

RESEARCH ARTICLE

Identifying the perceptive users for online social systems

Jian-Guo Liu^{1,3*}, Xiao-Lu Liu², Qiang Guo², Jing-Ti Han¹

1 Data Science and Cloud Service Research Centre, Shanghai University of Finance and Economics, Shanghai 200433, PR China, **2** Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China, **3** Department of Physics, Fribourg University, CH-1700 Fribourg, Switzerland

* liujg004@ustc.edu.cn



Abstract

In this paper, the perceptive user, who could identify the high-quality objects in their initial lifespan, is presented. By tracking the ratings given to the rewarded objects, we present a method to identify the user perceptibility, which is defined as the capability that a user can identify these objects at their early lifespan. Moreover, we investigate the behavior patterns of the perceptive users from three dimensions: User activity, correlation characteristics of user rating series and user reputation. The experimental results for the empirical networks indicate that high perceptibility users show significantly different behavior patterns with the others: Having larger degree, stronger correlation of rating series and higher reputation. Furthermore, in view of the hysteresis in finding the rewarded objects, we present a general framework for identifying the high perceptibility users based on user behavior patterns. The experimental results show that this work is helpful for deeply understanding the collective behavior patterns for online users.

OPEN ACCESS

Citation: Liu J-G, Liu X-L, Guo Q, Han J-T (2017) Identifying the perceptive users for online social systems. PLoS ONE 12(7): e0178118. <https://doi.org/10.1371/journal.pone.0178118>

Editor: Wen-Bo Du, Beihang University, CHINA

Received: January 16, 2017

Accepted: May 7, 2017

Published: July 13, 2017

Copyright: © 2017 Liu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work is supported by the National Natural Science Foundation of China (Grant Nos. 71271126, 61374177, 71371125), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shuguang Program Project of Shanghai Educational Committee (Grant No. 14SG42).

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

Introduction

Collective behaviors have been extensively investigated to quantitatively explore the behavior patterns of online social systems [1–3], such as the bursty nature, heavy-tailed distribution of inter-event time [4, 5], and so on. Many remarkable patterns and mechanisms of collective behaviors have been proposed, such as the task-based queuing model [6, 7], interest-driven model [8, 9] and radiation model [10], which are gradually applied in the rumor spreading [11, 12], disease spreading [13–15] and recommendation systems [16–18], etc.

Recently, the online user behavior patterns have attracted more and more attention [19–21]. The abundance of available information increases the difficulty in making choices for users: Buy objects, borrow DVDs, or watch movies. Nowadays, online rating systems provide channels for users to show their preferences in the form of ratings [22–24], which can be represented as growing weighted bipartite networks where users are linked with the rated objects over time and the weights are the ratings. Preferential attachment [25, 26], the users connect objects in terms of the object degree preferentially, has been widely used to interpret user rating or selecting behaviors, presenting a homogeneous population composed of users driven by

object popularity. Meanwhile, Liu *et al* [17] found that users are heterogeneous in selecting the rated objects: Some objects are collected by almost all users, while some small-degree objects are only collected by large-degree users, indicating that the users' tastes could be expressed by two categories: Popular one and special one. The work of Ni *et al* [27] also described this idea. Inspired by these work, we carry on investigating the heterogeneity [28–31] of users in their rating patterns. An interesting phenomenon is found: While the majority of users usually collect the popular objects, some users frequently attach to the high-quality objects (which is rewarded finally) when they are rarely concerned, in which the latter group of users are our focus in this paper.

We divide objects into two sets: High-quality objects and the others, in which the high-quality objects are defined as rewarded objects here, e.g. Oscars Award for film, Grammy Award for music, Emmy Award for television, Tony Award for theater, etc. There will be many users focusing on the rewarded objects when they become widely accepted, while an interesting phenomenon is found: There exist some users paying attention to the rewarded objects long before they actually be widely approved (finally rewarded), i.e. at their early lifespan. Here we present two definitions: Perceptive user and user perceptibility. Perceptive user is defined as the user who can make high appraisals of the rewarded objects long before they actually be rewarded. Meanwhile, the degree to which the user can identify the rewarded objects in their initial lifespan is defined as the user perceptibility.

Meanwhile, we present a method to identify the user perceptibility based on online user rating behaviors. Then we investigate the behavior patterns of the perceptive users from three aspects: User activity, correlation characteristics of user rating series and user reputation. Experimental results indicate that high perceptibility users show different behavior patterns than others. Finally, considering the hysteresis in finding the rewarded objects, we present a framework for identifying high perceptibility users based on users' behavior patterns.

Material and methods

Data sets

In this paper, two empirical data sets containing timestamps and ratings for movies: MovieLens and Netflix are investigated. The MovieLens data set is downloaded from the GroupLens (<http://www.grouplens.org>), consists 943,355 ratings given by 4,295 users to 3,706 movies during 1,039 days. The Netflix data set is provided by the Netflix Prize (<http://www.netflixprize.com>), consists 37,755,925 ratings delivered by 218,319 users on 7,803 movies during 2,241 days. The MovieLens and Netflix ratings are both given by the integer ratings scaling from 1 to 5. Meanwhile, each user has at least 50 ratings for two data sets. Here, two object sets mentioned above, high-quality objects and the others, are divided based on the Oscars awards. We select movies nominated for the best picture category at the Annual Academy Awards, popularly known as Oscars (<http://www.filmsite.org>), as the high-quality objects. There are 162 and 150 rewarded movies in the MovieLens and Netflix data sets, respectively.

Method description

The rating system can be modeled by a weighted bipartite network, where the users and objects are denoted by $U = \{u_1, u_2, \dots, u_{|U|}\}$, $O = \{o_1, o_2, \dots, o_{|O|}\}$. We use the Latin and Greek letters to represent the users and objects, respectively. The rating $r_{i\alpha}$ given by user u_i to object o_α is the weight of the link connecting nodes u_i and o_α in the bipartite network. The timestamp of rating $r_{i\alpha}$ is denoted by $t_{i\alpha}$ and the highest rating is recorded as r_h . The user set U_α is defined as the users who rate to object o_α , and the object set O_i is recorded as the objects rated by user u_i . In addition, the degrees of user u_i and object o_α are denoted as k_i and ρ_α , respectively. Two

object sets, rewarded and non-rewarded ones, are denoted by object set O_1 and O_2 , respectively, satisfying $O_1 \cup O_2 = O$ and $O_1 \cap O_2 = \{\emptyset\}$. What's more, the numbers of rewarded and non-rewarded objects are denoted by n_1 and n_2 , respectively, $n_1 + n_2 = |O|$.

For each rewarded object, we track the ratings given by users who give the highest rating r_h at the early lifespan of the object. The number of these links D_i created by user u_i can be expressed as,

$$D_i = \sum_{o_\alpha \in O_1} D_{i\alpha}, \tag{1}$$

$$D_{i\alpha} = \begin{cases} 1 & \text{if } r_{i\alpha} = r_h, t_{i\alpha} \leq t_{\alpha 1} + (t_{\alpha p_\alpha} - t_{\alpha 1}) * \theta \\ 0 & \text{else} \end{cases}, \tag{2}$$

where $D_{i\alpha}$ is a binary event to measure whether the user u_i can make a high evaluation of object o_α ($o_\alpha \in O_1$) during the initial θ ($0 < \theta < 1$) of its lifespan, $t_{\alpha 1}$ and $t_{\alpha p_\alpha}$ are the timestamps of the first and last ratings the object o_α received, respectively. The quantity D_i is the number of identifying rewarded objects at their early lifespan for user u_i and $0 \leq D_i \leq n_1$. Meanwhile, θ is a tunable parameter and the value of D_i increases with the parameter θ . It should be noted that there is no rating to be considered ($D_i = 0, i = 1, 2, \dots, |U|$) when $\theta = 0$ and the whole lifespan is viewed as the initial lifespan when $\theta = 1$.

Finally, we define the perceptibility p_i as the proportion of D_i in the number of rewarded objects n_1 for user u_i ,

$$p_i = D_i / n_1. \tag{3}$$

Results

The identification of the user perceptibility could quantitatively measure the degree to which the user can identify the rewarded objects in their lifespan. To qualitatively measure whether a user is a perceptive user, we introduce a free-parameter bootstrap analysis [32–34]. The bootstrap sampling results show that, for the MovieLens and Netflix data sets, there are 5 and 27 identified perceptive users, respectively (accounting for 0.12% and 0.012% of all users, respectively). Here the parameter θ is set to 0.3 and 0.6 for the MovieLens and Netflix data sets, respectively. It should be noted that a larger parameter θ for the Netflix data set is selected due to the few rewarded objects with regard to the size of the whole objects and ratings (150 rewarded objects, 7803 objects and 37755925 ratings).

Moreover, we investigate whether the identification of user perceptibility is of significance. To this end, we calculate the average perceptibility of the first L users who give the rating 5 (the highest rating) in order of time for each object in two empirical data sets, denoted by $\langle p^L \rangle_\alpha$ for objects o_α . The parameter L is set to 10 in the following analysis. All objects are divided into two groups based on their corresponding average perceptibility $\langle p^L \rangle$: Objects rated by high perceptibility users (recorded as object set Θ) and the others (recorded as object set Λ), in which the objects in set Θ are selected as top q ($0 < q < 1$) high $\langle p^L \rangle$ objects. Firstly, we track the links attached to all objects in the future time window and calculate the average degree $\langle \rho_O(t) \rangle$ of two divided object groups as a function of time t , in which the length of the future time window are 100 and 200 days for the MovieLens and Netflix data sets, respectively. Fig 1 (a) and 1(b) shows the degree evolution of two divided object groups with the parameter $q = 10\%$ for the MovieLens and Netflix data sets, respectively. One can find that the average degrees of objects in set Λ in the future time window are larger than those of objects in set Θ , showing that the objects rated by high perceptibility users become less popular than the others, indicating that user perceptibility has little impact on finding the popular objects.

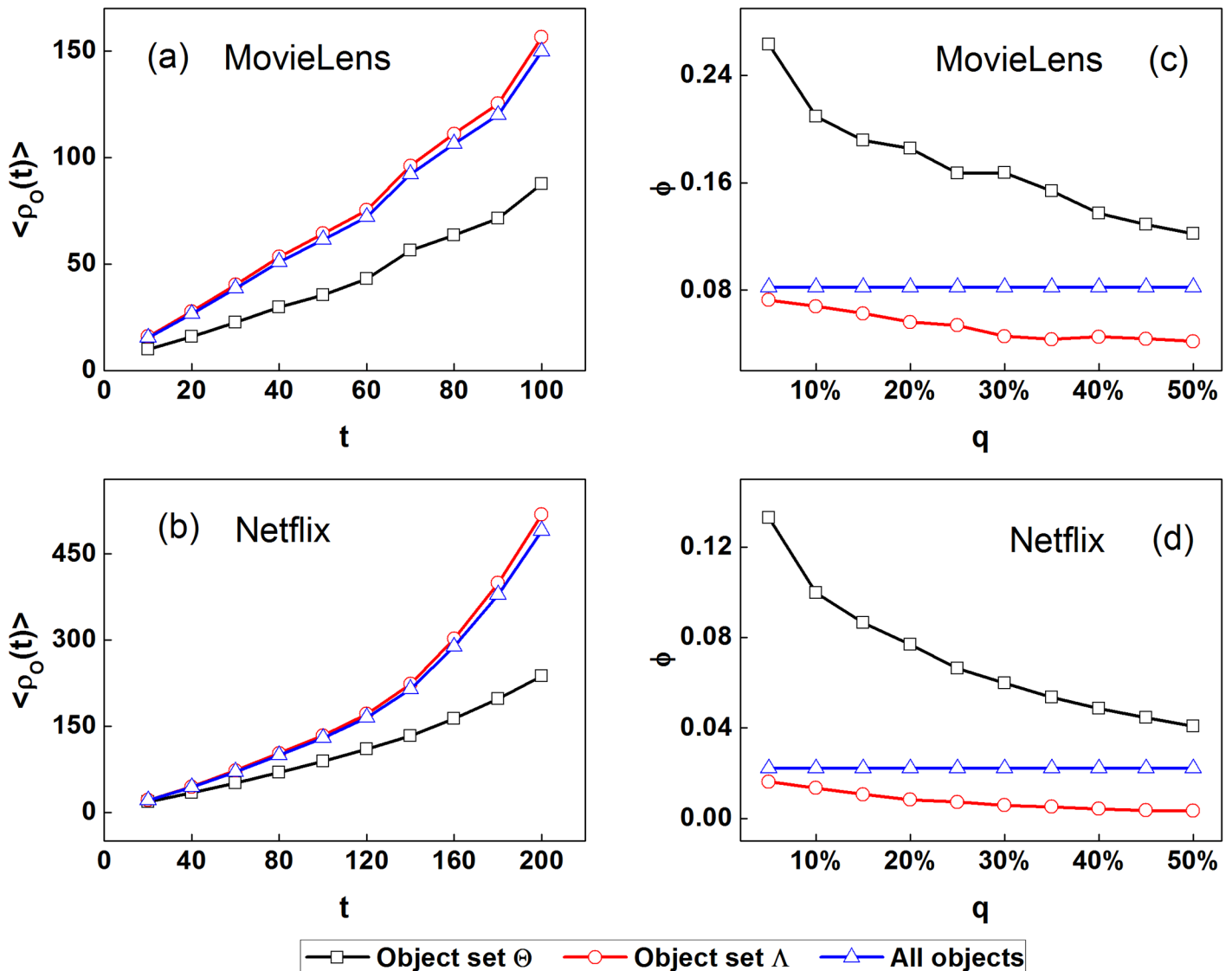


Fig 1. (Color online) Degree evolution $\langle \rho_o(t) \rangle$ of two divided object groups (a,b) and the ratio ϕ of rewarded objects in two divided object groups with different parameter q (c,d) for two empirical data sets, in which the time t is measured in days, and the parameter θ is set to 0.3 and 0.6 for the MovieLens and Netflix data sets, respectively. From the subplots (a,b) one can find that the average degrees of objects rated by high perceptibility users in the future time window are larger than those of the other objects. From the subplots (c,d) one can find that the ratio ϕ of rewarded objects in objects rated by high perceptibility users is higher than that in the other objects with different parameter q . The results indicate that the user perceptibility is helpful to find the potential rewarded objects.

<https://doi.org/10.1371/journal.pone.0178118.g001>

Subsequently, we investigate the ratio ϕ of rewarded objects in two divided object groups with different parameter q (Fig 1(c) and 1(d)). One can find that the ratio ϕ of rewarded objects in object set Θ is larger than that in object set Λ with different parameter q for two empirical data sets. For instance, the ratio ϕ of rewarded objects in object set Θ is larger than that in object set Λ by 263.0% and 722.0% with the parameter $q = 5\%$ for the MovieLens and Netflix data sets, respectively. Meanwhile, the ratios ϕ of rewarded objects in two divided object groups with $\theta = 0.2, 0.4$ for MovieLens and $\theta = 0.5, 0.7$ for Netflix show the similar results. Therefore, the results indicate that the user perceptibility is of significance in finding the rewarded objects rather than popular objects.

Furthermore, we investigate the relations between user perceptibility and user collective behavior patterns. All users are divided into two groups: High perceptibility users (recorded as user set Φ) and the others (recorded as user set Δ), in which the high perceptibility users are denoted as top q ($0 < q < 1$) high perceptibility users. We investigate the collective behavior patterns of two divided user groups from three aspects: User activity, correlation characteristics of user rating series and user reputation. User activity (denoted by k_U), namely user degree, is one of the most important user characteristics in social systems [27, 35]. The larger user degree, the more active the user would be. In our analysis, correlation characteristics of user rating series is described by detrended fluctuation analysis (short for DFA), which is widely used for analyzing the statistical self-affinity of a time series [36–39], calculated by the scaling exponent η . The quantity $\eta > 0$: $\eta < 0.5$ corresponds to anti-correlated series; $\eta = 0.5$ corresponds to uncorrelated white noise; $\eta > 0.5$ corresponds to correlated series. User reputation is proposed to measure the user ability of rating accurate assessments of various objects [40, 41]. So far, many reputation ranking methods have been widely investigated [42–44]. In this paper, we use the correlation based ranking algorithm [41] to calculate the user reputation denoted by μ . The quantity μ lies in $[0, 1]$ and larger μ means higher user reputation.

Fig 2 shows the average degree $\langle k_U \rangle$, scaling exponent $\langle \eta \rangle$, reputation $\langle \mu \rangle$ of two divided user groups with different parameter q for the MovieLens and Netflix data sets, respectively. One can find that the average $\langle k_U \rangle$, $\langle \eta \rangle$ and $\langle \mu \rangle$ of user set Φ (high perceptibility users) are larger than those of user set Δ (the other users) with different parameter q for two empirical data sets. For instance, the average $\langle k_U \rangle$, $\langle \eta \rangle$ and $\langle \mu \rangle$ of user set Φ are larger than the ones of

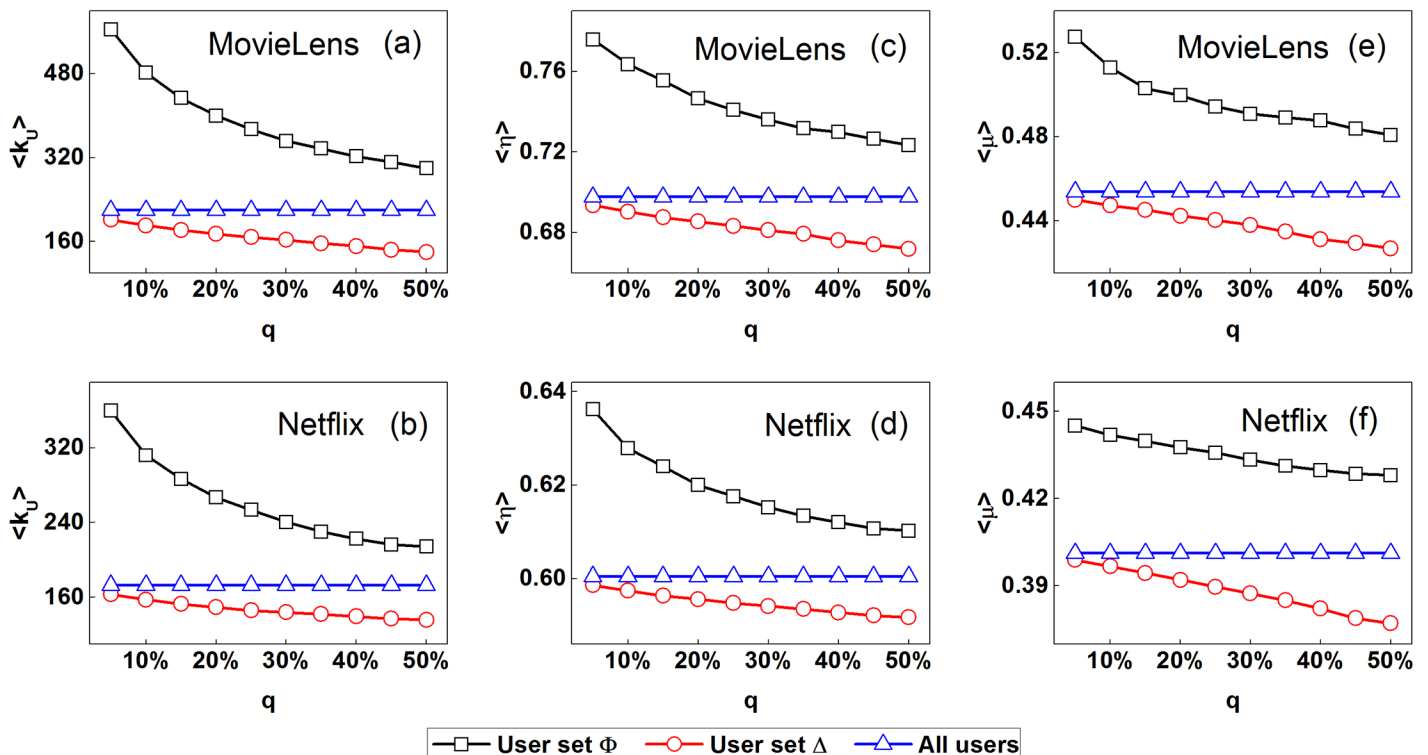


Fig 2. (Color online) The average degree $\langle k_U \rangle$, scaling exponent $\langle \eta \rangle$, reputation $\langle \mu \rangle$ of two divided user groups with different parameter q for (a, c, e) MovieLens and (b, d, f) Netflix data sets, in which the parameter θ is set to 0.3 and 0.6, respectively. One can find that the average $\langle k_U \rangle$, $\langle \eta \rangle$ and $\langle \mu \rangle$ of high perceptibility users are larger than those of the other users with different parameter q for two empirical data sets, which indicates that high perceptibility users show different collective behavior patterns than the other users: Larger activity, stronger correlation of rating series and higher reputation.

<https://doi.org/10.1371/journal.pone.0178118.g002>

user set Δ by 180.1%, 11.8% and 17.3%, respectively with the parameter $q = 5\%$ for the MovieLens data set. For the Netflix data set, the increases are 120.5%, 6.3% and 11.6%, respectively with the parameter $q = 5\%$. The collective behavior patterns of two divided user groups with $\theta = 0.2, 0.4$ for MovieLens and $\theta = 0.5, 0.7$ for Netflix show the similar results. The results indicate that high perceptibility users show larger activity, stronger correlation of rating series and higher reputation than other users.

A framework for high perceptibility user identification

High perceptibility users are denoted as top $q(0 < q < 1)$ high perceptibility users, the identification of high perceptibility users is closely linked with the identification of user perceptibility. User perceptibility is calculated by tracking the ratings to the rewarded objects, while the discovery of the rewarded objects has hysteresis. With the growing amount of new users, objects and the corresponding ratings, the rewarded objects of the current rating systems are uncertain. Thus, the user perceptibility and high perceptibility users cannot be identified in real time. In terms of the fact that high perceptibility users have specific collective behavior patterns, we develop a general framework for identifying high perceptibility users based on users' behavior patterns.

All users are divided into two groups: High perceptibility users and the others. Given that identifying high perceptibility users belongs to a classification problem, random forests [45], one of the most widely used machine learning [46, 47] methods, is introduced in our framework. The Data Flow Diagram (short for DFD) of the framework is shown in Fig 3. Firstly, the available ratings and the rewarded objects are calculated to identify the user perceptibility using the presented method (Process P1). Meanwhile, the available ratings are used to analyze the user collective behavior patterns from three aspects: Degree, DFA of rating series and reputation (Process P2). The process P1 and P2 could be performed simultaneously. Then, we use the

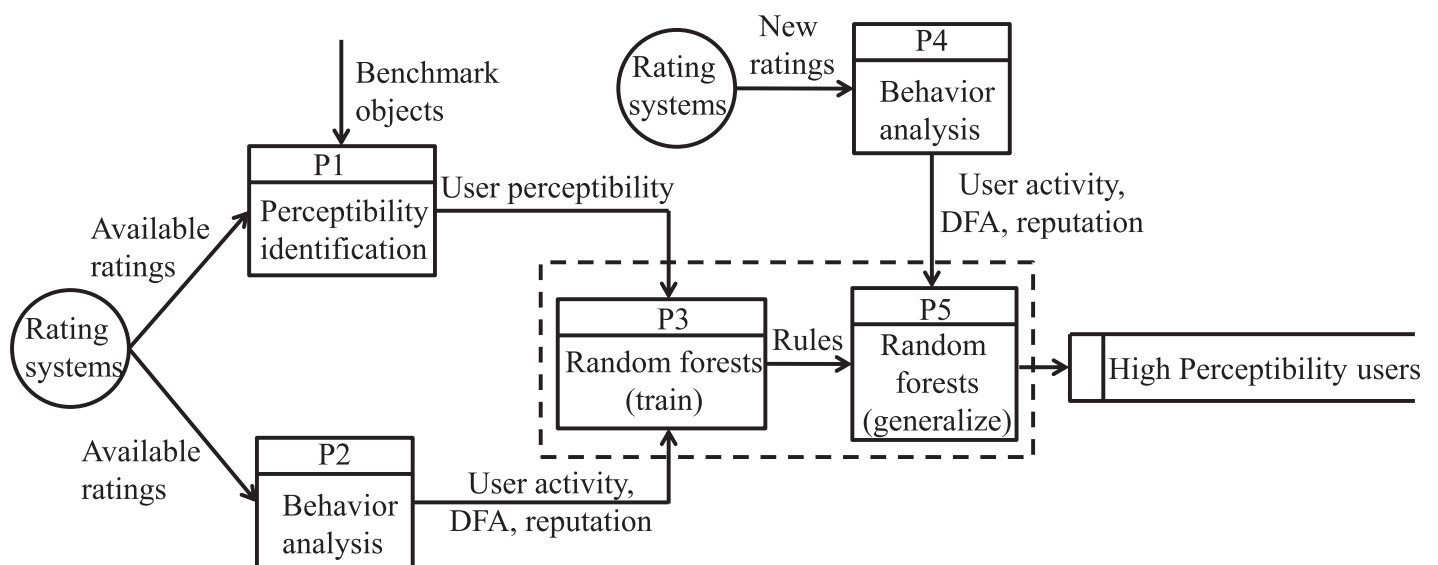


Fig 3. The Data Flow Diagram (DFD) of the framework for high perceptibility user identification. The available ratings in the rating systems, on the one hand, are applied with the rewarded objects to identify the user perceptibility by the presented method (Process P1). On the other hand, they are used to analyze the user collective behavior patterns described by three aspects: Activity, DFA of rating series and reputation (Process P2). Then, we use the random forests to train the obtained results containing the user perceptibility and behavior patterns (Activity, DFA of rating series and reputation) (Process P3). Afterwards, the high perceptibility users will be identified based on the user collective behavior patterns analysed from the new ratings in the rating systems (Process P4) by the generalization of random forests (Process P5).

<https://doi.org/10.1371/journal.pone.0178118.g003>

random forests to train the obtained results, which contain the user perceptability and behavior patterns (Process P3). When the rating systems generate new ratings, the user collective behavior patterns analysed based on the new ratings (Process P4) are used to identify high perceptability users in the current rating systems by the generalization of random forests (Process P5).

Moreover, we investigate the performance of high perceptability user identification using the presented framework. After identifying the user perceptability based on the rewarded objects and ratings, high perceptability users are classified as top $q(0 < q < 1)$ high perceptability users. We select 70% of user data (user perceptability and behavior patterns) as the training set S_{tr} and the remaining 30% as the test set S_{te} for the MovieLens and Netflix data sets, respectively. High perceptability users in the test set S_{te} are denoted as set H_{te} . Meanwhile, the identified high perceptability user set H'_{te} in the test set S_{te} will be predicted by the generalization of random forests after training the data of the training set S_{tr} . Then, the performance of high perceptability user identification is measured by the precision P , recall R and F-measure F ,

$$P = \frac{|H_{te} \cap H'_{te}|}{|H'_{te}|}, \tag{4}$$

$$R = \frac{|H_{te} \cap H'_{te}|}{|H_{te}|}, \tag{5}$$

$$F = \frac{2 * P * R}{P + R}, \tag{6}$$

where $|H_{te} \cap H'_{te}|$ is the number of high perceptability users in the identified high perceptability user set H'_{te} . $|H'_{te}|$ is the number of users in the identified high perceptability user set H'_{te} . And $|H_{te}|$ is the number of users in the high perceptability user set H_{te} . Precision P , recall R and F-measure F all lie in $[0, 1]$ and larger P , R or F represents better performance of high perceptability user identification. The precision P , recall R and F-measure F with different parameter q for two empirical data sets are shown in Fig 4, in which the parameter $q(0 < q < 1)$ represents the ratio of the high perceptability users in all users. One can find that the framework can perform well in identifying the high perceptability users. The precision P , recall R and F-measure F could reach $P = 0.68$, $R = 0.66$ and $F = 0.67$ with $q = 50\%$ for the MovieLens data set, and for

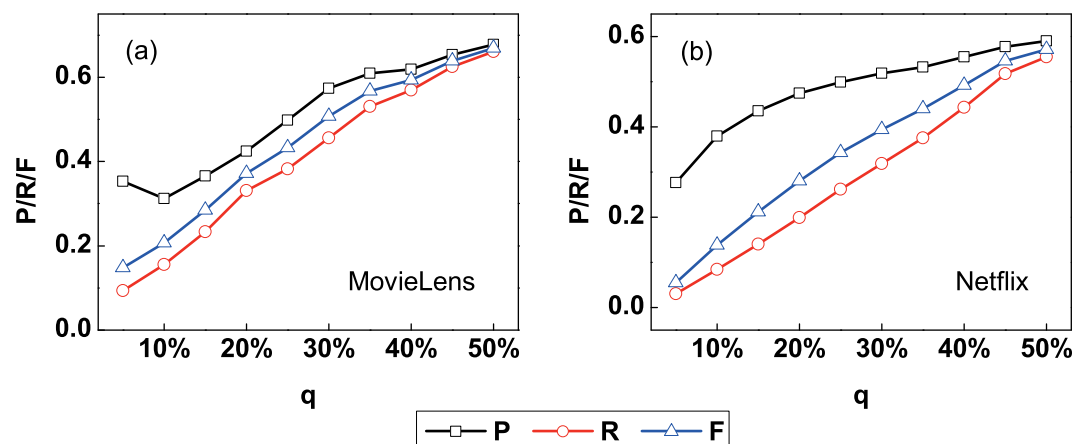


Fig 4. The precision P , recall R and F-measure F of identifying the high perceptability users in the framework with different parameter q for (a) MovieLens and (b) Netflix data sets, respectively. One can find that the precision P , recall R and F-measure F all increase with the parameter q in general, and the presented framework can perform well in identifying the high perceptability users.

<https://doi.org/10.1371/journal.pone.0178118.g004>

the Netflix data set, the performance achieves $P = 0.59$, $R = 0.55$ and $F = 0.57$. Meanwhile, the precision P , recall R and F-measure F all increase with the parameter q in general. The performances of high perceptibility user identification with different parameter θ indicate that larger precision P , recall R and F-measure F are obtained in the case of larger parameter θ with different parameter q .

We can obtain both the performance of high perceptibility user identification and importance of behavior patterns using random forests. Besides the random forests, we also use other machine learning methods including gradient boosting machine [48, 49] (short for GBM) and support vector machine [50, 51] (short for SVM) to identify the high perceptibility users. The

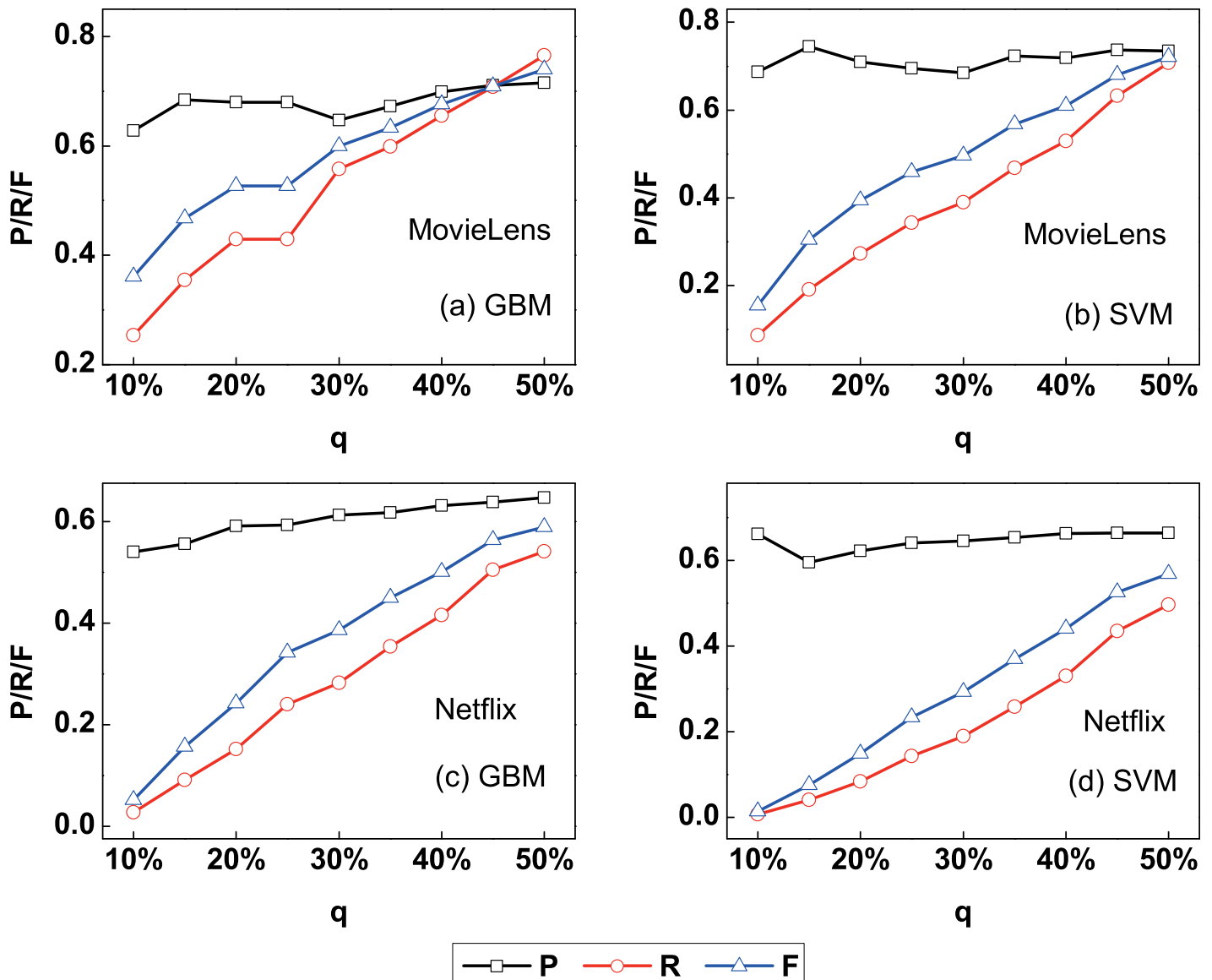


Fig 5. The precision P , recall R and F-measure F of identifying the high perceptibility users in the framework with different machine learning methods (GBM and SVM) and different parameter q for (a-b) MovieLens and (c-d) Netflix data sets, respectively. One can find that the recall R of high perceptibility user identification using GBM and SVM have little difference with the results using random forests. The precision P is better using GBM and SVM than using random forests.

<https://doi.org/10.1371/journal.pone.0178118.g005>

precision P , recall R and F-measure F of high perceptibility user identification are shown in Fig 5, from which one can find that the performance could reach $P = 0.72$, $R = 0.77$ and $F = 0.74$ using GBM and $P = 0.74$, $R = 0.71$ and $F = 0.72$ using SVM with $q = 50\%$ for the MovieLens data set. The recall of high perceptibility user identification using GBM and SVM have little difference with the results using random forests. While the precision of high perceptibility user identification using GBM and SVM are different, the precision P is large when the parameter q is small. The precision is better using GBM and SVM than using random forests.

Conclusion and discussions

In this paper, taking into account collective behavior patterns and the heterogeneity of online users, we present the definition of perceptive user, which is defined as the user who can make high evaluations of the rewarded objects at their early lifespan. In addition, user perceptibility is defined as the degree to which the user can identify the rewarded objects in their initial lifespan. Then, we present a method for identifying the user perceptibility by tracking the ratings given to rewarded objects and the timestamps. Meanwhile, to track out the relations between user perceptibility and user collective behavior patterns, we investigate the user behavior patterns from three aspects: User activity, correlation characteristics of user rating series and user reputation. The experimental results for the MovieLens and Netflix data sets indicate that high perceptibility users have larger activity, stronger correlation of rating series and higher reputation than the other users. For the MovieLens data set, the average $\langle k_U \rangle$, $\langle \eta \rangle$ and $\langle \mu \rangle$ of user set Φ (high perceptibility users) are larger than those of user set Δ by 180.1%, 11.8% and 17.3%, respectively, with the parameter $q = 5\%$, for example. Finally, given that there exists hysteresis in finding the rewarded objects, we present a general framework to identify the high perceptibility users in real time based on users' behavior patterns. The experimental results show that the framework can perform well in identifying the high perceptibility users. The precision P , recall R and F-measure F could reach $P = 0.72$, $R = 0.77$ and $F = 0.74$ with $q = 50\%$ for the MovieLens data set, for example.

The computational complexity of the method presented to identify the user perceptibility is $O(n_1 \cdot \langle \rho_O \rangle + n_1 \cdot |U|)$, where the first term accounts for the calculation of $D_{i\omega}$, the results whether the user can make a high evaluation for each rewarded object in its initial lifespan. And the second term accounts for the calculation of D_b , the number of identifying the rewarded objects at their early lifespan for each user. Substituting the inequality $\langle \rho_O \rangle \leq |U|$, we are left with $O(n_1 \cdot |U|)$. Due to n_1 is a constant value in a certain rating system, one has the fact that the computational complexity of the user perceptibility identification is $O(|U|)$, a linear function of the user size.

For a long time, popular objects are more concerned targets, while few users recognize and appreciate the rewarded objects when they are rarely concerned. The discovery of perceptive users and the identification of user perceptibility provides us a new perspective of understanding these special users. The results that user perceptibility can be helpful to find the potential rewarded objects indicate the identification of user perceptibility is of practical significance in e-commerce and marketing. Meanwhile, the presented framework for high perceptibility user identification, from investigating behavior patterns of two divided user groups to conversely identifying high perceptibility users based on the behavior patterns, gives us a systematic study of the perceptive users and it is also suitable for big data processing. In addition, the following points should be addressed in the future work. Firstly, the high-quality objects here are generated based on the rewarded ones, how to construct the high-quality object set is an open problem. Secondly, the user collective behavior patterns are investigated from three aspects in this paper, which may be found incompletely. As further improvement, we could consider more

dimensions to deeper explore the user behavior patterns. Thirdly, random forests is applied in the framework for high perceptibility user identification, in which econometrics and time series analysis could be emphasized as well in our future research.

Supporting information

S1 Text. The bootstrap analysis.

(TEX)

S2 Text. The ratio ϕ of benchmark objects in two divided object groups with different parameter θ .

(TEX)

S3 Text. The calculation process of DFA of user rating series.

(TEX)

S4 Text. The calculation process of user reputation.

(TEX)

S5 Text. The collective behavior patterns of two divided user groups with different parameter θ .

(TEX)

S6 Text. The identification performances of high perceptibility users with different parameter θ .

(TEX)

S7 Text. The importance of behavior patterns in random forests.

(TEX)

S8 Text. The identification performances of high perceptibility users with RBPD reputation ranking method.

(TEX)

S1 Fig. (Color online) Zipf plots of user perceptibility in real data and bootstrap analysis for (a) MovieLens and (b) Netflix data sets, in which the parameter θ is set to 0.3 and 0.6, respectively.

(EPS)

S2 Fig. The ratio ϕ of benchmark objects in two divided object groups with different parameter q for (a-b) MovieLens ($\theta = 0.2, \theta = 0.4$) and (c-d) Netflix ($\theta = 0.5, \theta = 0.7$) data sets, respectively.

(EPS)

S3 Fig. The average degree $\langle k_U \rangle$, scaling exponent $\langle \eta \rangle$, reputation $\langle \mu \rangle$ of two divided user groups with different parameter q in the case of $\theta = 0.2, \theta = 0.4$ for the MovieLens data set.

(EPS)

S4 Fig. The average degree $\langle k_U \rangle$, scaling exponent $\langle \eta \rangle$, reputation $\langle \mu \rangle$ of two divided user groups with different parameter q in the case of $\theta = 0.5, \theta = 0.7$ for the Netflix data set.

(EPS)

S5 Fig. The precision P , recall R and F-measure F of identifying high perceptibility users in the framework with different parameter q for (a-b) MovieLens ($\theta = 0.2, \theta = 0.3, \theta = 0.4$) and (c-d) Netflix ($\theta = 0.5, \theta = 0.6, \theta = 0.7$) data sets, respectively.

(EPS)

S6 Fig. The precision P , recall R and F-measure F of identifying the high perceptibility users in the framework with different parameter q for (a) MovieLens and (b) Netflix data sets, respectively.

(EPS)

S1 Table. The importance of user behavior patterns in random forests for the MovieLens data set.

(TEX)

S2 Table. The importance of user behavior patterns in random forests for the Netflix data set.

(TEX)

S3 Table. The importance of user behavior patterns in random forests for the MovieLens data set.

(TEX)

S4 Table. The importance of user behavior patterns in random forests for the Netflix data set.

(TEX)

Acknowledgments

We acknowledge Dr. Zhenhua Yang's help and GroupLens Research Group for providing us MovieLens data and the Netflix Inc. for Netflix data. This work is supported by the National Natural Science Foundation of China (Grant Nos. 71271126, 61374177, 71371125), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shuguang Program Project of Shanghai Educational Committee (Grant No. 14SG42).

Author Contributions

Conceptualization: Jian-Guo Liu, Xiao-Lu Liu.

Data curation: Jian-Guo Liu, Xiao-Lu Liu.

Formal analysis: Jian-Guo Liu, Xiao-Lu Liu, Qiang Guo, Jing-Ti Han.

Funding acquisition: Jian-Guo Liu, Qiang Guo, Jing-Ti Han.

Investigation: Jian-Guo Liu, Xiao-Lu Liu.

Methodology: Jian-Guo Liu, Xiao-Lu Liu, Qiang Guo, Jing-Ti Han.

Project administration: Jian-Guo Liu, Qiang Guo, Jing-Ti Han.

Resources: Jian-Guo Liu, Xiao-Lu Liu.

Software: Jian-Guo Liu, Xiao-Lu Liu.

Supervision: Jian-Guo Liu, Qiang Guo, Jing-Ti Han.

Validation: Jian-Guo Liu, Xiao-Lu Liu.

Visualization: Jian-Guo Liu, Xiao-Lu Liu.

Writing – original draft: Jian-Guo Liu, Xiao-Lu Liu, Qiang Guo, Jing-Ti Han.

Writing – review & editing: Jian-Guo Liu, Xiao-Lu Liu, Qiang Guo, Jing-Ti Han.

References

1. Oliveira J G, Barabási A L. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*. 2005; 437: 1251–1252. <https://doi.org/10.1038/4371251a> PMID: 16251946
2. Muchnik L, Aral S, Taylor S J. Social influence bias: a randomized experiment. *Science*. 2013; 341: 647–651. <https://doi.org/10.1126/science.1240466> PMID: 23929980
3. Wu Y, Zhou C, Xiao J, Kurths J, Schellnhuber HJ. Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences*. 2010; 107: 18803–18808. <https://doi.org/10.1073/pnas.1013140107>
4. Barabási A L. The origin of bursts and heavy tails in human dynamics. *Nature*. 2005; 435: 207–211. <https://doi.org/10.1038/nature03459> PMID: 15889093
5. Malmgren R D, Stouffer D B, Motter A E, Amaral L A. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*. 2008; 105: 18153–18158. <https://doi.org/10.1073/pnas.0800332105>
6. Vázquez A. Exact results for the Barabási model of human dynamics. *Physical Review Letters*. 2005; 95: 248701. <https://doi.org/10.1103/PhysRevLett.95.248701> PMID: 16384430
7. Vázquez A, Oliveira JG, Dezső Z, Goh K I, Kondor I, Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*. 2006; 73: 036127. <https://doi.org/10.1103/PhysRevE.73.036127>
8. Han X P, Zhou T, Wang B H. Modeling human dynamics with adaptive interest. *New Journal of Physics*. 2008; 10: 073010. <https://doi.org/10.1088/1367-2630/10/7/073010>
9. Zhao Z D, Yang Z, Zhang Z, Zhou T, Huang Z G, Lai Y C. Emergence of scaling in human-interest dynamics. *Scientific Reports*. 2013; 3: 3472. <https://arxiv.org/abs/1307.7796v1> <https://doi.org/10.1038/srep03472> PMID: 24326949
10. Simini F, González MC, Maritan A, Barabási A L. A universal model for mobility and migration patterns. *Nature*. 2012; 484: 96–100. <https://doi.org/10.1038/nature10856> PMID: 22367540
11. Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H, et al. Identification of influential spreaders in complex networks. *Nature Physics*. 2010; 6: 888–893. <https://doi.org/10.1038/nphys1746>
12. Moreno Y, Nekovee M, Pacheco A F. Dynamics of rumor spreading in complex networks. *Physical Review E*. 2004; 69: 066130. <https://doi.org/10.1103/PhysRevE.69.066130>
13. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. *Physical Review Letters*. 2001; 86: 3200. <https://doi.org/10.1103/PhysRevLett.86.3200> PMID: 11290142
14. Weng L, Menczer F, Ahn Y Y. Virality prediction and community structure in social networks. *Scientific Reports*. 2013; 3: 2522. <https://doi.org/10.1038/srep02522> PMID: 23982106
15. Liu J G, Lin J H, Guo Q, Zhou T. Locating influential nodes via dynamics-sensitive centrality. *Scientific Reports*. 2016; 6: 21380. <https://doi.org/10.1038/srep21380> PMID: 26905891
16. Zhou T, Kuscsik Z, Liu J G, Medo M, Wakleing J R, Zhang Y C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*. 2010; 107: 4511–4515. <https://doi.org/10.1073/pnas.1000488107>
17. Liu J G, Zhou T, Guo Q. Information filtering via biased heat conduction. *Physical Review E*. 2011; 84: 037101. <https://doi.org/10.1103/PhysRevE.84.037101>
18. Liu J G, Hou L, Pan X, Guo Q, Zhou T. Stability of similarity measurements for bipartite networks. *Scientific Reports*. 2016; 6: 18653. <https://doi.org/10.1038/srep18653> PMID: 26725688
19. Hou L, Pan X, Guo Q, Liu J G. Memory effect of the online user preference. *Scientific Reports*. 2014; 4: 6560. <https://arxiv.org/abs/1409.4403> <https://doi.org/10.1038/srep06560> PMID: 25308573
20. Yang Z, Zhang Z K, Zhou T. Anchoring bias in online voting. *Europhysics Letters*. 2012; 100: 68002. <http://iopscience.iop.org/article/10.1209/0295-5075/100/68002/meta>
21. Ji L, Liu J G, Hou L, Guo Q. Identifying the role of common interests in online user Trust formation. *PLoS ONE*. 2015; 10: e0121105. <https://doi.org/10.1371/journal.pone.0121105> PMID: 26161853
22. Zhang Q M, Zeng A, Shang M S. Extracting the information backbone in online system. *PLoS ONE*. 2013; 8: e62624. <https://doi.org/10.1371/journal.pone.0062624> PMID: 23690946
23. Zhang Y L, Guo Q, Ni J, Liu J G. Memory effect of the online rating for movies. *Physica A*. 2015; 417: 261–266. <https://doi.org/10.1016/j.physa.2014.09.012>
24. Medo M, Wakeling J R. The effect of discrete vs. continuous-valued ratings on reputation and ranking systems. *Europhysics Letters*. 2010; 91: 48004. <http://iopscience.iop.org/article/10.1209/0295-5075/91/48004/meta>
25. Barabási A L, Albert R. Emergence of scaling in random networks. *Science*. 1999; 286: 509–512. <https://doi.org/10.1126/science.286.5439.509> PMID: 10521342

26. Barabási A L. Scale-free networks: a decade and beyond. *Science*. 2009; 325: 412–413. <https://doi.org/10.1126/science.1173299> PMID: 19628854
27. Ni J, Zhang Y L, Hu Z L, Song W J, Hou L, Guo Q, et al. Ceiling effect of online user interests for the movies. *Physica A*. 2014; 402: 134–140. <https://doi.org/10.1016/j.physa.2014.01.046>
28. Albert R, Jeong H, Barabási A L. Error and attack tolerance of complex networks. *Nature*. 2000, 406: 378–382. <https://doi.org/10.1038/35019019> PMID: 10935628
29. Strogatz S H. Exploring complex networks. *Nature*. 2001, 410: 268–276. <https://doi.org/10.1038/35065725> PMID: 11258382
30. Kurant M, Thiran P, Hagmann P. Error and attack tolerance of layered complex networks. *Physical Review E*. 2007, 76: 026103. <https://doi.org/10.1103/PhysRevE.76.026103>
31. Motter A E, De Moura A P S, Lai Y C, Dasgupta P. Topology of the conceptual network of language. *Physical Review E*. 2002, 65: 065102. <https://doi.org/10.1103/PhysRevE.66.065102>
32. Efron B. Better bootstrap confidence intervals. *Journal of the American statistical Association*. 1987; 82: 171–185. <https://doi.org/10.2307/2289153>
33. Efron B. Bootstrap methods: another look at the jackknife. Springer New York. 1992. http://link.springer.com/chapter/10.1007/978-1-4612-4380-9_41
34. Athreya K B. Bootstrap of the mean in the infinite variance case. *The Annals of Statistics*. 1987; 15: 724–731. <http://www.jstor.org/stable/2241336> <https://doi.org/10.1214/aos/1176350371>
35. Liu X L, Guo Q, Hou L, Cheng C, Liu J G. Ranking online quality and reputation via the user activity. *Physica A*. 2015; 436: 629–636. <https://doi.org/10.1016/j.physa.2015.05.043>
36. Peng C K, Buldyrev S V, Havlin S, Simons M, Stanley H E, Goldberger A L. Mosaic organization of DNA nucleotides. *Physical Review E*, 1994, 49: 1685. <https://doi.org/10.1103/PhysRevE.49.1685>
37. Peng C K, Havlin S, Stanley H E, Goldberger A L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*. 1995; 5: 82. <https://doi.org/10.1063/1.166141> PMID: 11538314
38. Rybski D, Buldyrev S V, Havlin S, Liljeros F, Makse H A. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*. 2009; 106: 12640–12645. <https://doi.org/10.1073/pnas.0902667106>
39. Rybski D, Buldyrev S V, Havlin S, Liljeros F, Makse H A. Communication activity in a social network relation between long-term correlations and interevent clustering. *Scientific Reports*. 2012; 2: 560. <https://doi.org/10.1038/srep00560> PMID: 22876339
40. Laureti P, Moret L, Zhang Y C, Yu Y K. Information filtering via iterative refinement. *Europhysics Letters*. 2006; 75: 1006. <http://iopscience.iop.org/article/10.1209/epl/12006-10204-8/meta>
41. Zhou Y B, Lei T, Zhou T. A robust ranking algorithm to spamming. *Europhysics Letters*. 2011; 94: 48002. <http://iopscience.iop.org/0295-5075/94/4/48002>
42. Liao H, Zeng A, Xiao R, Ren Z M, Chen D, Zhang Y C. Ranking reputation and quality in online rating systems. *PLoS ONE*. 2014; 9: e97146. <https://doi.org/10.1371/journal.pone.0097146> PMID: 24819119
43. Gao J, Dong Y W, Shang M S, Cai S M, Zhou T. Group-based ranking method for online rating systems with spamming attacks. *Europhysics Letters*. 2015; 110: 28003. <http://iopscience.iop.org/0295-5075/110/2/28003>
44. Liu X L, Liu J G, Yang K, Guo Q, Han J T. Identifying online user reputation of user object bipartite networks. *Physica A*. 2017; 467: 508–516. <https://doi.org/10.1016/j.physa.2016.10.031>
45. Breiman L. Random forests. *Machine learning*. 2001; 45: 5–32.
46. Goldberg D E, Holland J H. Genetic algorithms and machine learning. *Machine learning*. 1988; 3: 95–99. <https://doi.org/10.1007/BF00113892>
47. Cortes C, Vapnik V. Support-vector networks. *Machine learning*. 1995; 20: 273–297. <https://doi.org/10.1007/BF00994018>
48. Friedman J H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 2001, 29: 1189–1232. <http://www.jstor.org/stable/2699986>
49. Friedman J H. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002, 38: 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
50. Suykens J A K, Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters*, 1999, 9: 293–300. <https://doi.org/10.1023/A:1018628609742>
51. Furey T S, Cristianini N, Duffy N, Bednarski D, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 2000, 16: 906–914. <https://doi.org/10.1093/bioinformatics/16.10.906> PMID: 11120680