Physica A 473 (2017) 546–560

Contents lists available at ScienceDirect

Physica A

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

Evaluating user reputation in online rating systems via an iterative group-based ranking method

Jian Gao [a,*], Tao Zhou [a,b]

[a] CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China
[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

HIGHLIGHTS

- We evaluated online reputation based on users' grouping behaviors instead of the traditional objects' quality-based assumption.
- We introduced an iterative reputation–allocation process, which improved the method's robustness in resisting spamming attacks.
- We found the users' grouping behaviors have advantages in designing better online reputation systems.

ARTICLE INFO

Article history:
Received 10 March 2016
Received in revised form 11 June 2016
Available online 12 January 2017

Keywords:
Rating systems
Reputation evaluation
Ranking method
Iterative refinement
Spamming attack

ABSTRACT

Reputation is a valuable asset in online social lives and it has drawn increased attention. Due to the existence of noisy ratings and spamming attacks, how to evaluate user reputation in online rating systems is especially significant. However, most of the previous ranking-based methods either follow a debatable assumption or have unsatisfied robustness. In this paper, we propose an iterative group-based ranking method by introducing an iterative reputation–allocation process into the original group-based ranking method. More specifically, the reputation of users is calculated based on the weighted sizes of the user rating groups after grouping all users by their rating similarities, and the high reputation users' ratings have larger weights in dominating the corresponding user rating groups. The reputation of users and the user rating group sizes are iteratively updated until they become stable. Results on two real data sets with artificial spammers suggest that the proposed method has better performance than the state-of-the-art methods and its robustness is considerably improved comparing with the original group-based ranking method. Our work highlights the positive role of considering users' grouping behaviors towards a better online user reputation evaluation.

1. Introduction

At the age of Internet, individual reputation plays the role of fundamental blocks in building up online ecosystems [1,2]. For example, the reputation mechanisms help to realize the governance structures of social commerce in the field of e-commerce [3], the online reputations of job applicants are helpful to employers in making better hiring choices, and the online reputations of users have been used to enhance the performance of recommender systems [4,5]. Moreover, online reputations can potentially affect the accuracy of the information that we obtained [6], since initially localized

* Corresponding author.
E-mail address: gaojian08@hotmail.com (J. Gao).

http://dx.doi.org/10.1016/j.physa.2017.01.055
0378-4371/

biased information or even rumors from untrusted users are much easier to propagate throughout the whole online social networks [7] due to the rapid development of online social media in recent years [8]. Therefore, new challenges arise that how to build and maintain user reputation in online communities? Normally, quantifying the credibility of users by analyzing their behavioral patterns is a good way to address this issue [9,10].

In recent years, many online platforms have implemented web-based rating systems [1,11,12], such as Amazon, eBay, MovieLens, Netflix, Epinions, etc. On these collaborative online communities, users are encouraged to give ratings to various objects to share their feedbacks and shopping experiences. On the one side, the ratings that an object have received are the direct reflections of the object's true quality, which has a significant impact on users' purchasing decisions [13,14]. Usually, the following buyers prefer the higher rated objects to the lower rated ones [15], and the objects with higher ratings have advantage in receiving the following recommendation [11]. On the other side, the user rating behaviors can also be the reflections of their credibility [16]. In reality, there are some users who give unreasonable ratings due to their poor judgments [17,18] or give maximal/minimal ratings to purposefully guide the public choices [19,20]. The behaviors of these users have negative effect on the reliability of online rating systems [21,22], and these noisy and biased ratings are widely found in practical systems [23,24]. Therefore, how to extract credible information from abundant feedbacks, measure the credibility of users, and filter out untrusted users are becoming urgent tasks for online rating systems [25,26].

One of the most straightforward ways to cope with these concerns is introducing online user reputation evaluation systems into the web-based rating platforms [27]. Some previous works have been done in modeling and managing online reputation [28–30]. These reputation evaluation systems are capable of decision support for Internet-mediated services and help to maintain the healthy development of online rating systems and recommender systems. As the core of online reputation systems, a variety of user reputation evaluation methods have been proposed [31,32], among which the quality-based ranking methods are the most popular ones. Underlying an assumption that each object has a most objective rating that best reflects its quality [33], the quality-based ranking methods measure a user's reputation by the difference between the rating values and the estimated objects' quality values [34]. These methods include the iterative refinement (IR) method [35], an improved IR method [36], the correlation-based ranking (CR) method [37], the reputation redistribution ranking (RR) method [38] and the other seven methods [32,39]. These aforementioned methods are well-performed in online user reputation evaluation, however, some of them may not converge and some others are not robust to spamming attacks [39,40]. Moreover, due to the fact that the online rating system is fundamentally a socialized information collection platform, one object should accept multiple reasonable ratings [33] since the ratings are subjective and can be affected by users' background and some other factors [15,24]. Therefore, the assumption of the quality-based ranking methods is worthy of scrutiny.

Very recently, a group-based ranking (GR) method [41] is proposed without following the debatable assumption in quality-based methods. In the GR method, the reputation of users is calculated by the corresponding group sizes after grouping all users according to their rating similarities. Specifically, users are assigned with high reputation if they always fall into large rating groups. Although this method has better performance in evaluating user reputation on data sets with spamming attacks, it is not robust for plenty of large-degree spammers mainly due to the reason that the ratings are evenly contributed regardless of users' reputation in calculating the corresponding group sizes in the one-step process. In fact, the ratings from users of higher reputation should play a more important role in dominating the corresponding group sizes while the contributions of lower reputation users' ratings should be reduced. Moreover, recent literatures on physical dynamics, such as the HITS algorithm with iterative refinement procedure [42] and the original resource-allocation process [43,44], also provide us a promising way to improve the robustness of the original method by adopting a reputation–allocation procedure.

In this paper, we propose an iterative group-based ranking (IGR) method by introducing an iterative reputation–allocation process into the original group-based ranking method. Specifically, ratings from higher reputation users are assigned with larger weights in calculating the corresponding group sizes. Both the user reputation and the group sizes are iteratively updated until they become stable. In fact, the proposed IGR method is partially inspired by the group-based ranking method, the quality-based ranking methods and iterative refinement procedures. When tested on two real data sets with artificial spammers, the proposed IGR method has better performance than the state-of-the-art methods in online user reputation evaluation and its robustness in resisting a large number of spamming attacks is considerably improved compared with the original group-based ranking method. Further, we provide some insights on the mechanisms and analyze the characteristics of the proposed method and some other benchmark methods. Our work provides a further understanding on some reputation evaluation methods and highlights the positive role of considering users' grouping behaviors in designing better reputation evaluation systems.

The rest of this paper is organized as follows. Section 2 presents related works on online ranking methods and iterative refinement procedures. The proposed iterative group-based ranking method is described in Section 3. The data sets and metrics for performance evaluation are introduced in Section 4. The experiments and results are shown in Section 5. At last, the conclusions and discussion are given in Section 6.

2. Related work

This work is particularly inspired by a large body of prior works, including the quality-based ranking methods, the group-based ranking method and the iterative refinement procedures. In this section, we will first define some basic notations for the online rating systems. Then, three representative quality-based ranking methods will be briefly reviewed together,

namely, the iterative refinement (IR) [35], the correlation-based ranking (CR) [37], and the reputation redistribution (RR) [38]. Hereafter, we will describe the group-based ranking (GR) [41] and two iterative refinement procedures, namely, the HITS algorithm [42] and the original resource-allocation process [43,44].

2.1. Basic notation

Some basic notations for the online rating systems and the user reputation evaluation methods are introduced in the following. The online rating system can be described by a weighed bipartite network $G = \{U, O, E\}$, where $U = \{U_1, U_2, \ldots, U_m\}$, $O = \{O_1, O_2, \ldots, O_n\}$ and $E = \{E_1, E_2, \ldots, E_l\}$ are sets of users, objects and ratings, respectively. The degree of user i is denoted as k_i, and the degree of object α is denoted as k_α. In this paper, Greek and Latin letters are respectively used for object-related and user-related indices. Further, the bipartite network can be naturally represented by a rating matrix A for discrete rating systems. Generally, star ratings and other non-numerical ratings are mapped to numerical ratings in order to facilitate the calculation although this mapping is complex and may bring positive bias [45]. The entry of the rating matrix $A_{i\alpha} \in \Omega = \{\omega_1, \omega_2, \ldots, \omega_z\}$ is the weight of the link connecting user i and object α, with $A_{i\alpha}$ being equal to the corresponding rating value from user i to object α. Based on the analysis of the bipartite network and the equivalent rating matrix, a reputation evaluation method will assign user i with a reputation value R_i, which measures his/her credibility.

2.2. Quality-based ranking method

All the quality-based ranking methods have an underlying assumption that each object α is associated with a most objective rating that best reflects its true quality Q_α. As it is really hard to tell the true quality of an object, the estimated object quality is usually used as an alternative. In general, the estimated object quality $\hat{Q}_\alpha$ of object α is defined as the object's weighted average rating. Mathematically, it reads

$$\hat{Q}_\alpha = \frac{\sum_{i \in U_\alpha} R_i A_{i\alpha}}{\sum_{i \in U_\alpha} R_i}, \tag{1}$$

where $A_{i\alpha}$ is the rating to object α from user i with reputation R_i, and U_α is the set of users who have rated object α.

The iterative refinement (IR) method [35] calculates the user reputation and the object quality in an iterative way. Specifically, a user's reputation is inversely proportional to the difference between the rating vector and the corresponding objects' estimated quality vector. Mathematically, the difference f_i for user i is defined as

$$f_i = \frac{1}{k_i} \sum_{\alpha \in O_i} (A_{i\alpha} - \hat{Q}_\alpha)^2, \tag{2}$$

where O_i is the set of objects that rated by user i, and $\hat{Q}_\alpha$ is the estimated quality value of object α being rated by user i. Initially, all users are assigned with the same reputation, e.g., $R_i = 1$. Then, the reputation of user i is iteratively updated according to

$$R_i = (f_i + \varepsilon)^{-\beta}, \tag{3}$$

where β is a tunable parameter, whose optimal value is $\beta = 1$ [38]. The iteration goes according to Eqs. (1)–(3) until both $\hat{Q}_\alpha$ and R_i converge.

The correlation-based ranking (CR) method [37] and the reputation redistribution ranking (RR) method [38] are based on the same framework. Therefore, in the following only RR is introduced as a representative. In RR, each user i is initially with reputation $R_i = k_i/n$, which can be essentially seen as the user's activity. The estimated quality of object i is $\hat{Q}_i$, which can be calculated by Eq. (1). To obtain the reputation R_i for user i in a step, a so-called temporal reputation TR_i is calculated, which is the Pearson correlation coefficient between the rating vector A_i and the estimated objects' quality vector $\hat{Q}_i$. Mathematically, the temporal reputation TR_i is defined as

$$TR_i = \frac{1}{k_i} \sum_{\alpha \in O_i} \left(\frac{A_{i\alpha} - \mu(A_i)}{\sigma(A_i)} \right) \left(\frac{\hat{Q}_\alpha - \mu(\hat{Q}_i)}{\sigma(\hat{Q}_i)} \right), \tag{4}$$

where μ and σ are functions of mean value and standard deviation, respectively. If TR_i is smaller than 0, TR_i is reset as 0, leading TR_i being in the range [0, 1]. Then, the reputation R_i is recalculated by nonlinearly redistributing TR_i via

$$R_i = TR_i^\theta \frac{\sum_j TR_j}{\sum_j TR_j^\theta}, \tag{5}$$

where θ is a tunable parameter. In each step, both $\hat{Q}_\alpha$ and R_i are updated by Eqs. (1), (4) and (5) until the change of the estimated quality $|\hat{Q} - \hat{Q}'| = \sum_\alpha (\hat{Q}_\alpha - \hat{Q}'_\alpha)^2/n$ is smaller than a threshold value, e.g. $\Delta = 10^{-4}$. Here, $\hat{Q}'$ denotes the

estimated object quality vector in the previous step, and the parameter θ is set as its optimal value $\theta = 3$ [38]. Note that the RR method degenerates to the CR method when $\theta = 1$. In another word, CR is a special case of RR. In fact, there is an improved version of RR called IARR2 by introducing two penalty factors to enhance the performance [38]. However, considering IARR2 can degenerate to RR and CR and all these methods are based on the same framework, we mainly study the CR and RR methods as two examples.

2.3. Group-based ranking method

The group-based ranking (GR) method [41] quantifies the user reputation by calculating the corresponding group sizes after grouping users based on their rating similarities instead of following the debatable assumption as the previous quality-based ranking methods do. Specifically, the GR method works as follows: Firstly, all users are grouped according to their ratings. Users who gave the rating ω_s to object α are put into the group $\Gamma_{s\alpha}$:

$$\Gamma_{s\alpha} = \{U_i \mid A_{i\alpha} = \omega_s, i = 1, 2, \ldots, m\}. \tag{6}$$

Secondly, the size of each group is calculated as $\Lambda_{s\alpha} = |\Gamma_{s\alpha}|$, namely, the number of users who gave the rating ω_s to object α. Thirdly, by normalizing Λ per column, a rating–rewarding matrix Λ^* is defined as

$$\Lambda^*_{s\alpha} = \frac{\Lambda_{s\alpha}}{k_\alpha}, \tag{7}$$

where k_α is the degree of object α. Fourthly, the original rating matrix A is mapped to a rewarding matrix A' referring to Λ^*. The rewarding that a user obtains from the rating $A_{i\alpha}$ is defined as

$$A'_{i\alpha} = \Lambda^*_{s\alpha}, \tag{8}$$

with the constrain that $A_{i\alpha} = \omega_s$. Note that, if user i has not yet rated object α, the value of $A'_{i\alpha}$ is null and it will be ignored in the following calculation.

Then, based on the consideration that small mean and large variance of the rewarding correspond to a user's untrustworthy rating behavior due to its deviation from the majority, the reputation R_i of user i is defined as the inverse of the coefficient of variation [46] of the rewarding vector A'_i. Mathematically, it reads

$$R_i = \frac{\left(\sum_{\alpha \in O_i} A'_{i\alpha}\right)^2}{\sum_{\alpha \in O_i} \left(k_i^2 A'_{i\alpha} - k_i \sum_{\alpha \in O_i} A'_{i\alpha}\right)^2}, \tag{9}$$

where k_i is the degree of object α being rated by user i. Finally, all users are sorted in ascending order by their reputations, and the top-L users with the lowest reputations will be detected as the spammers.

2.4. Iterative refinement procedures

The iterative refinement is an iterative method, which is original proposed to improve the accuracy of numerical solutions to linear systems [47]. From the network perspective, many iterative methods have been proposed in the following years, such as the HITS algorithm with iterative refinement procedure [42] for the hyperlinked pages and the original resource-allocation process [43,44] for the bipartite networks.

The HITS algorithm with iterative refinement procedure [42] is a link-based model, which was first used to identify the hubs pages that link to many related authorities in the context of the WWW. As there is a mutually reinforcing relationship between the Hubs and the authorities, the iterative algorithm which can maintain and update numerical weights for each page was introduced to break this circularity. Specifically, the relationship between the hubs and the authorities is described as follows. First, the weights of the hub i and the authority α are respectively initialized as $x^{(i)}$ and $y^{(\alpha)}$ after their normalization. Then, the weight $x^{(i)}$ of the hub i is updated by

$$x^{(i)} \leftarrow \sum_{\alpha \in Y^{(i)}} y^{(\alpha)}, \tag{10}$$

where $Y^{(i)}$ denotes the set of the authorities that pointed by the hub i. Conversely, the weight $y^{(\alpha)}$ of the authority α is updated by

$$y^{(\alpha)} \leftarrow \sum_{i \in X^{(\alpha)}} x^{(i)}, \tag{11}$$

where $X^{(\alpha)}$ denotes the set of the hubs that pointed by the authority α. Both of the two weights $x^{(i)}$ and $y^{(\alpha)}$ are iterative updated by Eqs. (10) and (11) until a fixed point is reached. The HITS algorithm is a representative iterative refinement procedure and it has found widely applications, such as the Google search engine.

The original resource-allocation process [43,44] is another well-known iterative refinement method, which is equivalent to the one-step random walk in networks starting from the common neighbors. In a "user–object" bipartite network, the resource-allocation process works as follows: First, the resource of object α is initialized as f_α. Then, all the resources are allocated via the following process

$$f'_\alpha = Wf_\alpha, \tag{12}$$

where f'_α is the final resource that located on object α, and W is the transformation matrix. Specifically, the entry $\omega_{\alpha\beta}$ of the transformation matrix W is defined as

$$\omega_{\alpha\beta} = \sum_{i=1}^{m} \frac{A_{i\alpha}A_{i\beta}}{k_i}, \tag{13}$$

where k_i is the degree of user i. In fact, $\omega_{\alpha\beta}$ measures the similarity between object α and object β by summing their contribution from all two-step paths [44,48]. The resource-allocation process can be also used in solving many network-related problems, such as recommendation [49], link predication [50], and community detection [51].

3. Iterative group-based ranking method

The iterative group-based ranking (IGR) method is mainly inspired by the original group-based ranking (GR) method [41] and the iterative refinement procedures [42–44] when reallocating users' reputation based on the consideration of their rating values and previous reputations, referring to some quality-based methods, such as the iterative refinement (IR) [35] and the correlation-based ranking (CR) [37]. Although the proposed IGR method is based on the similar framework as the original GR method, it has several distinguishing characteristics, differentiating it from the original one. In this section, we will describe the proposed IGR method in detail.

3.1. User rating group

The web-based online rating system is a fundamental socialized information collection platform, where users can give multiple reasonable ratings to the same object based on their own experience and judgments [15]. Therefore, the underlying assumption of the quality-based methods, i.e., the quality of each object can be represented by a most objective rating, has faced a huge challenge [41]. From another point of view on group level, it is much easier for users to make a decision on the object with aggregate ratings than the one with scattered ratings mainly due to the fact that the choice of the majority is the best reflection of whether an object is worth to buy. Therefore, referring to two mild assumptions in crowdsourcing that the ratings of the object with high true quality tend to be the same and the users of high credibility tend to agree with others in rating objects [33], the sizes of the user rating groups can be used to quantify the credibility of users.

Different users have different online behavioral patterns in rating objects. The proposed IGR method works by grouping users into different rating groups according to their historical ratings. First, the rating vector A_i of user i is mapped to a rating-object matrix $B^{(i)}$, whose entry $B^{(i)}_{s\alpha}$ is defined as

$$B^{(i)}_{s\alpha} = \begin{cases} 1 & \text{if } A_{i\alpha} = \omega_s, \\ - & \text{otherwise,} \end{cases} \tag{14}$$

where $A_{i\alpha}$ is the rating value ω_s given by user i to object α. Here, the symbol "$-$" stands for a non-value, which should be ignored in the calculation (the same below). Then, the user rating groups can be obtained based on the rating-object matrix, namely, users who gave the same rating ω_s to object α belong to the user rating group $\Gamma_{s\alpha}$. Mathematically, the entry of $\Gamma_{s\alpha}$ is defined as

$$\Gamma_{s\alpha} = \{U_i | B^{(i)}_{s\alpha} = 1\} \tag{15}$$

where $B^{(i)}_{s\alpha}$ is the entry of the rating-object matrix of user i. Obviously, the number of different user rating groups that a user i belongs to is determined by the user degree k_i. Besides, all the k_α users who have rated the object α are put into s different user rating groups, where s is the number of different rating values.

3.2. Weighted group size

The size of the user rating group is a direct reflection of user reputation. Therefore, the calculation of the corresponding group sizes plays an important role in the user reputation evaluation as the ratings to objects of high true quality tend to be the same, leading high credible users being in the large rating groups. However, in the previous work [41], all ratings are evenly contributed in calculating the sizes of the user rating groups without considering the reputation of users who gave these ratings. That is to say the rating from a spammer has the same contribution as the one from a credible user. However, from view on group level, the ratings from users of higher reputation should play a more important role in dominating the

Table 1
Some basic characteristics of the two real data sets. m is the number of users. n is the number of objects. l is the number of all ratings. $\langle k_U \rangle$ is the average degree of users. $\langle k_O \rangle$ is the average degree of objects. $S = l/mn$ is the sparsity of the corresponding bipartite network.

Data set	m	n	l	$\langle k_U \rangle$	$\langle k_O \rangle$	S
MovieLens	943	1682	100 000	106	60	0.0630
Netflix	3000	2779	197 248	66	71	0.0237

corresponding group sizes while the contribution of the lower reputation users' ratings should be reduced. In another word, the size of the user rating group should be determined not only by the ratings of the users but also by their reputations.

In the proposed IGR method, the weighted size of the user rating group $\Gamma_{s\alpha}$ is calculated by considering both the rating-object matrix $B^{(i)}$ and the user reputation R_i. Specifically, after the initial configuration that each user i has equal reputation, e.g. $R_i = 1$, the weighted size $\Lambda_{s\alpha}$ of the user rating group $\Gamma_{s\alpha}$ is defined as

$$\Lambda_{s\alpha} = \sum_{i=1}^{m} R_i B^{(i)}_{s\alpha}, \tag{16}$$

where $B^{(i)}$ is the rating-object matrix of user i with the reputation R_i, and m is the number of all users.

Considering that the distribution of the object degree is heterogeneous in real online rating systems [52], where only a few objects are rated by a large number of users and most of objects are rated by a few users, the absolute sizes of the user rating groups are not comparable for different objects. Therefore, a rating–rewarding matrix Λ^* is established by normalizing the original weighted rating group size matrix Λ by column. Mathematically, it reads $\Lambda^*_{s\alpha} = \Lambda_{s\alpha}/k_\alpha$, where k_α is the degree of object α. Then, the original rating matrix A is mapped to a rewarding matrix A' referring to the rating–rewarding matrix Λ^*. Specifically, the rewarding $A'_{i\alpha}$ that user i obtains from the rating $A_{i\alpha}$ is defined as

$$A'_{i\alpha} = \begin{cases} \Lambda^*_{s\alpha} & \text{if } A_{i\alpha} = \omega_s, \\ - & \text{otherwise,} \end{cases} \tag{17}$$

where $A_{i\alpha}$ is the rating from user i to object α, and ω_s is the corresponding rating value. In fact, the rewarding matrix A' is a more direct reflection of the user's reputation comparing with the original rating matrix.

3.3. Reputation–allocation process

The reputations of all users can be reallocated according to their rewarding vectors. On the one side, if the mean value of a user's rewarding is small, most of the ratings must be deviated from the majority, indicating the user's poor reputation since high credibility users tend to agree with others in rating objects. On the other side, if the rewarding that a user obtains from his ratings varies largely, the user is also untrustworthy as the variation of the rewarding indicates the user's unstable rating behaviors.

Based on the two considerations that larger mean value and smaller variation of a user's rewarding correspond to a user's better credibility, in the proposed IGR method, the reputation R_i of user i is re-allocated via

$$R_i = \frac{\mu(A'_i)}{\sigma(A'_i)}, \tag{18}$$

where μ and σ are respectively mean value and standard deviation. Specifically, the mean value of the rewarding vector A'_i is defined as

$$\mu(A'_i) = \sum_{\alpha} \frac{A'_{i\alpha}}{k_i} \tag{19}$$

and the standard deviation of the rewarding vector A'_i is defined as

$$\sigma(A'_i) = \sqrt{\frac{\sum_{\alpha}(A'_{i\alpha} - \mu(A'_i))^2}{k_i}}, \tag{20}$$

where $A'_{i\alpha}$ is the rewarding that user i obtains from the rating $A_{i\alpha}$, and k_i is the degree of user i.

Note that, there is a mutually reinforcing relationship between the user reputation R and the user rating group size Λ. In fact, recent literatures on the iterative refinement procedures (e.g., the HITS algorithm [42] and the resource-allocation process [42–44]) and the user reputation evaluation (e.g., the iterative refinement (IR) [35] and the reputation redistribution (RR) [38]) inspired us to adopt the reputation–allocation procedure to break this circularity by maintaining and updating the reputation for each user.

Specifically, in the proposed IGR method, the user reputation R and the user rating group size Λ are iteratively updated according to Eqs. (16)–(18) until the change of the user reputation is smaller than the threshold value, e.g. $\Delta = 10^{-4}$. Here,

the change of the user reputation is defined as $|R-R'| = \sum_i (R_i - R'_i)^2/m$, where R' denotes the user reputation vector at the previous iteration step, and m is the number of all users. Finally, all users are sorted by their final reputations in ascending order, and the top-L users with the lowest reputations will be detected as spammers. Note that, when there is no iteration, the proposed IGR method degenerates to the original GR method.

4. Data set and metric

4.1. Real rating data

We consider two commonly used data sets in real online rating systems, namely, MovieLens and Netflix. Both of the two data sets contain ratings on movies based on a 5-point rating scale with 1 being the worst and 5 being the best. MovieLens data set is provided by GroupLens project at University of Minnesota (www.grouplens.org). Herein, we only use a small subset, which is sampled and extracted from the original data with the constraint that each user has at least 20 ratings and each movie is rated by at least one of these users. In the subset, 100 000 ratings are given by 943 users to 1682 movies. Netflix is a huge data set released by the DVD rental company Netflix for its Netflix Prize contest (www.netflixprize.com). We extracted a small data set by random choosing 3000 users who have at least 20 ratings and took all 2779 movies that rated by at least one of these users. Finally, there are 197 248 ratings in the Netflix data set. Compared with Netflix, MovieLens has larger average degree of users, smaller average degree of objects, and higher sparsity of the bipartite network. The basic statistics of the data sets are summarized in Table 1.

4.2. Artificial rating data

To test the performance of different ranking methods, one way is to calculate the ranks of all users and compare them with the ground truth. However, in practice, we are unable to know the ground true ranks of users in advance. As an alternative, we manipulate the real data set by adding artificial spammers and test to what extent these spammers can be detected by a ranking method. In fact, two types of distorted ratings, namely, malicious ratings and random ratings, are widely found in real online rating systems [53,54]. The malicious ratings are from spammers who always gives minimum (maximum) allowable ratings to push down (up) certain target objects while the random ratings mainly come from the test engineers or some naughty users who give meaningless ratings randomly.

As real spammers are unknown, to generate artificial rating data sets for testing, we add either type of artificial spammers (i.e. malicious or random) at one time into the original data. In the implementation, we randomly select d users and turn them into spammers by replacing their original ratings with distorted ones: (i) integer 1 or 5 with the same probability for malicious spammers, and (ii) random integers in the set $\{1, 2, 3, 4, 5\}$ for random spammers. If d is no more than a spammers original degree k, we randomly select his/her d ratings and replace them with distorted ratings with the unselected ratings being ignored. If d is larger than a spammers original degree k, we first select the rest number of uncollected objects, then replace all his/her ratings and unrated ones with distorted ratings. In this way, the ratio of artificial spammers can be calculated by $p = d/m$, where m is the number of all users.

4.3. Evaluation metric

Two widely used metrics are applied to evaluate the performance of the ranking, namely, recall [55] and AUC (the area under the ROC curve) [56]. The recall only focuses on the top-L ranks and its value measures to what extent the spammers can be ranked at the top. Mathematically, the value of recall is defined as

$$R_c(L) = \frac{d'(L)}{d}, \tag{21}$$

where $d'(L) \leq d$ is the number of detected artificial spammers in the top-L ranking list. In the following experiments, the length of the ranking list is set as $L = d$, at which setting the recall is equivalent to another accuracy metric named precision [55]. Larger value of recall R_c indicates higher accuracy of the ranking.

Next, we introduce the L-independent metric AUC. Given the ranks of all users, the value of AUC can be essentially seen as the probability that the reputation of a randomly chosen spammer is lower than the reputation of a randomly chosen normal user (non-spammer) [1]. To calculate the value of AUC, at each time a pair of spammer and normal user are picked and their reputations are compared. If among N independent comparisons, there are N' times the spammer has a lower reputation and N'' times they have the same reputation, the value of AUC is defined as

$$AUC = \frac{N' + 0.5N''}{N}. \tag{22}$$

The value of AUC should be about 0.5 if all normal users and spammers are ranked randomly. Therefore, the more the value of AUC exceeds 0.5, the better the ranking method performs.

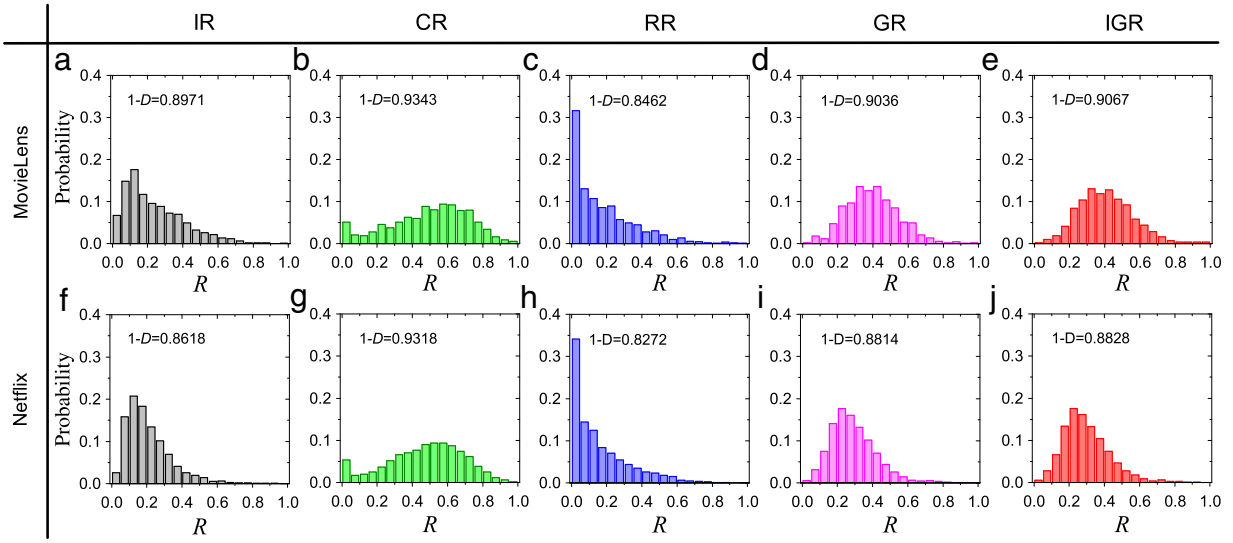

Fig. 1. The probability distribution of all users' reputations after applying different reputation evaluation methods on the two real rating data sets, MovieLens and Netflix. Subfigures (a–e) are for MovieLens; subfigures (f)–(j) are for Netflix. R is the reputation of users. $1-D$ is the Simpson's index of diversity.

4.4. Self-consistency metric

For the reputation evaluation methods, there is an intuition that a user of higher rating error should have a lower reputation or vice versa. That is to say, for a well-performed reputation evaluation method, the reputation should be negatively correlated with the rating error. Here, the rating error δ_i of user i refers to the degree of deviation after comparing the user rating vector A_i with the estimated object quality vector $\hat{Q}$. Mathematically, it reads

$$\delta_i = \frac{\sum_{\alpha \in O_i} |A_{i\alpha} - \hat{Q}_\alpha|}{k_i}, \tag{23}$$

where $\hat{Q}_\alpha = \sum_{i \in U_\alpha} A_{i\alpha}/k_\alpha$ is the average rating that object α is rated, and O_i is the set of objects being rated by user i.

In fact, the correlation between the user rating error δ_i and the user reputation R_i measures the self-consistent of a ranking method. That is because the calculation of the user rating error δ_i depends on the estimated object's quality value $\hat{Q}$ and further the user reputation R_i is considered in calculating $\hat{Q}$. Therefore, the higher value of the correlation between the user rating error δ_i and the user reputation R_i indicates the more self-consistent of a ranking method.

5. Experiments and results

5.1. Reputation evaluation

First, we consider the probability distribution of all users' reputations after applying different reputation evaluation methods on the two real rating data sets. Results are shown in Fig. 1. It can be seen that in IR the reputation is Poisson-like distributed whereas in CR, GR and IGR the reputation is normal-like distributed. By contrast, in RR the reputation is exponential-like distributed with the reputation of most users being zero (see Fig. 1(c) and (h)). To quantify the diversity of all users' reputation based on its probability distribution, we calculate the Simpson's index of diversity, which is denoted as $1-D$ [57]. Higher values of $1-D$ suggest more distinguishable of the obtained reputation. In CR, the values of $1-D$ are the highest as 0.9343 and 0.9318 respectively for MovieLens and Netflix. In GR and IGR, the values of $1-D$ are nearly the same as around 0.90 and 0.88 respectively for MovieLens and Netflix. In RR, the values of $1-D$ are the lowest, suggesting that reputations of users in RR are the least distinguishable. Actually, the reputation that a well-performed evaluation method assigns should be distinguishable. Therefore, the CR, GR and IGR methods meet the criteria better.

Then, we show the relation between the rating error δ and the user reputation R, i.e. the self-consistency index, for different methods in Fig. 2(a) and (d). Note that GR and IGR both assign a high reputation to users of small rating errors and a stably low reputation to users of high rating errors. By contrast, the other three quality-based ranking methods, i.e., IR, CR and RR, are not stable in dealing with users of high rating errors, as indicated by the high variation of R when δ is large. To quantify the relation, we additionally calculate the Pearson correlation coefficient ρ between R and δ. Results are shown in the first row of Table 2. The values of ρ are -0.8166 and -0.8201 (-0.7353 and -0.7629) respectively for GR and IGR

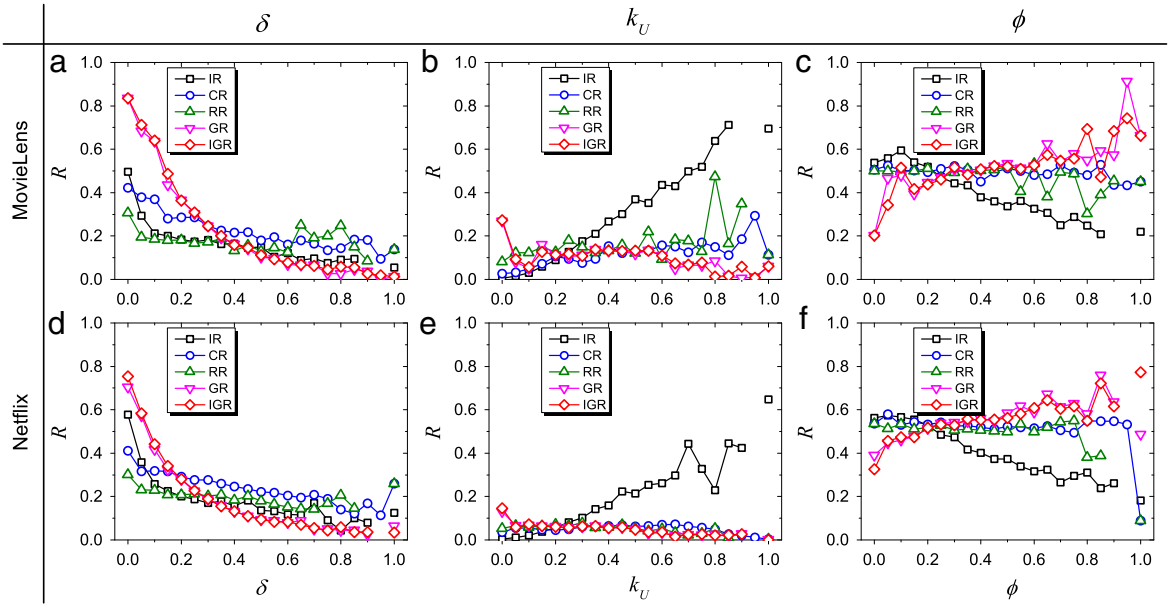

Fig. 2. The relation between R and δ, k_U and ϕ, respectively. Subfigures (a–c) are for MovieLens; subfigures (d–f) are for Netflix. δ is the vector of users' rating errors, k_U is the degree of users, and ϕ is the degree of trend following. For comparison, δ, k_U and ϕ are respectively normalized. As the three normalized indicators are continuous, we respectively divide them into bins and evaluate the mean reputation of users in the same bins.

Table 2
Pearson correlation coefficient ρ between the user reputation R and the user rating error δ, the user degree k_U and the degree of trend following ϕ, respectively. The highest values of the correlation in each row are emphasized in bold.

Metrics	MovieLens					Netflix				
	IR	CR	RR	GR	IGR	IR	CR	RR	GR	IGR
$\rho(\delta, R)$	−0.4471	−0.4537	−0.3189	−0.8166	**−0.8201**	−0.4640	−0.3926	−0.2812	−0.7353	**−0.7629**
$\rho(k_U, R)$	**0.8759**	0.2318	0.1719	−0.0519	−0.0419	**0.7868**	0.0538	0.0040	−0.0950	−0.0904
$\rho(\phi, R)$	**−0.4746**	−0.0244	−0.0287	0.2141	0.2048	**−0.3793**	−0.0428	−0.0569	0.2368	0.2157

in MovieLens (Netflix), showing that IGR performs better than GR. The highest negative correlations also suggest the best self-consistent of both the GR and IGR methods.

5.2. Effect of degree

First, we consider the effect of user degree k_U on determining the corresponding user reputation R under different ranking methods. Fig. 2(b) and (e) show the relations between k_U and R. It is worthy to notice that R in IR is positively correlated with k_U as the correlations are respectively 0.8759 and 0.7868 for MovieLens and Netflix (see the second row of Table 2). In fact, the degree k_U can be essentially seen as a user's activity. Therefore, the result indicates that IR prefers users with high activity as it assigns a higher reputation to active users than inactive ones. By contrast, for the other four methods, there is no obvious degree preference as the correlations are all around 0. The main reason for these observations is that the user reputation in IR is inversely proportional to the least mean square of the difference between the rating and the estimated object quality. As the difference is degree-dependent, large-degree users get a high reputation through the iteration. While CR and RR calculate the correlation, and GR and IGR calculate the mean and standard deviation, which are all independent of user degree. In practice, there is another understanding of such positive correlation for IR. The user degree can be roughly seen as a reflection of buyers' experiences as large-degree users receive more information and they are experienced customers. Hence, large-degree users are considered to have better judgment and their reputations should be higher. However, the straightforward user degree along is not enough to deal with the problem as it is much harder to dig out large-degree spammers.

Then, we study how the degree of trend following affects the user reputation evaluation. The so-called degree of trend following measures to what extent a user would like to collect objects of high popularity. Usually, the popularity of an object can be represented by its degree. Therefore, the degree of trend following can be calculated by the average degree of objects that rated by the user. Mathematically, the degree of trend following ϕ of user i is defined as

$$\phi_i = \frac{\sum_{\alpha \in O_i} k_\alpha}{k_i}, \tag{24}$$

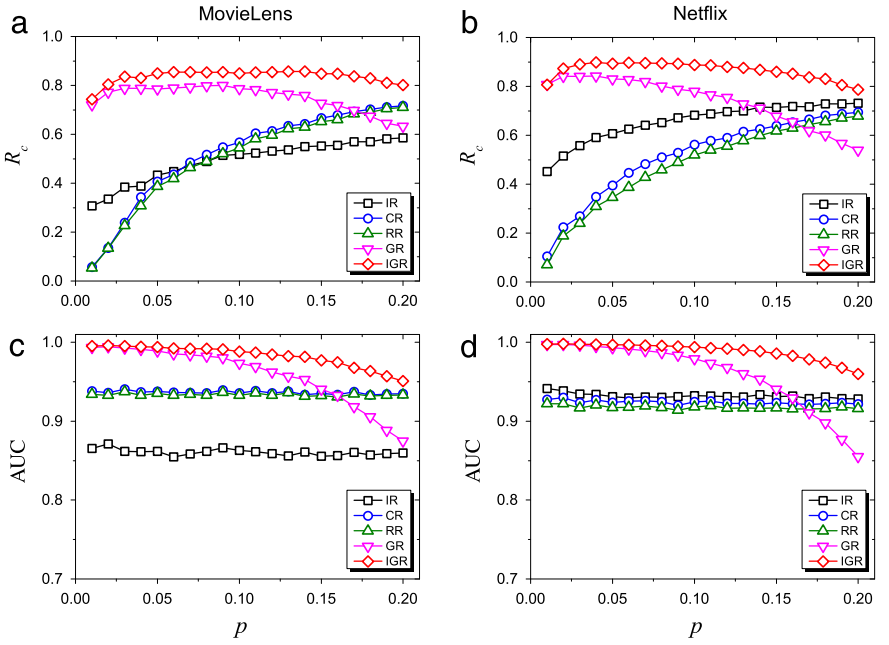

Fig. 3. Performance of different methods on data sets with malicious spamming. Subfigures (a) and (b) are for R_c; subfigures (c) and (d) are for AUC. p is the ratio of malicious spammers. Results are averaged over 100 independent realizations.

where O_i is the set of objects that rated by user i, k_i is the degree of user i, and k_α is the degree of object α. In Fig. 2(c) and (f), we show the relations between the user reputation R and the degree of trend following ϕ. The corresponding values of the correlation coefficients are shown in the third row of Table 2. It can be seen that the values of ρ are -0.4746 and -0.3793 respectively for MovieLens and Netflix, suggesting that the reputation in IR is negatively correlated with the degree of trend following. In GR and IGR, R is weak positively correlated with ϕ as the value of ρ is around 0.2. In CR and RR, the value of ρ is around 0, meaning that R is independent of ϕ.

To better understand these observations, we focus on the mechanisms of these methods. In IR, the ratings from a user of larger ϕ have less chance in dominating the corresponding object estimated quality, resulting in the user's lower reputation. In GR and IGR, a larger ϕ ensures a more stable grouping, which ensures a user's higher reputation. For a more intuitive understanding, we consider the real meaning of the differences in the correlations. Users who always buy objects of high popularity have public taste and the information that they receive is popular to audience. Therefore, it is much harder for them to get higher reputation compared with the users who choose low popular objects as their unique tastes bring them richer information. By contrast, in GR and IGR, users with larger degree of trend following have better grouping behaviors in collecting objects, which bring them higher reputations.

5.3. Malicious spamming analysis

To evaluate the performance of different methods in resisting malicious spamming, we calculate the values of recall R_c and AUC based on the generated artificial data sets with malicious spammers. Results are shown in Fig. 3. When focusing on the top ranks as indicated by the values of R_c in Fig. 3(a) and (b), the proposed IGR method is much more robust than GR especially when the ratio of spammers p is larger, and they both perform the best when p is relatively small. In addition, one can find that CR and RR have similar performance, which increases as p increases. The performance of IR depends on the data sets, and overall it outperforms CR and RR. When focusing on the overall performance of the ranking as indicated by AUC in Fig. 3(c) and (d), the proposed IGR method has the best performance as the values of AUC are over 0.95. Once again, IGR is found to be more robust than GR especially when p is relatively large. Moreover, CR and RR are all very robust as the AUC values are about 0.92, and the performance of IR depends on the data sets. Based on these observations, one can conclude that the proposed IGR method outperforms the original GR method and the other three quality-based methods in resisting malicious spamming. In addition, it is notable that the robustness of IGR is remarkably improved compared with GR although the performance of both the two degreases when p is large. Considering that the ratio of spammers is usually small in real rating systems, the remarkable improvement is meaningful in real applications.

For a more intuitive understanding of how different ranking methods work in resisting malicious spamming, we show the effect of the user degree k_U on evaluating user reputation in parameter spaces (R, k_U) in Fig. 4. Note that IR gives a high reputation to large-degree spammers as it prefers users of large degree. Hence, for users of close degree, IR can accurately distinguish spammers from normal users as shown in Fig. 4(a) and (f). Despite of this, IR gives a relatively higher reputation

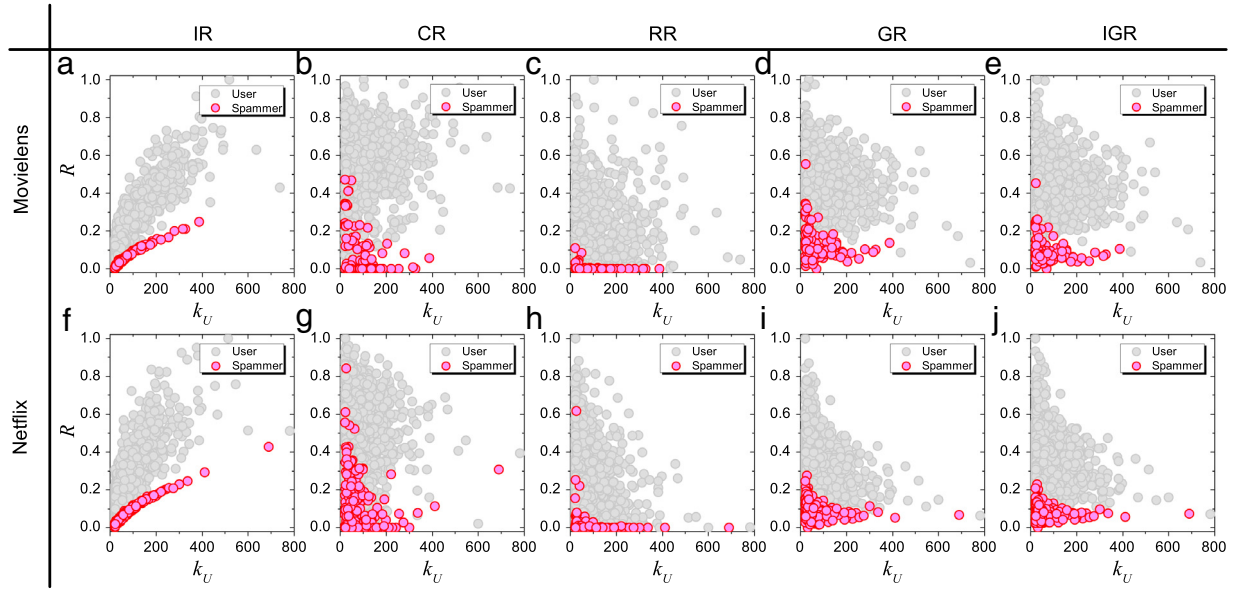

Fig. 4. The relation between the user reputation and the user degree. R is the reputation of users, obtained by applying different methods on data sets with malicious spamming. k_U is the degree of users. The data points colored gray and pink stand for normal users and malicious spammers, respectively. The ratio of spammers is set as $p = 0.1$. The results in each subfigure is for one realization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

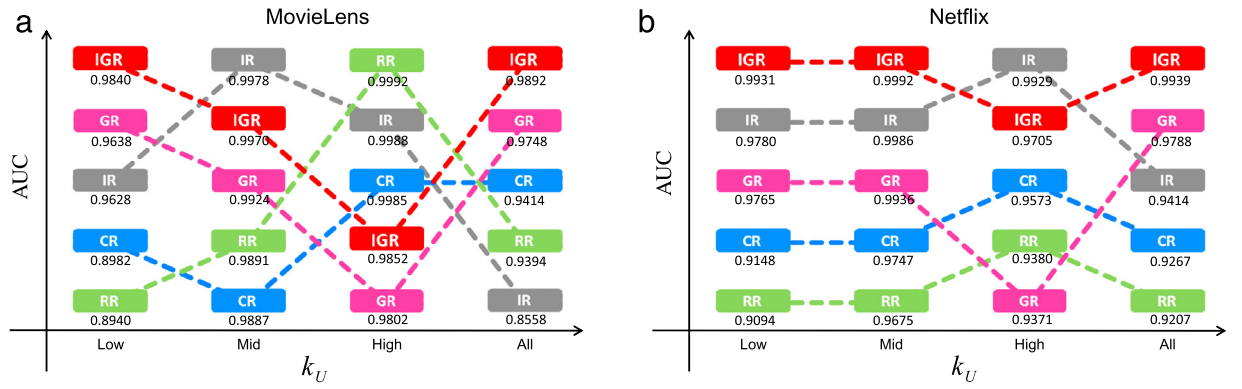

Fig. 5. Comparison of difference methods in ranking malicious spammers with different degree k_U. Subfigures (a) and (b) are for MovieLens and Netflix, respectively. According to k_U, All users are divided into three subgroups, namely, Low, Mid and High. In each subgroup, AUC is calculated after applying different ranking methods. Accordingly, the relative ranks of these methods are obtained. The ratio of spammers is set as $p = 0.1$. Results are averaged over 100 independent realizations.

to large-degree spammers. Meanwhile, CR gives a higher reputation to all users, especially some small-degree spammers, as indicated by most of dots being in the middle and top of Fig. 4(c) and (h). In other words, the mean value of all users' reputation is relatively higher in CR (see Fig. 1(b) and (g)). By contrast, RR over limits all users reputation, as indicated by most dots being in the bottom of Fig. 4(c) and (h), although it gives a lower reputation to most of the spammers. In RR, the reputations of a lot of users are zero (see Fig. 1(c) and (h)), which results in its high false positive rate in spam detection. Although both GR and IGR slightly prefer small-degree users as they give low reputations to large-degree users (see Fig. 4(d) and (i) for GR, and see Fig. 4(e) and (j) for IGR), the reputations in GR and IGR are normal-like distributed and the spammers are always assigned with low reputations. These characteristics ensure both GR and IGR perform the best.

To quantify the effects of the user degree on ranking, we divide all users into three subgroups, namely, Low, Mid and High, according to their degrees. As the evidence of the heavy-tailed (i.e., stretched exponential) distribution of the user degree [52], only a small number of users have large degree. To balance the sizes of each subgroups, the intervals of the user degree k_U for groups Low, Mid and High are respectively set as $[k_{\min}, k_{\min} + 0.1(k_{\max} - k_{\min}))$, $[k_{\min} + 0.1(k_{\max} - k_{\min}), k_{\min} + 0.3(k_{\max} - k_{\min}))$ and $[k_{\min} + 0.3(k_{\max} - k_{\min}), k_{\max}]$, where $k_{\min}$ and $k_{\max}$ are the minimum and maximum values of k_U. In each subgroup, AUC is calculated after applying the five methods. Accordingly, the relative ranks of these methods are obtained as shown in Fig. 5(a) and (b) for MovieLens and Netflix, respectively. IR has a limited performance for Low and Mid-degree spammers. CR and GR have a good performance for High-degree spammers but a poor performance for

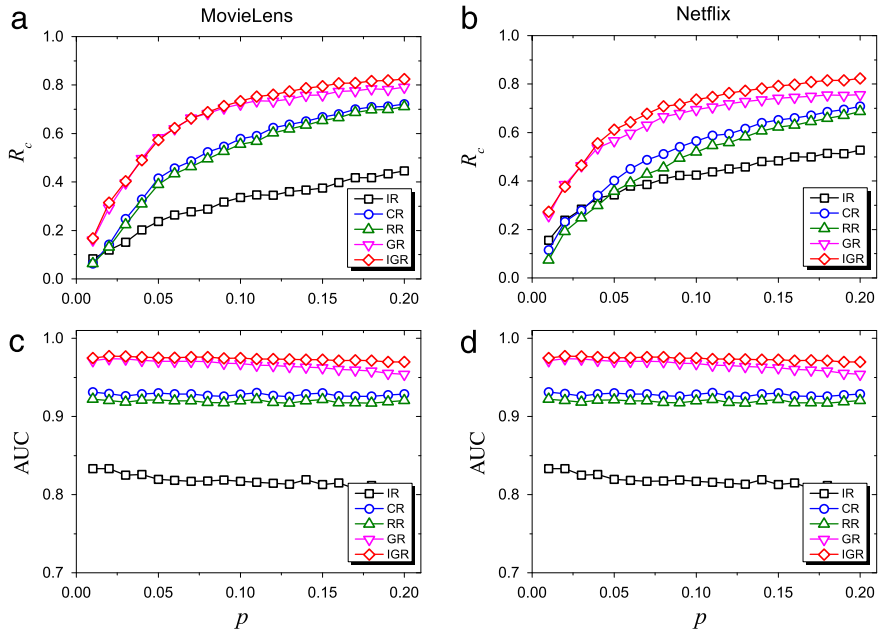

Fig. 6. Performance of different methods on data sets with random spamming. Subfigures (a) and (b) are for R_c; subfigures (c) and (d) are for AUC. The parameter p is the ratio of random spammers. Results are averaged over 100 independent realizations.

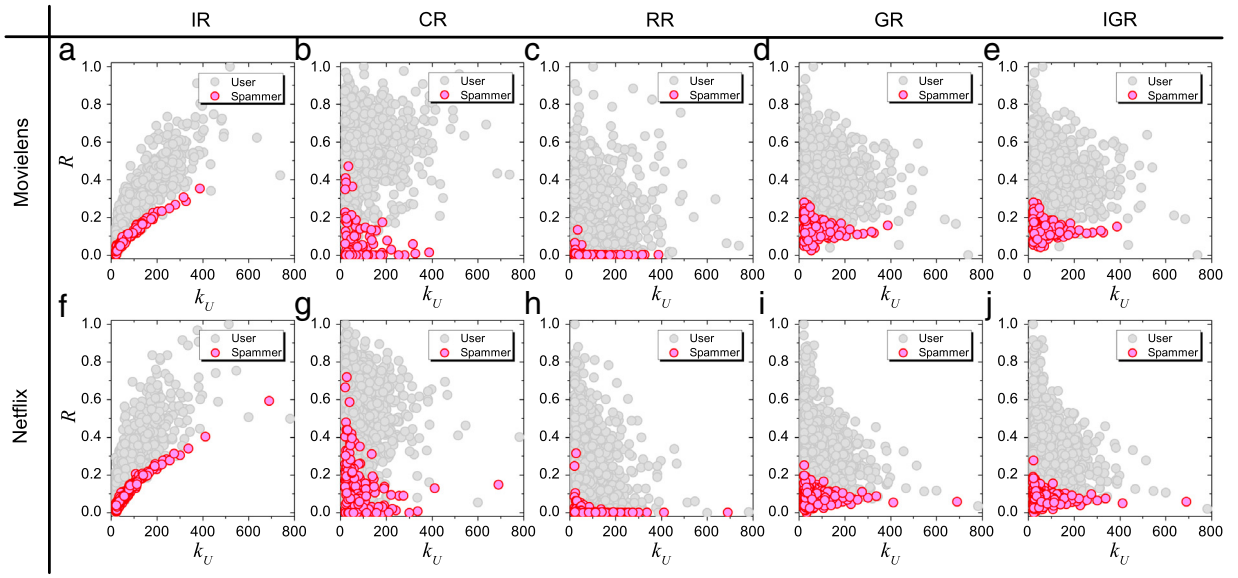

Fig. 7. The relation between the user reputation and the user degree. R is the reputation of users, obtained by applying different methods on data sets with random spamming. k_U is the degree of users. The data points colored gray and pink stand for normal users and random spammers, respectively. The ratio of spammers is set as $p = 0.1$. Results in each subfigures are for one realization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Low-degree spammers. By contrast, GR and IGR outperform the other methods for Low-degree spammers and IGR is better than GR. In ranking All spammers, the order of these methods from the worst to the best is IR, RR, CR, GR and IGR, meaning that IGR is the best in resisting malicious spamming.

5.4. Random spamming analysis

For the evaluation of different methods in resisting random spamming, we first generate artificial data sets with random spammers and then calculate the values of recall R_c and AUC accordingly. Results are shown in Fig. 6. When focusing on the top ranks as indicated by the values of R_c in Fig. 6(a) and (b), the proposed IGR method is much more robust than GR

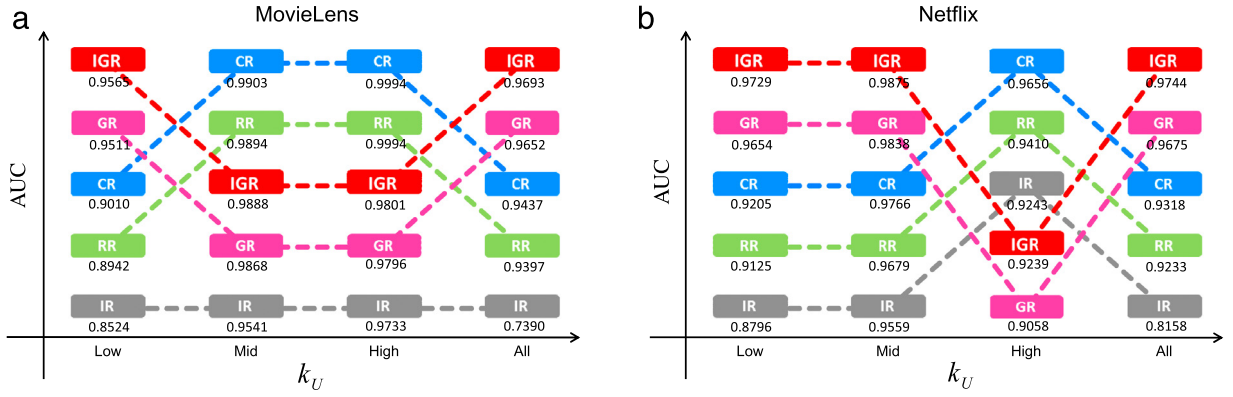

Fig. 8. Comparison of difference methods in ranking random spammers with different degree k_U. Subfigures (a) and (b) are for MovieLens and Netflix, respectively. According to k_U, All users are divided into three subgroups, namely, Low, Mid and High. In each subgroup, AUC is calculated after applying different ranking methods. Accordingly, the relative ranks of these methods are obtained. The ratio of spammers is set as $p = 0.1$. Results are averaged over 100 independent realizations.

especially when p is larger. Both IGR and GR have better performance than the other methods. CR is on a par with RR, and they both outperform IR. Furthermore, we note that the value of R_c increases as p increases. Specifically, the value of R_c has a rapid growth when p is approaching a value around 0.05. Afterwards, the value of R_c becomes stable. These results suggest that there are some real random spammers in the original data sets, and the ratio is about 0.05. When focusing on the overall performance as indicated by the values of AUC in Fig. 6(c) and (d), GR and IGR remarkably outperform the other methods by giving a robust AUC value around 0.96. CR and RR are slightly inferior as the AUC values are about 0.92. For IR, the AUC value is significant lower, indicating its limited performance. In short, the group-based ranking methods outperform the quality-based ranking methods in resisting random spamming. Further, it should be noted that the proposed IGR method performs better than GR although the advantage is limited when p is small. However, the improvement is notable due to the difficulty in detecting a relatively small number of random spammers in real rating systems.

To better understand how these methods work in resisting random spamming, we show the effect of the user degree on reputation evaluation in parameter spaces (R, k_U) in Fig. 7. It can be seen that IR gives a high reputation to large-degree spammers due to its preference to users of large degree (see Fig. 7(a) and (f)). As indicated by the pink dots randomly occupy the diagram regardless of the user degree (see Fig. 7(b) and (g)), the user degree k_U and the reputation R for CR are not correlated, suggesting that CR has no obvious degree preference as it gives high R to some users regardless of their degrees. RR over limits all users' reputations by giving a almost zero reputation to lots of users (see Fig. 7(c) and (h)), which increases the false positive rate in spamming detection. In GR and IGR, the reputation is normal-like distributed and the spammers are always assigned with a low reputation (see Fig. 7(d) and (i) for GR, and see Fig. 7(e) and (j) for IGR). These characteristics ensure both GR and IGR perform better than the other benchmark methods.

In order to quantify the effects of k_U on ranking users' reputations, we show the relative ranks of different methods by AUC after dividing all users into three subgroups according to the user degree in Fig. 8(a) and (b) for MovieLens and Netflix, respectively. It can be seen that IR has a limited performance for Low and Mid-degree spammers. CR and GR have a good performance for High degree-spammers but a poor performance for Low-degree spammers. By contrast, both IGR and GR outperform the other methods for Low-degree spammers, and IGR has better performance than GR. In ranking All spammers, the order of these methods from the worst to the best is IR, RR, CR, GR and IGR, showing that the proposed IGR method is once again the best one in resisting random spamming.

6. Conclusions and discussion

In summary, we have proposed an iterative group-based ranking method in user reputation evaluation by introducing an iterative reputation–allocation process into the original group-based ranking method. Specifically, when calculating the weighted user rating group sizes, the ratings are assigned with larger weights if they come from users with higher reputation, otherwise the ratings are assigned with smaller weights. In the iterations, the user reputation and the corresponding group sizes are iteratively calculated until they become stable. Extensive experiments on two real data sets with artificial spamming suggest that the proposed iterative method remarkably outperforms the previous quality-based ranking methods in evaluating user reputation, and its robustness is largely improved compared with the original group-based ranking method. Further, we provided some insights on the mechanisms and analyzed the characteristics of these methods. This work emphasizes the positive role of considering users' grouping behaviors in designing better reputation evaluation methods.

In fact, the proposed method is based on the same framework as the original group-based ranking method and is inspired by the HITS algorithm with iterative refinement procedure [42] and the original resource-allocation process [43,44]. As an improvement, the proposed method is more robustness in resisting a large number of spamming attacks compared to the

original one. That is mainly because the ratings from users with poor reputation have less chance in forming big groups and the reputation is reallocated in an iterative manner. Even though the number of spammers increases, the reputation of spammers decays through the iterations and the effects of the spam ratings on the whole system are well restricted. Together, it is noted that the performance of both the proposed method and the original group-based method decreases when the ratio of malicious spammers is dramatically increased. The main reason may be that plenty of malicious ratings have dominated the user rating group and the whole rating systems are completely destroyed. Considering that the ratio of spammers is usually small in real systems, the improvement of the proposed method is still helpful in real applications.

From the macro analysis, the proposed iterative group-based ranking method as well the original one are distinguishable from the quality-based ranking methods as the former two assign reputations to users by considering their grouping behaviors while the latter ones are based on the estimation of objects' true qualities. Moreover, both the stability of assigning low reputation to users with high rating errors and the independence of the reputation from the user degree ensure the effective of the iterative group-based ranking method in resisting spamming attacks. Our work provides a further understanding on the mechanism of some user reputation evaluation methods and prompts us to pay more attention to the grouping behaviors of users in improving the algorithmic performance. Indeed, based on the estimated user reputation, our method can also be applied in uncovering the quality of online products as the quality-based methods do through the reputation weighted average rating. In addition, the proposed method has advantages in real applications as it is not only accurate in ranking, but also easier to be implemented.

Traditionally, a well-performed method should be convergent to a unique reputation vector, however, most of the previous reputation-based ranking methods cannot guarantee the convergence [39]. Although extensive simulations suggest that the proposed method can converge, we still expect further theoretical analysis to justify it. Moreover, the previous studies either assume the continuums of the rating values such as the correlation-based ranking method or depend on the assumption of the discrete rating systems such as the group-based ranking method [41]. Actually, how the continuous vs. discrete-valued ratings affect the user reputation evaluation is still an open issue [58]. Besides, the effect of numerical ratings on building the ranks of users is also worth of further investigation [45]. As future works, we could consider applying the proposed method to the rating systems with higher-resolution scales [59] and designing more reputation evaluation methods that can make best use of users' grouping behaviors [60].

Acknowledgments

The authors acknowledge Shuhong Chen for useful suggestions. This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11222543 and 61433014). T.Z. acknowledges the Special Project of Sichuan Youth Science and Technology Innovation Research Team (Grant No. 2013TD0006) and the Program for New Century Excellent Talents in University (Grant No. NCET-11-0070).

References

[1] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.
[2] P. Resnick, K. Kuwabara, R. Zeckhauser, E. Friedman, Reputation systems, Commun. ACM 43 (12) (2000) 45–48.
[3] S.-R. Yan, X.-L. Zheng, Y. Wang, W.W. Song, W.-Y. Zhang, A graph-based comprehensive reputation model: Exploiting the social context of opinions to enhance trust in social commerce, Inform. Sci. 318 (2015) 51–72.
[4] J. O'Donovan, B. Smyth, Trust in recommender systems, in: Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI'05, ACM, New York, NY, USA, 2005, pp. 167–174.
[5] X.-L. Zheng, C.-C. Chen, J.-L. Hung, W. He, F.-X. Hong, Z. Lin, A hybrid trust-based recommender system for online communities of practice, IEEE Trans. Learn. Technol. 8 (4) (2015) 345–356.
[6] J. Weng, C. Miao, A. Goh, Improving collaborative filtering with trust-based metrics, in: Proceedings of the 2006 ACM Symposium on Applied Computing, SAC'06, ACM, New York, NY, USA, 2006, pp. 1860–1864.
[7] J. Gao, Y. Hu, T. Zhou, Bootstrap percolation on spatial networks, Sci. Rep. 5 (2015) 14662.
[8] Q. Tang, B. Gu, A.B. Whinston, Content contribution for revenue sharing and reputation in social media: A dynamic structural model, J. Manage. Inf. Syst. 29 (2) (2012) 41–76.
[9] A. Jøsang, R. Ismail, C. Boyd, A survey of trust and reputation systems for online service provision, Decis. Support Syst. 43 (2) (2007) 618–644.
[10] C.-H. Lai, D.-R. Liu, C.-S. Lin, Novel personal and group-based trust models in collaborative filtering for document recommendation, Inform. Sci. 239 (2013) 31–49.
[11] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, Recommender systems survey, Knowl.-Based Syst. 46 (2013) 109–132.
[12] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Comput. 7 (1) (2003) 76–80.
[13] G. Bente, O. Baptist, H. Leuschner, To buy or not to buy: Influence of seller photos and reputation on buyer trust and purchase behavior, Int. J. Hum.-Comput. Stud. 70 (1) (2012) 1–13.
[14] D.-H. Park, J. Lee, I. Han, The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement, Int. J. Electron. Commer. 11 (4) (2007) 125–148.
[15] L. Muchnik, S. Aral, S.J. Taylor, Social influence bias: A randomized experiment, Science 341 (6146) (2013) 647–651.
[16] S. Xie, G. Wang, S. Lin, P.S. Yu, Review spam detection via temporal pattern discovery, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12, ACM, New York, NY, USA, 2012, pp. 823–831.
[17] P.-A. Chirita, W. Nejdl, C. Zamfir, Preventing shilling attacks in online recommender systems, in: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management, WIDM'05, ACM, New York, NY, USA, 2005, pp. 67–74.
[18] A. Mukherjee, B. Liu, N. Glance, Spotting fake reviewer groups in consumer reviews, in: Proceedings of the 21st International Conference on World Wide Web, WWW'12, ACM, New York, NY, USA, 2012, pp. 191–200.
[19] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, M. Gonçalves, Detecting spammers and content promoters in online video social networks, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'09, ACM, New York, NY, USA, 2009, pp. 620–627.

[20] D. Fraga, Z. Bankovic, J. Moya, A taxonomy of trust and reputation system attacks, in: 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, (TrustCom), IEEE, 2012, pp. 41–50.
[21] S.S. Standifird, Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings, J. Manag. 27 (3) (2001) 279–295.
[22] Y. Sun, Y. Liu, Security of online reputation systems: The evolution of attacks and defenses, IEEE Signal Process. Mag. 29 (2) (2012) 87–97.
[23] R.Y. Toledo, Y.C. Mota, L. Martínez, Correcting noisy ratings in collaborative recommender systems, Knowl.-Based Syst. 76 (2015) 96–108.
[24] Z. Yang, Z.-K. Zhang, T. Zhou, Anchoring bias in online voting, Europhys. Lett. 100 (6) (2012) 68002.
[25] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, H.W. Lauw, Detecting product review spammers using rating behaviors, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10, ACM, New York, NY, USA, 2010, pp. 939–948.
[26] C.-J. Zhang, A. Zeng, Behavior patterns of online users and the effect on information filtering, Physica A 391 (4) (2012) 1822–1830.
[27] G. Ling, I. King, M.R. Lyu, A unified framework for reputation estimation in online rating systems, in: Proceedings of the 32rd International Joint Conference on Artificial Intelligence, IJCAI'13, AAAI, Beijing, China, 2013, pp. 2670–2676.
[28] M.K. Chang, W. Cheung, M. Tang, Building trust online: Interactions among trust building mechanisms, Inf. Manag. 50 (7) (2013) 439–445.
[29] W. Jiang, J. Wu, F. Li, G. Wang, H. Zheng, Trust evaluation in online social networks using generalized flow, IEEE Trans. Comput. 65 (3) (2016) 952–963.
[30] C. Martinez-Cruz, C. Porcel, J. Bernabé-Moreno, E. Herrera-Viedma, A model to represent users trust in recommender systems using ontologies and fuzzy linguistic modeling, Inform. Sci. 311 (2015) 102–118.
[31] K. Fujimura, T. Nishihara, Reputation rating system based on past behavior of evaluators, in: Proceedings of the 4th ACM Conference on Electronic Commerce, EC'03, ACM, New York, NY, USA, 2003, pp. 246–247.
[32] X.-L. Liu, Q. Guo, L. Hou, C. Cheng, J.-G. Liu, Ranking online quality and reputation via the user activity, Physica A 436 (2015) 629–636.
[33] Y. Tian, J. Zhu, Learning from crowds in the presence of schools of thought, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12, ACM, New York, NY, USA, 2012, pp. 226–234.
[34] H. Liao, A. Zeng, Y.-C. Zhang, Towards an objective ranking in online reputation systems: the effect of the rating projection, arXiv:1411.4972.
[35] P. Laureti, L. Moret, Y.-C. Zhang, Y.-K. Yu, Information filtering via iterative refinement, Europhys. Lett. 75 (6) (2006) 1006.
[36] C. de Kerchove, P. Van Dooren, Iterative filtering for a dynamical reputation system, arXiv:0711.3964.
[37] Y.-B. Zhou, T. Lei, T. Zhou, A robust ranking algorithm to spamming, Europhys. Lett. 94 (4) (2011) 48002.
[38] H. Liao, A. Zeng, R. Xiao, Z.-M. Ren, D.-B. Chen, Y.-C. Zhang, Ranking reputation and quality in online rating systems, PLoS One 9 (5) (2014) e97146.
[39] R.-H. Li, J.X. Yu, X. Huang, H. Cheng, Robust reputation-based ranking on bipartite rating networks, in: Proceedings of the 2012 SIAM International Conference on Data Mining, ICDM'2012, SIAM, Anaheim, California, USA, 2012, pp. 612–623.
[40] M. Allahbakhsh, A. Ignjatovic, An iterative method for calculating robust rating scores, IEEE Trans. Parallel Distrib. Syst. 26 (2) (2015) 340–350.
[41] J. Gao, Y.-W. Dong, M.-S. Shang, S.-M. Cai, T. Zhou, Group-based ranking method for online rating systems with spamming attacks, Europhys. Lett. 110 (2) (2015) 28003.
[42] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.
[43] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Phys. Rev. E 75 (2007) 021102.
[44] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Phys. Rev. E 76 (2007) 046115.
[45] R. Centeno, R. Hermoso, M. Fasli, On the inaccuracy of numerical ratings: Dealing with biased opinions in social networks, Inf. Syst. Front. 17 (4) (2015) 809–825.
[46] L.I.-K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1) (1989) 255–268.
[47] J.H. Wilkinson, Rounding Errors in Algebraic Processes, Courier Corporation, 1994.
[48] L.-J. Chen, Z.-K. Zhang, J.-H. Liu, J. Gao, T. Zhou, A vertex similarity index for better personalized recommendation, Physica A 466 (2017) 607–615.
[49] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[50] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[51] Y. Pan, D.-H. Li, J.-G. Liu, J.-Z. Liang, Detecting community structure in complex networks via node similarity, Physica A 389 (14) (2010) 2849–2857.
[52] M.-S. Shang, L. Lü, Y.-C. Zhang, T. Zhou, Empirical analysis of web-based user-object bipartite networks, Europhys. Lett. 90 (4) (2010) 48006.
[53] N. Jindal, B. Liu, Opinion spam and analysis, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM'08, ACM, New York, NY, USA, 2008, pp. 219–230.
[54] F. Ricci, L. Rokach, B. Shapira, Introduction to recommender systems handbook, in: Recommender Systems Handbook, Springer, 2011, pp. 1–35.
[55] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. 22 (1) (2004) 5–53.
[56] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[57] E.H. Simpson, Measurement of diversity, Nature 163 (1949) 688.
[58] M. Medo, J.R. Wakeling, The effect of discrete vs.continuous-valued ratings on reputation and ranking systems, Europhys. Lett. 91 (4) (2010) 48004.
[59] E. Svensson, Comparison of the quality of assessments using continuous and discrete ordinal rating scales, Biom. J. 42 (2000) 417.
[60] X. Shi, J. Zhu, R. Cai, L. Zhang, User grouping behavior in online forums, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'09, ACM, New York, NY, USA, 2009, pp. 777–786.