Without losing any generality, news is used as an example in this paper. When a piece of news is submitted or approved by a user, it will be forwarded to her followers. The diffusion of news thus depends on the structure of leader-follower network, with higher transmission probability of news if users with higher similar tastes are linked [20].

Recently, an adaptive news recommendation model is proposed [21]. In this model, when a user reads news, she can either "approve" or "disapprove" it. If approved, the news will be forwarded to her followers. With the spreading of news, the leader-follower network will be updated, that is, the least suitable leader of a user will be replaced by a better one according to the quality of the leader. The quality of her leader is measured by the similarity based on their past assessments on news. The model has been extensively tested by additional aspects like users' reputation [22], implicit ratings [23], local topology optimization [24], leadership structure [25] and link reciprocity [26]. More recently, Cimini et al. [27] introduced two settings for modeling users' tastes and showed that the heterogeneous setting of users' tastes was closer to the real case than homogeneous setting.

Confirmed by many empirical analysis, it is now well-known that the activity frequency of online users are heterogenous [28–30]. However, in the original adaptive news recommendation model and the following studies, users are randomly selected to be active (i.e. submitting or reading news), which indicates that the activity frequency of users are set to be homogenous. In this paper, we find that the propagation of news is seriously delayed when some classic similarity metrics are applied in the heterogenous users activity setting. Moreover, the recommendation accuracy (i.e. approval fraction of the news) is lowered as well. To solve this problem, we propose a new similarity measure which takes user activity into account. The simulation shows that the propagation delay is considerably shortened and recommendation accuracy is largely improved. Finally, we introduce a more general similarity definition in which the weight on users' news assessments and users' activity is tunable. With this, the effectiveness (accuracy) and efficiency (time delay) of the information propagation is further improved.

Materials and Models

Empirical analysis

To begin our analysis, we study the distribution of user activity frequency in real systems. Here, we consider the dataset of digg.com [31], which contains 3,018,197 votes on 3,553 popular stories made by 139,409 distinct users over a period of a month in 2009. In this dataset, users and stories form a bipartite network in which a link between a user i and a story α exists if user i reads story α. The degree of a user represents the number of stories read by her (i.e. the activity of this user). Figure 1 shows the degree distribution in three observation time windows. Clearly, the distribution follows a power-law with exponent around -2. The results confirm that users are very heterogenous in the frequency of their online activity.

Furthermore, we employ the Kendall's tau coefficient (τ) to calculate users' correlation of activity frequency in two adjacent periods. The length of the period is set as one day, so we are actually calculating the Kendall's tau coefficient of users' activity frequency in each day and the previous day. As shown in Fig. 2, τ is always larger than zero, which means that users' activity frequency is positively correlated in time. Moreover, we observe that there is some periodic fluctuation in Fig. 2. With the actual dates, we check carefully the reason for this periodic fluctuation. We find that each period is one week, and the correlation is higher in weekdays than in weekends. We conjecture it is because people's live is regular in weekdays but diverse in weekends.

Model description

In the original adaptive news recommendation model [21] and the following studies [22–27], users' activity frequencies are assumed to be homogenous, which is inconsistent with the results of above empirical study. To make the information propagation model closer to the real system, we introduce the heterogeneity of users' activity frequency to it. Our model will be directly built on the original news-sharing model in ref. [21]. The system consists of U users. Each of them is connected by directed links to L other users, who represent her news sources and to whom we refer as her *leaders*. The value of L is fixed as users can follow a limited number of sources. Users receive pieces of news from their leaders, and eventually assess them. In addition, they can introduce new content to the system.

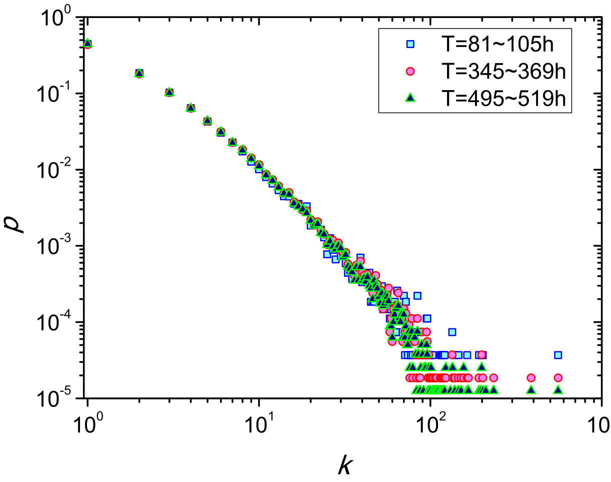

Figure 1. The distribution of activity frequency of users on Digg.com over a period of a month in 2009.
doi:10.1371/journal.pone.0096614.g001

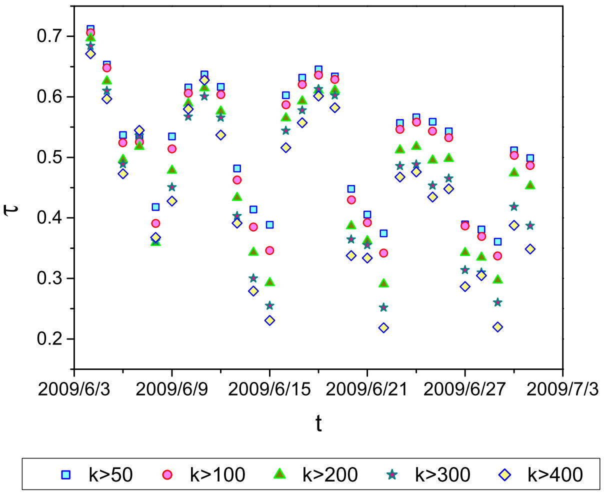

Figure 2. The correlation of the activity frequency of users in two adjacent days. The actually dates are marked in the horizontal axis. $k>x$ in the legend means that we only take into account the users with total activity being larger than x.
doi:10.1371/journal.pone.0096614.g002

Evaluation of news α by user i ($e_{i\alpha}$) is either $+1$ (liked), -1 (disliked) or 0 (not read yet). The set of evaluations from any pair of users i and j is the basis to compute their similarity of their interests (or reading tastes), which is denoted as s_{ij}. The explicit recipes to compute users' similarity are presented in the next section. Note that, apart from their evaluations, no other information about users is assumed by the model.

Users' activity. In each time step of the simulation, a given user is active with probability p_A. When active, a user reads the top R news from her recommendation list, immediately forwarding the ones she likes to her followers. In addition, with probability p_S she submits a new piece of news. Different from the original model in [21], the users' activity frequency is drawn from a power-law distribution as $P(p_A)\sim p_A^{-\gamma}$ where $\gamma=2$.

Propagation of news. When news α is introduced to the system by user i at time t_α, it is forwarded from i to the users j who have selected her as a leader, with a *recommendation score* proportional to their similarity s_{ij}. If this news is later liked by one of her followers j, it is similarly passed further to this user's followers q, with recommendation score proportional to s_{jq}, and so on. For a generic user q at time t, a news α is recommended to her according to its current score:

$$R_{q\alpha}(t)=\delta_{e_{q\alpha},0}\,\lambda^{t-t_\alpha}\sum_{l\in L_q} s_{ql}\,\delta_{e_{l\alpha},1} \qquad (1)$$

where L_q is the set of leaders of user q. δ is a Dirac delta function with only two possible values: 0 and 1. If user q has not read news α, $\delta_{e_{q\alpha},0}=1$ since $e_{q\alpha}=0$ and if q has read news α, $\delta_{e_{q\alpha},0}=0$ since $e_{q\alpha}\neq 0$. Similarly, $\delta_{e_{l\alpha},1}=1$ if user l likes news α, otherwise $\delta_{e_{l\alpha},1}=0$. To make the fresh news fast accessed, recommendation scores are damped with time ($\lambda\in(0,1]$ is the damping factor).

Leader selection. The model is adaptive. Initially, each user randomly select L other users as her leaders. Leader-follower connections are periodically rewired to make the social network approach an optimal state where only highly similar users are connected [32]. In each rewiring, for user i, her current leader j with the lowest similarity value is replaced with a new user (q) if $s_{iq}>s_{ij}$. There are different selection strategies for picking new candidate leaders, which are discussed in detail in [22,24,26]. In this paper we employ a hybrid strategy in which the user q is picked at random in the network with probability 0.1, otherwise she is selected among the leaders' leaders and followers of user i to maximize s_{iq}. This mechanism well mimics users establishing mutual friendship relations, searching for friends among friends of friends, and having casual encounters which may lead to long-term relationships. In addition, it is an excellent compromise between computational cost and system's performance [24].

Measure of users' similarity

An essential ingredient of the social recommendation algorithms is the estimated similarity of users' reading tastes, which regulates the news' flow over the system by determining the leaders' selection from users (i.e., the link structure of the network) and recommendation scores of news. Since only users' ratings and records of activities are known, the similarity of a pair of users has to be estimated from their assessments on news, which in our case can be either approved, disapproved, or not rated.

The first similarity measure considered is introduced in [21] as

$$s_{ij}^{(0)}=\frac{|A_i\bigcap A_j|+|D_i\bigcap D_j|}{|N_i\bigcap N_j|}\left(1-\frac{1}{\sqrt{|N_i\bigcap N_j|}}\right), \qquad (2)$$

where $A_i(A_j)$ is the set of news approved by user i (user j), $D_i(D_j)$ is the set of news disapproved by user i (user j), and $N_i(N_j)$ is the set of news read by user i (user j). The term $1-1/\sqrt{N_i\bigcap N_j}$ is used to remove the effect of statistical fluctuation. If a user i and a user j share a small number of commonly read news, they are more likely to achieve "perfect" similarity 1. After multiplying this term, the similarity measure will give this user pair a very low similarity value. In sampling of n trials, the typical relative fluctuation is of the order of $1/\sqrt{n}$. Therefore, we select the above form.

In [33], it is shown that Eq. (2) works well only in the system where tastes of users are homogeneously distributed, i.e., each user has the same number of interested fields. To achieve a more accurate leader assignment in the system where users can have different number of tastes, an asymmetric similarity measure is defined as

$$s_{ij}^{(1)}=\frac{|A_i\bigcap A_j|}{|A_j|}\left(1-\frac{1}{\sqrt{|A_j|}}\right). \qquad (3)$$

$1-1/\sqrt{|A_j|}$ here is also used to remove the effect of statistical fluctuation.

In this paper, we consider the systems where users' activity distribution is uneven. Some users can be extremely active and read many news, so that their followers can constantly receive fresh news. On the other hand, if one user is connected to many inactive leaders, the news received by her will be very limited. Therefore, recommending highly inactive users is meaningless. Considering this, we modified the similarity in Eq. (3) as

$$s_{ij}^{(2)}=s_{ij}^{(1)}H_j(t), \qquad (4)$$

where $H_j(t)$ is a measurement of user j's activity. Actually, there are many other previous works showing that online users' activity frequency is unevenly distributed[34,35].

Users' active frequencies p_A are users' inherent feature and unknown by the recommender system. We design the following way to estimate p_A of the users. Instead of taking the whole history into account, we only use the recent record of activities within a time window $[t',t-1]$ with length T (In our simulation, $T=250$ generally works best, see Supporting Information S1). The estimated probability for user j to get online is

$$P_j(t)=\frac{1}{T}\sum_{w=t-T}^{t-1}\frac{f_j(w)}{t-w}, \qquad (5)$$

where $f_j(w)$ is user j's online times from time w to time $t-1$. If $f_j(w)$ equals 0 for each w, we set $P_j(t)$ as $\frac{1}{T(T+1)}$. In Eq. (5), users' recent record plays a more important role. This is very useful in real systems, since the correlation of real users' activity frequency is generally high in short term (See Fig. 2).

However, $P_j(t)$ cannot be directly used as $H_j(t)$ in Eq. (4). Since $P_j(t)\sim P(p_A)$, $P_j(t)$ follows a power-law distribution. If it is used as $H_j(t)$, some users with high activity will dominate the similarity matrix and be always selected as the leaders of others. In order to solve this problem, we proposed a logarithmic way to embed $P_j(t)$ in $H_j(t)$. After normalization, it reads

$$H_j(t)=\frac{-(\ln P_j(t)-\ln x_0)}{\ln x_0}, \qquad (6)$$

where x_0 is the possible lowest value of $P_j(t)$, set as $\frac{1}{T(T+1)}$. After simplification, $H_j(t)=1+\log_{T(T+1)}P_j(t)$. In this definition, $H_j(t)$ can distinguish different users and the most inactive users are punished severely. However, the majority gets $H_j(t)$ over 0.5, as shown in Fig. 3. We rewrite $s_{ij}^{(2)}$ as

$$s_{ij}^{(2)}=s_{ij}^{(1)}\left(1+\log_{T(T+1)}P_j(t)\right). \qquad (7)$$

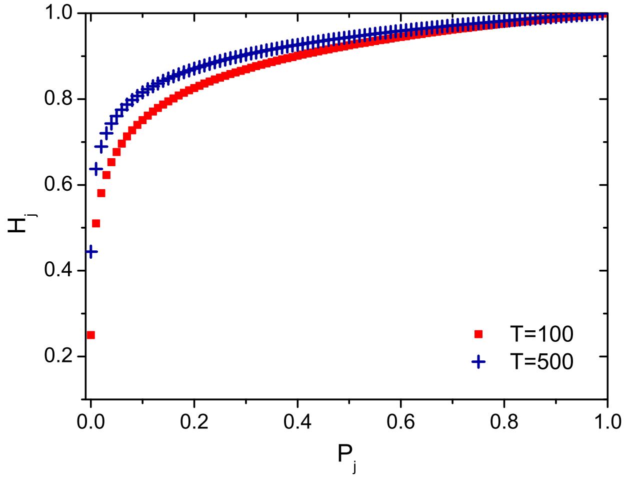

Figure 3. The relation between P_j and H_j.
doi:10.1371/journal.pone.0096614.g003

Table 1. List of parameters used in simulations.

parameter	symbol	value
Number of users	U	3498
Number of leaders per user	L	10
Dimension of taste vectors	D	20
Minimum active elements per vector	D_{min}	4
Maximum active elements per vector	D_{max}	8
Users' approval threshold	Δ	3
Index of distribution	γ	-2
Probability of submitting a news	p_S	$p_A/40$
Number of news read when active	R	3
Damping of recommendation score	λ	0.9
Base similarity for users	s_0	10^{-7}
Period of the rewiring	u	10

doi:10.1371/journal.pone.0096614.t001

In our simulation, we actually compare our method $s^{(2)}$ to with some start-of-the-art similarity methods based on both news-reading and topology. The results show that $s^{(2)}$ can outperform others, see Supporting Information S1. In the following, we will study the behavior of the system under these similarity metrics. For numerical tests of the model, we use an agent-based framework.

Agent-based simulations

To model users' judgments of read news we use a vector model where tastes of user i are represented by a D-dimensional taste vector $\vec{a}_i=(a_i^1,\ldots,a_i^D)$ and attributes of news α are represented by a D-dimensional attribute vector $\vec{b}_\alpha=(b_\alpha^1,\ldots,b_\alpha^D)$. Similar vector models are often used in semantic approaches to recommendation [36]. Opinion of user i about news α is based on the overlap of the user's tastes and the news's attributes, which can be expressed by the scalar product

$$\Omega_{i\alpha}=\vec{a}_i\cdot\vec{b}_\alpha. \tag{8}$$

We assume that user i approves news α only when $\Omega_{i\alpha}\geq\Delta$, disapproves otherwise, where Δ is the users' approval threshold: the higher it is, the more demanding the users are. Here, we adopt the heterogeneous setting of the taste/attribute vectors.

Each user has preference for a variable number of D available tastes. Each taste vector has a different number of elements equal to one (active tastes, denoted as d) and the remaining elements are zero. In this paper, we assume $D_{min}\leq d\leq D_{max}$ where D_{min} and D_{max} are the minimum and maximum number of active tastes that users can have, respectively. Moreover, we assume that each news' attribute vector has a fixed number D_{min} of active attributes (number of ones), which are randomly chosen among the active tastes of the user who submits it.

Simulation runs in discrete time steps. Assuming no a priori information, the starting network configuration is given by randomly assigning L leaders to each user. Then in each simulation step, an individual user is active with probability p_A. When active, the user reads and evaluates the R top-recommended news she has received and with probability p_S submits a new news. Connections are rewired every u simulation steps. Parameter values used in all following simulations are given in Table 1. A detailed study of the effect of the parameters on the model can be seen in [24].

Metrics

Three metrics are used to measure the performance of the recommendation models: *approval fraction* ψ, *average differences* δ, and *average delay* ϕ.

Approval fraction ψ. The ratio of news' approvals to all assessments is an obviously important measure of the method's performance. This number, referred to as approval fraction, tells us how often users are satisfied with the news they get recommended. The higher the ψ is, the more accurate the recommendation is, and users are more satisfied correspondingly. It can be defined as

$$\psi=\frac{\sum_{i\alpha}\delta_{e_{i\alpha},1}}{\sum_{i\alpha}\delta_{|e_{i\alpha}|,1}}, \tag{9}$$

where term $\delta_{e_{i\alpha},1}$ equals one if user i approved news α and zero otherwise, and term $\delta_{|e_{i\alpha}|,1}$ equals one if user i has rated news α and zero otherwise.

Average differences δ. In the computer simulation, we have the luxury of knowing users' taste vectors and hence we can compute the number of differences between the taste vector of a user and the taste vectors of the user's authorities. By averaging over all users, we obtain the average number of differences. Obviously, the less are the differences, the better is the assignment of authorities. The average differences are defined as

$$\delta=\frac{1}{U}\sum_i\sum_{l\in L_i}\frac{\|\vec{a}_i-\vec{a}_l\|}{|L_i|}, \tag{10}$$

where L_i is the set of leaders of user i, and $\vec{a}_i$ ($\vec{a}_l$) is the taste vector of user i (user l).

Average delay ϕ. The freshness of the news is very important. Once the news becomes old, it is of no interest to users. The average delay measures the novelty of the news read by users. A small average delay indicates that users are always reading fresh news. The average delay is defined as

Table 2. Three evaluation metrics on different similarity measures.

	$s^{(0)}$			$s^{(1)}$			$s^{(2)}$		
	Pos.	**Neg.**	**Uncorr.**	**Pos.**	**Neg.**	**Uncorr.**	**Pos.**	**Neg.**	**Uncorr.**
Average delay ϕ	247.08	426.52	1058.10	377.23	881.12	1383.01	103.93	109.30	123.53
Average differences δ	6.1838	7.5234	8.3693	5.7045	6.7641	6.9015	5.5779	5.5033	5.3161
Approval fraction ψ	0.4561	0.3034	0.2280	0.3873	0.3324	0.3198	0.4274	0.4187	0.3502

doi:10.1371/journal.pone.0096614.t002

$$\phi = \frac{\sum_{i\in U_s}\sum_{\alpha\in N_i}(t_{i\alpha}-t_\alpha)}{\sum_{i\in U_s}|N_i|}, \tag{11}$$

where N_i is the set of news read by user i, $t_{i\alpha}$ is the time when i reads news α, t_α is the submitted time of news α, and U_s is set of users. The smaller the ϕ is, the news read by users will be fresher.

Results and Discussion

We now study the described adaptive social recommender system under different definitions of the similarity measure employed. For comparison, initial conditions and parameters for all simulations are identical, as listed in Table 1. We obtained the average differences, approval fraction and average delay resulting from each similarity definition. We first consider the case where the user activity and the number of user interest are uncorrelated and the results are shown in the uncorrelated case in Table 2. As expected, $s^{(1)}$ enjoys a higher approval fraction and a smaller average difference than $s^{(0)}$. The results are consistent with ref. [27,33]. However, $s^{(0)}$ results in a smaller average delay than $s^{(1)}$. Among these methods, $s^{(2)}$ performs the best in all three metrics. These results suggest that introducing the users' activity to the similarity measure can significantly speed up the propagation of news in online systems so that users mostly receive fresh news. Moreover, it improves the leader assignment, resulting in a more accurate recommendation of news for users.

One concern of the new similarity measure is that the network updating algorithm might focus too strongly on the activity frequency of leaders rather than the taste overlap of the leader and follower, putting the information recommendation accuracy at risk. Accordingly, we further introduce a parameter η to adjust the effect of users' activity in the $s^{(2)}$ similarity calculation:

$$s^{(3)}_{ij} = s^{(1)}_{ij}\left(1+\log_{T(T+1)}P_j\right)^{\eta}. \tag{12}$$

Parameter η controls the weight assigned to $H_j(t)$. When $\eta = 0$, Eq. (12) reduces to $s^{(1)}$, and when $\eta = 1$, Eq. (12) reduces to $s^{(2)}$.

The stationary values of average delay ϕ and approval fraction ψ obtained by using $s^{(3)}$ under different η are reported in Fig. 4. One can immediately see that there is a maximum approval fraction when adjusting η. On the other hand, the average delay drop dramatically once $\eta > 0$ and then decreases monotonously with η. Compared to the case where the users' activity is not considered in the similarity calculation ($\eta = 0$), the approve fraction can be improved and average delay can be considerably shortened. Interestingly, the optimal η is around 1, corresponding to the case of $s^{(2)}$.

We further consider the situation where the user activity and the number of user interest are correlated. Three cases are considered in this paper: positive correlation, negative correlation, and no correlation. We first compare three evaluation metrics under different correlation settings (see Table 2). One can immediately see that both positive and negative correlation between the user activity and the number of user interests can significantly shorten the average delay ϕ of the news. However, the delay from $s^{(0)}$ and $s^{(1)}$ are still longer than $s^{(2)}$. The advantage of $s^{(2)}$ can also be found in the average difference δ. A lower average difference indicates that the network is adapted to a better state for news propagation. We can also see that $s^{(2)}$ enjoys the highest approval fraction ψ in almost all cases. When the correlation between the

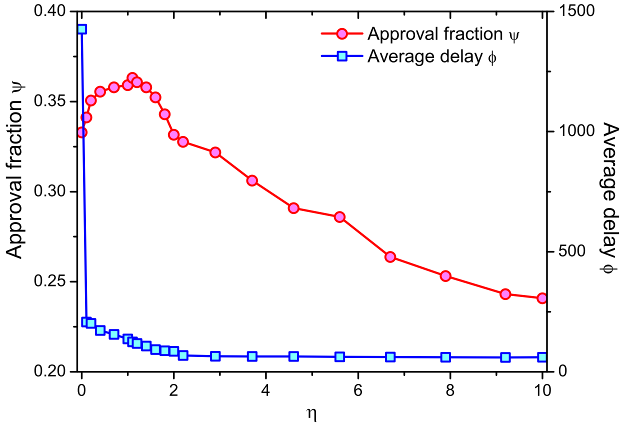

Figure 4. Stationary values (simulation step #15000) of *average delay* ϕ and *approval fraction* ψ in the adaptive system ruled by $s_{ij}^{(3)}$, and for different values of the parameter η.
doi:10.1371/journal.pone.0096614.g004

user activity and the number of user interests is positive, the approval fraction of $s^{(0)}$ is a bit higher than that of $s^{(2)}$. However, the delay in $s^{(0)}$ in this case is more than twice longer than that of $s^{(2)}$. Taken together, $s^{(2)}$ is a very effective and robust similarity measure for recommending leaders in online social systems.

The leader-follower networks after the systems reach stable state is studied. We first investigate the in-degree distribution of nodes (i.e. the distribution of number of followers). The results show that the largest in-degree in $s^{(2)}$ is smaller than that in $s^{(0)}$ and $s^{(1)}$. This is because users in $s^{(2)}$ select leaders according to not only the similarity but also the activity frequency. It is more difficult for the largest in-degree nodes in $s^{(2)}$ to attract as many followers as in $s^{(0)}$ since these users have both high similarity to others and high activity frequency. We then study some properties of the users of different activities in Fig. 5. Fig. 5(a) shows the relation between user activity and the number of her followers. As discussed above, if the leaders of a user are with low activity, the user may have no news to read and the propagation of news will be largely delayed. This happens a lot in the original similarity measure $s^{(0)}$ and $s^{(1)}$ (See the flat curves of them in Fig. 5(a)). We didn't plot the curves of negative and uncorrelated cases in $s^{(0)}$ and $s^{(1)}$ because they are as flat as in the positive correlation case. In $s^{(2)}$, the users with higher activity frequency have more followers, which makes users with rich number of news to read. Moreover, we observe that the users with higher activity and fewer interests (see the negative correlation case when p_A is large) have more followers. In ref. [33], it is already pointed out that the users with few interests are good information resource and should be selected as leaders (since they are specialized in their fields). As shown in Fig. 5(a), $s^{(2)}$ recommends the users with high activity and few interests as leaders to others. This again supports that $s^{(2)}$ is a good similarity measure.

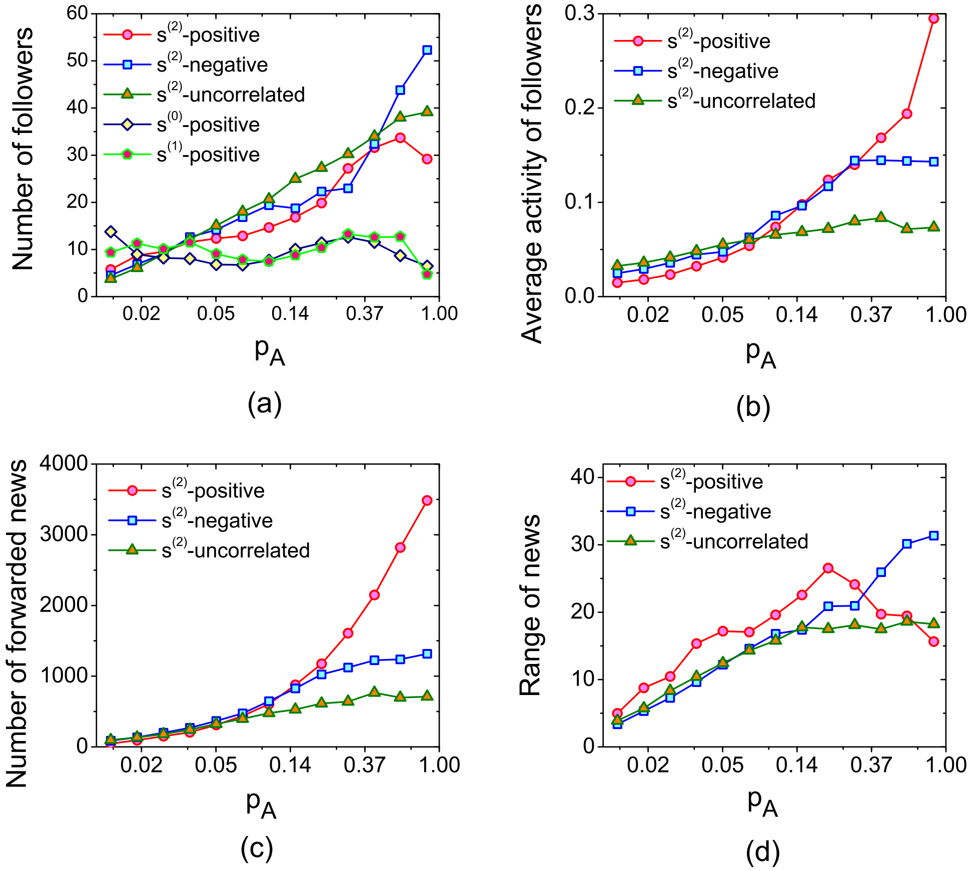

Figure 5. The (a) number of followers, (b) average activity of followers, (c)number of forwarded news and (d) spreading range of users with different activities (p_A).
doi:10.1371/journal.pone.0096614.g005

In Fig. 5(b), we present the relation between users' activity and the average activity of their followers. We can see that the users with higher activity and wider interests (i.e. large p_A in positive correlation) have more active followers. Generally, the interests of the followers are wider than that of the leaders [33]. The followers of large p_A users will have wide interests and thus high activity.

Moreover, it is interesting to identify which kind of users can be the information hubs in online social networks. We first investigate the number of forwarded news of different users. As shown in Fig. 5(c), the users with higher activity and wider interests forward more news. With wider interests, these users are more likely to approve the news from their leaders, which results in a large number of forwarded news from them.

However, the results in Fig. 5(d) indicate that the users with high activity and wide interests are actually not information hubs. We report the spreading range when the news is originated from different users in Fig. 5(d). The spreading range here is defined as the number of users who finally read the news. As discussed above, the users with fewer interests are more specialized in their fields and their followers are more likely to approve the news from them. Therefore, the news originated from users with higher activity and fewer interests will spread wider. The results imply that the active and specialized users are the information hubs in online social networks.

Conclusion

In this paper, we study a new multi-agent based model for information propagation and recommendation on online social network. The original online information propagation model was proposed in ref. [21] where users' activity frequency is assumed to be homogeneously distributed. Since the empirical study of the online news-sharing systems suggests that users' activity frequency distribution actually follows a power-law distribution, we introduce the heterogeneity to users' activity frequency distribution to the model in ref. [21]. We find that previous similarity methods for leader recommendation connects many users to inactive leaders, resulting serious delay of information propagation and low approval fraction of news.

To solve this problem, we propose a new similarity measure which takes users' activity frequency into account. With the new similarity measure, the suitability of a leader is evaluated according to not only the similarity but also the activity frequency. The numerical simulation shows that our method can outperform the existing ones in network optimization for information recommendation, in both approval fraction and information delay. Finally, we introduce a parameter to adjust the effect of users' activity in the similarity calculation. We find that the leader recommendation can be further improved by this parameter.

Since real online users have heterogenous activity frequency, we believe that our method will be very useful from practical point of view. Since real online news-sharing systems can be different from current model in parameter settings or even news propagation mechanism, the optimal weight of users' activity frequency in the similarity calculation should be determined by some preliminary testings. One possible way is to implement the method first on a small subset of users. After obtaining the optimal balance between users' activity frequency and similarity from the learning procedure, the method can be applied to the whole systems.

Supporting Information

Supporting Information S1 Supporting text and figures. (RAR)

Author Contributions

Conceived and designed the experiments: DBC AZ YCZ. Performed the experiments: DBC GNW. Analyzed the data: DBC GNW AZ YF. Wrote the paper: DBC AZ YCZ.

References

1. Lü L, Medo M, Yueng CH, Zhang YC, Zhang ZK, et al. (2012) Recommender systems. Phys Rep 519: 1–49.
2. Biancalana C, Gasparetti F, Micarelli A, Sansonetti G (2013) An approach to social recommendation for context-aware mobile services. ACM Trans Intell Syst Technol 4: 10.
3. De A, Ganguly N, Chakrabarti S (2013) Discriminative link prediction using local links, node features and community structure. arXiv: 1310.4579.
4. Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 22: 5–53.
5. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17: 734–749.
6. Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7: 76–80.
7. Breese J, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. of the 14^{th} Conf. on Uncertainty in Artificial Intelligence. ACM Press, pp. 45–53.
8. Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Syst 22: 89–115.
9. Maslov S, Zhang YC (2001) Extracting hidden information from knowledge networks. Phys Rev Lett 87: 248701.
10. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
11. Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, et al. (2010) Solving the apparent diversity accuracy dilemma of recommender systems. Proc Natl Acad Sci USA 107: 4511–4515.
12. Niemann K, Wolpers M (2013) A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. In: Proc. of 19^{th} ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, pp. 955–963.
13. Zhang CJ, Zeng A (2012) Behavior patterns of online users and the effect on information filtering. Physica A 391: 1822–1830.
14. Zhou Y, Lü L, Liu W, Zhang J (2013) The power of ground user in recommender systems. PLoS ONE 8: e70094.
15. Quijano-Sanchez L, Recio-Garcia JA, Diaz-Agudo B, Jimenez-Diaz G (2013) Social factors in group recommender systems. ACM Trans Intell Syst Technol 4: 8.
16. Sinha R, Swearingen K (2001) Comparing recommendations made by online systems and friends. In: Proc. of the DELO-SNSF Workshop on Personalisation and Recommender Systems in Digital Libraries. Dublin City University, pp. 1–6.
17. Golbeck J (2008) Weaving a web of trust. Science 321: 1640–1641.
18. Kazienko P, Musiał K, Kajdanowicz T (2011) Multidimensional social network in the social recommender system. IEEE T Syst Man Cy A 41: 746–759.
19. Chaoji V, Ranu S, Rastogi R, Bhatt R (2012) Recommendations to boost content spread in social networks. In: Proc. of the 21st international conference on World Wide Web. ACM Press, pp. 529–538.
20. Macskassy SA, Michelson M (2011) Why do people retweet? anti-homophily wins the day! In: Proc. of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM). AAAI Press, pp. 209–216.
21. Medo M, Zhang YC, Zhou T (2009) Adaptive model for recommendation of news. EPL 88: 38005.
22. Cimini G, Medo M, Zhou T, Wei D, Zhang YC (2011) Heterogeneity, quality, and reputation in an adaptive recommendation model. Eur Phys J B 80: 201–208.
23. Wei D, Zhou T, Cimini G, Wu P, Liu W, et al. (2011) Effective mechanism for social recommendation of news. Physica A 390: 2117–2126.
24. Cimini G, Chen DB, Medo M, Lü L, Zhang YC, et al. (2012) Enhancing topology adaptation in information-sharing social networks. Phys Rev E 85: 046108.
25. Zhou T, Medo M, Cimini G, Zhang ZK, Zhang YC (2011) Emergence of scale-free leadership structure in social recommender systems. PLoS ONE 6: e20648.
26. Chen DB, Gao H (2012) An improved adaptive model for information recommending and spreading. Chin Phys Lett 29: 048901.
27. Cimini G, Zeng A, Medo M, Chen DB (2013) The role of taste affinity in agent-based models for social recommendation. Adv Complex Syst 16: 1350009.
28. Zhou T, Kiet HAT, Kim BJ, Wang BH, Holme P (2008) Role of activity in human dynamics. EPL 82: 28002.

29. Yu J, Hu Y, Yu M, Di Z (2010) Analyzing netizens view and reply behaviors on the forum. Physica A 389: 3267–3273.
30. Goncalves B, Perra N, Vespignani A (2011) Modeling users activity on twitter networks: validation of dunbars number. PLoS ONE 6: e22656.
31. Lerman K, Ghosh R (2010) Information contagion: an empirical study of the spread of news on digg and twitter social networks. In: Proc. of the Fourth International AAAI Conference on Weblogs and Social Media. AAAI Press, pp. 90–97.
32. Weng L, Ratkiewicz J, Perra N, Goncolves B, Castillo C, et al. (2013) The role of information diffusion in the evolution of social network. In: KDD'13. ACM Press, pp. 356–364.
33. Chen DB, Zeng A, Cimini G, Zhang YC (2013) Adaptive social recommendation in a multiple category landscape. Eur Phys J B 86: 61.
34. Zhao ZD, Cai SM, Huang J, Fu Y, Zhou T (2012) Scaling behavior of online human activity. EPL 100: 48004.
35. Lev M, Sen P, Lucas CP, Saulo DSR, José AJ, et al. (2013) Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Sci Rep 3: 1783.
36. Zorrilla M, Mazón J, Ferrández O, Garrigós I, Daniel F, et al, editors (2012) Business Intelligence Applications and the Web: Models, Systems and Technologies. Hershey, PA: IGI Global Press.