Do recommender systems benefit users?

Chi Ho Yeung

Department of Science and Environmental Studies,

The Hong Kong Institute of Education,

10 Lo Ping Road, Taipo, Hong Kong

(Dated: July 20, 2015)

Recommender systems are present in many web applications to guide our choices. They increase sales and benefit sellers, but whether they benefit customers by providing relevant products is questionable. Here we introduce a model to examine the benefit of recommender systems for users, and found that recommendations from the system can be equivalent to random draws if one relies too strongly on the system. Nevertheless, with sufficient information about user preferences, recommendations become accurate and an abrupt transition to this accurate regime is observed for some algorithms. On the other hand, we found that a high accuracy evaluated by common accuracy metrics does not necessarily correspond to a high real accuracy nor a benefit for users, which serves as an alarm for operators and researchers of recommender systems. We tested our model with a real dataset and observed similar behaviors. Finally, a recommendation approach with improved accuracy is suggested. These results imply that recommender systems can benefit users, but relying too strongly on the system may render the system ineffective.

Introduction

Almost all popular websites employ recommender systems to match users with items [1–4]. For instance, news websites analyze the reading history of individuals and recommend news which match their interests [5]; online social networks recommend new friends to individuals based on their existing friends [6]. Most commonly, online retailers analyze the purchase history of customers and recommend products to them to increase their own sales [7–9]. These examples show an increasingly crucial role of recommender systems in our daily life, influencing our various choices.

Due to their broad applications, great efforts have been devoted to study recommendation

algorithms and to improve their accuracy [4]. Researchers in computer science, mathematics and management science employ various mathematical tools such as Bayesian approach and matrix factorization to derive recommendation algorithms [4, 10–12]. Recently, physicists and complex system scientists started to work in the area and incorporated physical processes such as mass diffusion and heat conduction to recommender system [13]. Nevertheless, the main goal of these studies is limited to recommendation accuracy, but their genuine benefits are less examined.

Although recommender systems have been shown to benefit retailers, whether the recommended products are relevant to customers is questionable [9, 14]. On one hand, many recommendation algorithms are based on product similarity and the recommended products may be redundant since they are similar to the already purchased products [13]. On the other hand, instead of specific products which match individual needs, many recommender systems can only recommend popular but potentially irrelevant products [9, 15]. Nevertheless, users may be tempted to purchase the products due to recommendations, and in this case recommender systems benefit sellers but not customers.

In this paper, we introduce a simple model to examine the relevance between the recommended products and the preferences of users. Unlike empirical studies where the true user preference is unknown, each user in the model is characterized by a taste and the true recommendation accuracy can be measured. We found that recommendations can be either random or very accurate depending on the frequency the users select a product without recommendations. For some algorithms, an abrupt increase in accuracy is observed when this frequency exceeds a threshold. On the other hand, we found that a high accuracy indicated by common evaluation metrics does not necessarily imply to a high real accuracy. We tested our model using the MovieLens dataset [16] and observed similar behaviors. Finally, a recommendation approach based on our findings was suggested which outperforms conventional approaches.

Model

Specifically, we consider a group of N users selecting products from a group of M items. Each user i and item α is characterized by one of the G *tastes* or *genres*, denoted by g_i and g_α respectively. For instance, in terms of movies, these tastes may correspond to science fictions,

romantic comedies or thrillers. The case where users have multiple tastes are described in Section C.

At each time step, a user i is randomly drawn. With a fraction f_{sel} of the times, user i chooses a product matching his/her own taste without using the recommender system. This is the conventional way to purchase a product and we call f_{sel} the *frequency of deliberate selection*. On the other hand, with a fraction $1 - f_{\text{sel}}$ of the times, user i buys a product following the recommender system. In both cases, a product in his/her collection is randomly removed since all products are assumed to be consumable and can be brought and consumed for more than once. In this case, the total number of products collected by user i remains constant at k_i, which simplifies our model as network growth is not required and N and M remain constant. The above procedures are repeated for a large number of times per user.

We remark that the recommender system has no direct knowledge of user taste and product genre, it can only infer user preferences through his/her purchase history. Since f_{sel} is the frequency a user makes purchases in the absence of recommender systems, on average at least f_{sel} of the purchases of user i must match his/her taste; f_{sel} is thus proportional to the amount of *available hints* the recommender systems can exploit. We further define *recommendation accuracy* A_{rec} to be the fraction of recommended products which match the taste of the user, and our goal is to examine A_{rec} to reveal the benefit of recommender systems to users.

For simplicity, we employ the common *Item-based Collaborative Filtering* (ICF) [17] to be the recommendation algorithm in our model. ICF provides personalized recommendations to users by computing similarity between their purchased products with other products. We first denote the similarity between item α and β at time t to be $s_{\alpha\beta}(t)$. As shown by previous studies [17], the performance of the algorithm is strongly dependent on the definition of similarity. To shown that our results are relevant to different recommendation algorithms, we will employ two definitions of similarity, namely the *common neighbor* (CN) *similarity*, given by

$$s_{\alpha\beta}^{(\text{CN})}(t) = \sum_{i=0}^{N} a_{i\alpha}(t)a_{i\beta}(t), \tag{1}$$

and the *cosine similarity* [17], given by

$$s_{\alpha\beta}^{(\text{cosine})}(t) = \frac{1}{\sqrt{k_\alpha k_\beta}} \sum_{i=0}^{N} a_{i\alpha}(t)a_{i\beta}(t). \tag{2}$$

The adjacency variable $a_{i\alpha}(t) = 1$ if item α is collected by user i at time t, and otherwise $a_{i\alpha}(t) = 0$. The recommendation score $r_{i\alpha}(t)$ of product α for user i at time t is given by

$$r_{i\alpha}(t) = \sum_{\beta=1}^{M} a_{i\beta}(t) s_{\alpha\beta}(t) = \sum_{\beta \in C_i(t)} s_{\alpha\beta}(t), \tag{3}$$

where $C_i(t)$ is the set of products collected by user i at time t. Finally, the product with the highest score not yet collected by the user is recommended.

Results

A. Random versus accurate recommendations

To examine the benefit of recommender system to users, we first study the dependence of recommendation accuracy A_{rec} on the frequency f_{sel} of deliberate selection. The higher the value of f_{sel}, the more often the user chooses a product of a matching taste without recommendation, and the more the information for the recommender system to exploit. If recommender systems work perfectly, $A_{\mathrm{rec}} = 100\% = 1$ whenever $f_{\mathrm{sel}} > 0$ as there exists non-zero information about user tastes in the dataset; on the other hand, if recommender systems do not work at all, recommendations are always random, and $A_{\mathrm{rec}} = 1/G$ independent of f_{sel}.

As shown in Fig. 1, the recommendation accuracy falls between the two extreme cases. The common neighbor similarity is employed in Fig. 1(a), and $A_{\mathrm{rec}} \approx 1/G$ which corresponds to the case of random recommendations when f_{sel} is less than a threshold. When f_{sel} increases beyond the threshold, recommendation accuracy increases abruptly to $A_{\mathrm{rec}} = 1$, which corresponds to a case of perfect recommendation. As shown in Fig. 1(b), cosine similarity is employed and a similar dependence of A_{rec} on f_{sel} is observed, though the transition between the two phases is more gentle. We remark that $A_{\mathrm{rec}} = 1$ is an artifact of the model since each user and product is categorized by only one taste, and after users and products of the same taste formed an isolated bipartite cluster, only products within the cluster are recommended and lead to a persistent perfect accuracy.

The accuracy A_{rec} is also dependent on the number of taste group G. Intuitively, the threshold value for perfect recommendation decreases with G, since it seems easier to identify an item with the correct taste out of a smaller number of taste groups. However,
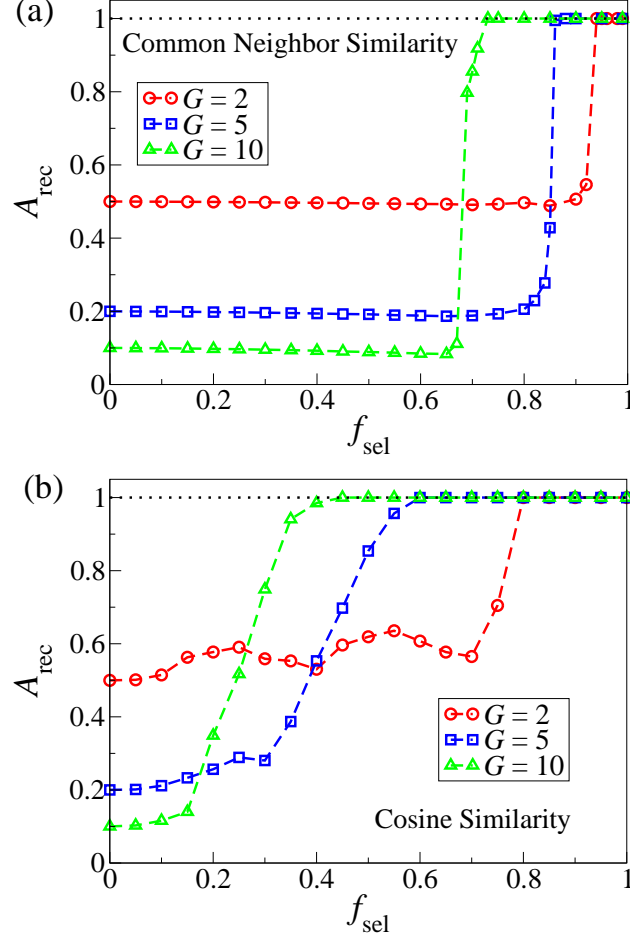
FIG. 1: The accuracy A_{rec} of the recommender system as a function of f_{sel} for different number of taste groups G. The simulation results were obtained with $N = 2000$ users and $M = 100$ products. Each user collects $k = 7$ products and is updated 1×10^5 times. Each data point was averaged over 50 instances. The common neighbor similarity Eq. (1) and the cosine similarity Eq. (2) were employed in (a) and (b) respectively.

simulated results in both Fig. 1(a) and (b) show that the threshold value increases when G decreases. It is because users collect products of both relevant and irrelevant taste; when G is small, the irrelevant products belong to a small number of taste groups, and there exists a strong connection between users and each irrelevant taste group, making it difficult for the recommender system to identify these false connections. In short, the more diverse and distinct the users and products, the less amount of hints are required to provide correct recommendations.

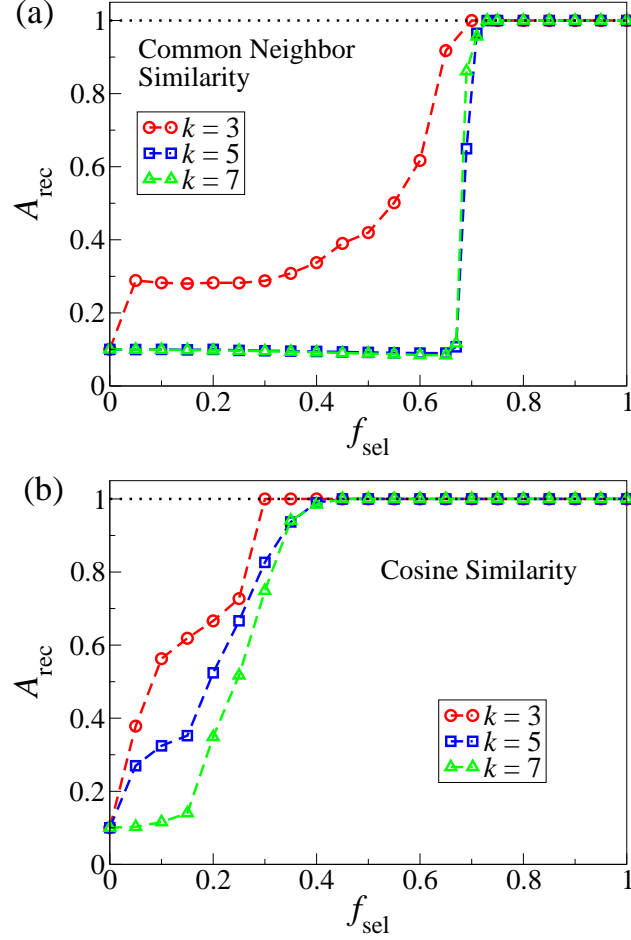Other than the number of taste group, recommendation accuracy also depends on the

FIG. 2: The accuracy A_{rec} of the recommender system as a function of f_{sel} for different values of k, the number of products collected per user. The simulation results were obtained with $N = 2000$, $M = 100$ and $G = 10$. Each user was updated 1×10^5 times, and each data point was averaged over 50 instances. The common neighbor similarity Eq. (1) and the cosine similarity Eq. (2) were employed in (a) and (b) respectively.

number of items collected by each user. For simplicity, all users collect the same number of items, i.e. $k_i = k$ for $\forall i$. As shown in Fig. 2(a) and (b), perfect recommendation is more difficult to be achieved for cases with larger k, where the stronger connection between users and irrelevant taste groups is again the reason. These results imply that when users collect a large number of products, false connections exist and may impact negatively on the recommender system. Hence, instead of drawing recommendations based on all the available data, an algorithm which effectively eliminates the false connections may lead to a high recommendation accuracy.

The above results suggest that recommender systems may provide irrelevant recommendations when users do not provide sufficient hints about their taste. On the other hand, given sufficient hints, recommender systems well utilize the information to match users with products. The amount of hints required for accurate recommendation is different for different algorithms and systems.

B. Estimated accuracy versus real accuracy

In real systems, since the real preference of users is unknown, there is no way to measure the real recommendation accuracy. Various metrics are thus introduced to evaluate recommendation accuracy. Nevertheless, whether these metrics correctly measure real accuracy is questionable. Since user taste and product genre are defined in our model, we can compare the accuracy measured by these metrics with the real accuracy.

One common metric to evaluate recommendation accuracy is AUC, i.e. the area under the *receiver operating curve* (ROC). When recommendations are made for user i, AUC is computed as the probability that a correct product α is ranked higher than an arbitrary product γ, given by

$$AUC_{i\alpha} = \frac{n(r_{i\gamma} < r_{i\alpha}) + 0.5n(r_{i\gamma} = r_{i\alpha})}{M - k_i} \tag{4}$$

where $n(r_{i\gamma} < r_{i\alpha})$ is the number of products with score $r_{i\gamma}$ lower than the score $r_{i\alpha}$ of the correct product, and $n(r_{i\gamma} = r_{i\alpha})$ is the number of items which tie with the correct item. Based on the definition of *correct predictions*, we compute two AUC measures - (i) the conventional estimated AUC_{est}, obtained by dividing the dataset into a training set and a probe set; links in the probe set are removed and are considered to be correct predictions if their existence are predicted; and (ii) the real AUC_{real} which quantifies the accuracy of the algorithm in recommending products of a matching taste.

The dependence of AUC_{est} and AUC_{real} on f_{sel} is shown in Fig. 3. As we can see, $AUC_{\text{real}} \approx 0.5$ when f_{sel} is small since recommendations are random (see Fig. 1) and the products of a matching taste are randomly ranked in the recommendation list. However, AUC_{est} is much higher and is not consistent with AUC_{real}. The reason for a large AUC_{est} at small f_{sel} is the frequent application of recommender systems, such that user purchases are strongly influenced by the algorithms regardless of their true preference. In this case,
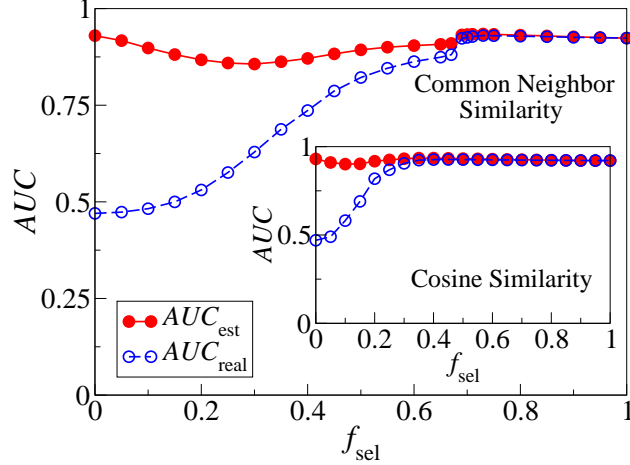
FIG. 3: The two different AUC measures, AUC_{est} and AUC_{real}, as a function of f_{sel}, obtained by ICF with common neighbor similarity and cosine similarity (inset) on systems with $N = 2000$, $M = 100$, $k = 3$ and $G = 10$.

products which do not match their preference but are consistent with the algorithms are also collected by the users. This favors the evaluation by AUC_{est} using a random probe set, and lead to a high AUC_{est} even random recommendations are indeed provided.

When f_{sel} increases, AUC_{est} decreases since the user-product relations become less influenced by the recommender system. At the same time, AUC_{real} increases since more hints about the user tastes are present. We remark that although $A_{\mathrm{rec}} \approx 1/G$ when f_{sel} is smaller than the threshold (see Fig. 1(a)), the corresponding AUC_{real} is increasing in the same regime. Finally, AUC_{real} and AUC_{est} become consistent when f_{sel} further increases and the system achieves perfect recommendation.

The above results imply that the conventional evaluation of recommendation accuracy may not necessarily reflect the true accuracy. Indeed, AUC_{est} may over-estimate the accuracy of the algorithm, especially in cases where users rely frequently on the recommender system and do not reveal their own taste by deliberately selecting products. This serves as an alarm for researchers and operators of recommender systems. Alternative evaluations are therefore necessary to supplement conventional accuracy metrics to quantify the benefit of recommender systems for users.
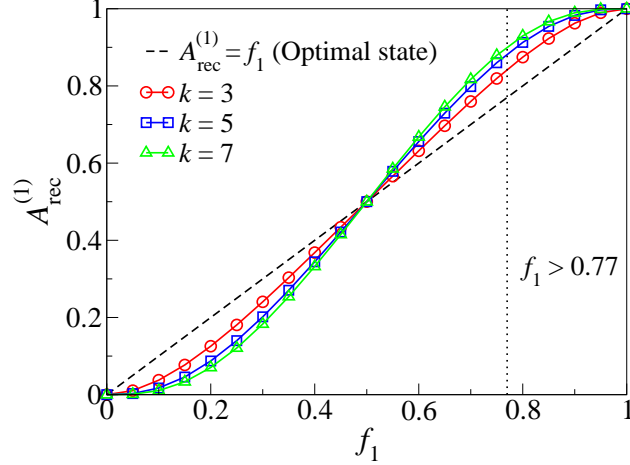
FIG. 4: The fraction $A_{\text{rec}}^{(1)}$ of recommended items in taste 1 as a function of f_1, the fraction of the selected products in taste 1. The simulations are obtained with $N = 2000$, $M = 100$, $G = 10$ and $f_{\text{sel}} = 0.95$ for $5 \times 10^4 N$ updates averaged over 50 instances. Only results obtained with common neighbor similarity are shown.

C. Users with multiple tastes

Ordinary users usually have more than one interests, for instance, a user may be interested in both scientific fiction and action movies. To model this scenario, we assume that each user is characterized by two tastes, which we denote by taste 1 and taste 2. Similar to the previous case, with f_{sel} of the times, the user selects a product in the absence of recommender systems; otherwise, the recommendation algorithm is applied. When a user selects a product, f_1 of the selected products are in taste 1 and the rest are in taste 2. To simplify the model, we only study cases with large f_{sel}, with which perfect recommendation is achieved in the original single-taste system.

Since a fraction f_1 of the selected products of the user are in taste 1, the ratio $f_1/(1 - f_1)$ corresponds to his/her preference between the two tastes. If optimal recommendations are achieved, f_1 of the recommended products should be in taste 1 and $1 - f_1$ of them should be in taste 2. Nevertheless, as shown in Fig. 4, the fraction $A_{\text{rec}}^{(1)}$ of the recommended products in taste 1 does not coincide with the optmial line $A_{\text{rec}}^{(1)} = f_1$. For instance, when f_1 is small, the recommendations are mainly in taste 2. It leads to a sub-optimal state which under-represent the minority taste, i.e. taste 1 when $f_1 < 0.5$, among the recommended products. Similarly, taste 2 is under-represented when $f_1 > 0.5$. As we can see in Fig. 4, the difference

between $A_{\text{rec}}^{(1)}$ and f_1 is larger when k is larger. This implies an increasing difficulty for the recommender system to identify a secondary taste if the user-product connections are denser. We remark that the results by employing the common neighbor similarity and the cosine similarity are almost identical.

On the other hand, one may expect a perfect recommendation regime at $f_1 f_{\text{sel}} > f_{\text{sel}}^*$, where f_{sel}^* denotes the threshold value, or equivalently the smallest f_{sel} at which the system achieves perfect recommendation in the corresponding single-taste scenario. For the system parameters employed in Fig. 4, $f_{\text{sel}}^* \approx 0.73$, but perfect recommendations in taste 1 are not achieved with $f_1 > f_{\text{sel}}^*/f_{\text{sel}} = 0.77$ (indicated by the dotted line in Fig. 4) due to the presence of taste 2.

D. Tests with empirical datasets

Finally, we incorporate our model with a real dataset obtained from MovieLens [16]. Since user taste and product genre are unknown in real systems, we again randomly divide the dataset into a training set and a probe set, and consider the recommended movie to be correct only if it was collected by the user and received a rating of 3 (in a scale from 1 to 5) from the user as recorded in the data. Similar to our model, with f_{sel} of the times, a user deliberately selects a correct movie and otherwise the recommendation algorithm is applied. For those users who rated at least two movies with a score of 3 or above, we set their degree to be $k_i - 1$ such that an un-collected correct movie always exists. As in the previous simulations, a user randomly removes one of his/her collected movies when he/she obtains a new movie; the system is then repeatedly updated.

As shown in Fig. 5(a), the accuracy A_{rec} obtained by both similarity definitions starts at a low value and increases with f_{sel}. Nevertheless, it does not show an abrupt jump to a high value similar to previous simulations but a plateau at small f_{sel} and a small jump at large f_{sel} are observed in the case with cosine similarity. These results again suggest that sufficient hints about user taste are essential for the system to obtain accurate recommendations. When f_{sel} approaches 1, A_{rec} decreases since users have collected most of the correct movies through deliberate selection and it becomes more difficult for the recommender system to identify the fewer correct items among all the other items.

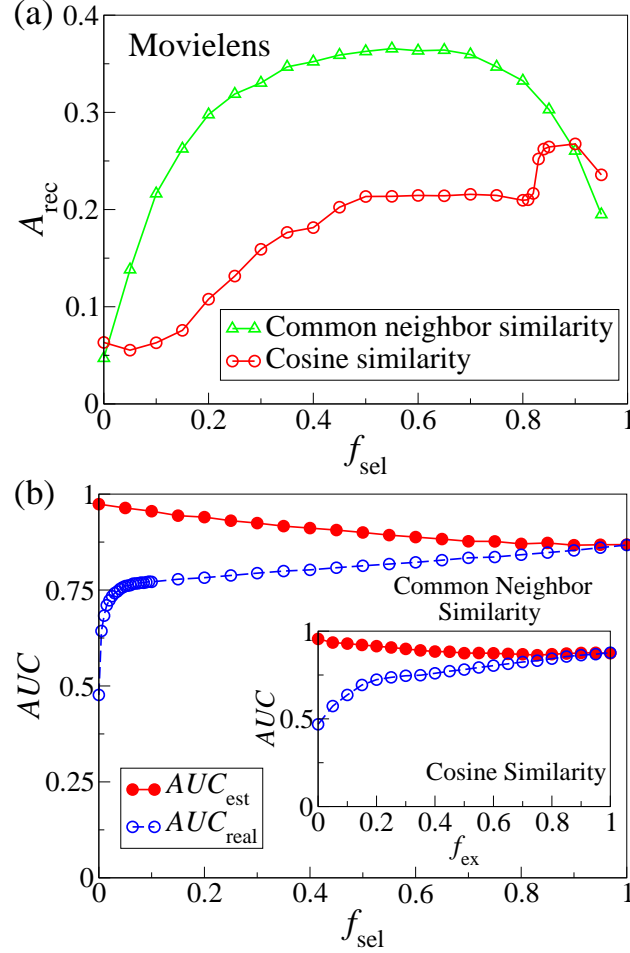As shown in Fig. 5(b), the dependence of AUC_{est} and AUC_{real} on f_{sel} is similar to that

FIG. 5: (a) The recommendation accuracy A_{rec} as a function of f_{sel}, obtained by incorporating our model with the MovieLens dataset with 944 users and 1683 products, and 5000 updates per user. (b) The corresponding estimated AUC_{est} and the real AUC_{real} as a function of f_{sel}.

observed from the previous simulations. When f_{sel} is small, the conventional AUC metric over-estimates the accuracy of the recommender system. Especially, AUC_{est} is highest when AUC_{real} is lowest, and $AUC_{\mathrm{est}} = AUC_{\mathrm{real}}$ only when $f_{\mathrm{sel}} = 1$. This suggests that conventional metrics may again be over-estimate recommendation accuracy in real systems.

E.   A recommendation algorithm with improvement

Based on the previous results, we slightly modify the ICF algorithm to improve the recommendation accuracy. The rationale is simple – since products deliberately selected by users usually match their taste, we simply give a higher weight to these products during
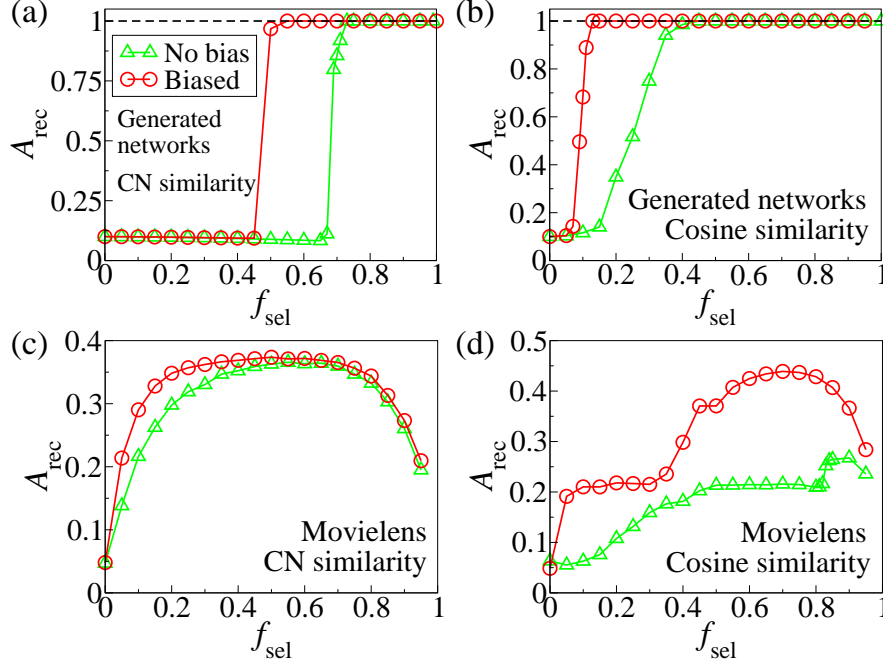
FIG. 6: The accuracy A_{rec} of the original ICF compared with ICF biased on products collected via deliberate selection (with $b = 2$ in Eq. (5)). The results are obtained by (a) the common neighbor (CN) similarity and (b) the cosine similarity on generated networks with $N = 2000$, $M = 100$, $k = 7$ and $G = 10$. The corresponding results on the MovieLens dataset are shown in (c) and (d).

the computation of recommendation scores, by modifying the adjacency variable $a_{i\alpha}(t)$ as follows:

$$
a_{i\alpha}(t) = \begin{cases} 0 & \text{if } \alpha \notin C_i(t), \\ 1 & \text{if } \alpha \in C_i(t) \text{ via recommendation}, \\ b & \text{if } \alpha \in C_i(t) \text{ via selection}, \end{cases} \tag{5}
$$

where $C_i(t)$ is again the set of products collected by user i at time t, and $b > 1$ is the *bias* on products collected via deliberate selection. The recommendation score of an item are then computed by the same formula Eq. (3). The recommendation accuracy obtained by the modified algorithm is compared to that of the original algorithm in Fig. 6. As we can see from Fig. 6(a) and (b), perfect recommendations are achieved at a smaller f_{sel} when selected products are weighed more in the algorithm. Similar results are observed with the MovieLens datasets as shown in Fig. 6(c) and (d). These results imply that products deliberately chosen by users are essential information to improve recommendation accuracy.

Discussion

To reveal the benefit of recommender systems for users, we studied a simple model where users either choose their own products or follow the recommendations from the system. Our results show that the recommendations may be equivalent to random draws if users rely too strongly on the recommender system and do not reveal their own taste by deliberately selecting products. On the other hand, if sufficient information about their taste is present, recommendation systems are able to achieve high accuracy in matching appropriate products to users. For some recommendation algorithms, the increase in accuracy is abrupt once the amount of available information exceeds a threshold. These results imply that recommender systems can benefit users, but relying too strongly on the system may render the system ineffective.

On the other hand, our study reveals the difficulties to obtain a realistic and accurate evaluation of recommendation accuracy. Since real user preference is unknown, evaluation of recommender algorithms usually involves removing a set of existing data and quantifies their accuracy by their success to retrieve the removed set. Our results show that such metrics do not necessarily reflect and may over-estimate the true accuracy of the algorithm. This is because the choice of products collected by users was previously influenced by the recommendation algorithms; the presence of these products may not reflect their true preference and may favor the evaluation by the conventional accuracy metrics. The disagreement between the estimated and the real accuracy was observed in simulations with both generated network and a real dataset. These results imply that a high recommendation accuracy indicated by the conventional metrics may not necessarily imply a benefit for users. Alternative evaluations are necessary to supplement these metrics in order to quantify the effectiveness of the recommender systems.

Additional information

Author contributions C. H. Y. designed the research, performed the experiments, analyzed the data and wrote the manuscript.

Competing financial interests The author declares no competing financial interests.

[1] Resnick, P. & Varian, H. R. Recommender Systems. *Commun. ACM* 56–58 (1997).

[2] Schafer, J. B., Konstan, J. & Riedl, J. Recommender Systems in E-Commerce. *Proceedings of the 1st ACM conference on Electronic commerce* 158–166 (1999).

[3] Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B. *Recommender systems handbook* (Springer, New York, USA, 2011).

[4] Lu, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K. & Zhou, T. Recommender Systems. *Physics Reports* **519**, 1–49 (2012).

[5] Liu, J., Pedersen, E. & Dolan, P. Personalized News Recommendation Based on Click Behavior. *Proceedings of the 15th international conference on Intelligent user interfaces* 31–40 (2010).

[6] Chen, J., Geyer, W., Dugan, C., Muller, M. & Guy, I. Make New Friends, but Keep the Old: Recommending People on Social Networking Sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 201–210 (2009).

[7] Brynjolfsson, E., Hu, Y. & Smith, M. D. Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. *Management Science* **49**, 1580–1596 (2003).

[8] Chen, P.-Y., Wu, S.-Y. & Yoon, J. The impact of online recommendations and consumer feedback on sales. *Proceedings of the 25th International Conference on Information Systems* 711–724 (2004).

[9] Fleder, D. & Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* **55**, 697–712 (2009).

[10] Takács, G., Pilászy, I., Németh, B. & Tikk, D. On the Gravity Recommendation System. *Proc. KDD Cup Workshop at SIGKDD, San Jose, California* 22V30 (2007).

[11] Blei, D. L., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *The Journal of Machine Learning Research* **3**, 993–1022 (2003).

[12] Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. USA* **101**, 5228–5235 (2004).

[13] Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R. & Zhang, Y.-C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. USA* **107**, 4511–4515 (2010).

[14] Liang, T.-P., Lai, H.-J. & Ku, Y.-C. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings. *Journal of Management Information Systems* 45–70 (2006).

[15] Zeng, A., Yeung, C. H., Shang, M.-S. & Zhang, Y.-C. The reinforcing influence of recommendations on global diversification. *Europhys Lett.* **97**, 18005 (2012).

[16] The website of MovieLens:. `http://grouplens.org/datasets/movielens/` (2015).

[17] Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International Conference on World Wide Web* 285–295 (2001).

Do recommender systems benefit users?

Chi Ho Yeung

Department of Science and Environmental Studies,

The Hong Kong Institute of Education,

10 Lo Ping Road, Taipo, Hong Kong

(Dated: June 19, 2015)

Recommender systems are present in many web applications to guide our choices. They increase sales and benefit sellers, but whether they benefit customers by providing relevant products is questionable. Here we introduce a model to examine the benefit of recommender systems for users, and found that recommendations from the system can be equivalent to random draws if one relies too strongly on the system. Nevertheless, with sufficient information about user preferences, recommendations become accurate and an abrupt transition to this accurate regime is observed for some algorithms. On the other hand, we found that a high accuracy evaluated by common accuracy metrics does not necessarily correspond to a high real accuracy nor a benefit for users, which serves as an alarm for operators and researchers of recommender systems. We tested our model with a real dataset and observed similar behaviors. Finally, a recommendation approach with improved accuracy is suggested. These results imply that recommender systems can benefit users, but relying too strongly on the system may render the system ineffective.

Introduction

Almost all popular websites employ recommender systems to match users with items [1–4]. For instance, news websites analyze the reading history of individuals and recommend news which match their interests [5]; online social networks recommend new friends to individuals based on their existing friends [6]. Most commonly, online retailers analyze the purchase history of customers and recommend products to them to increase their own sales [7–9]. These examples show an increasingly crucial role of recommender systems in our daily life, influencing our various choices.

Due to their broad applications, great efforts have been devoted to study recommendation algorithms and to improve their accuracy [4]. Researchers in computer science, mathematics and management science employ various mathematical tools such as Bayesian approach and matrix factorization to derive recommendation algorithms [4, 10–12]. Recently, physicists and complex system scientists started to work in the area and incorporated physical processes such as mass diffusion and heat conduction to recommender system [13]. Nevertheless, the main goal of these studies is limited to recommendation accuracy, but their genuine benefits are less examined.

Although recommender systems have been shown to benefit retailers, whether the recommended products are relevant to customers is questionable [9, 14]. On one hand, many recommendation algorithms are based on product similarity and the recommended products may be redundant since they are similar to the already purchased products [13]. On the other hand, instead of specific products which match individual needs, many recommender systems can only recommend popular but potentially irrelevant products [9, 15]. Nevertheless, users may be tempted to purchase the products due to recommendations, and in this case recommender systems benefit sellers but not customers.

In this paper, we introduce a simple model to examine the relevance between the recommended products and the preferences of users. Unlike empirical studies where the true user preference is unknown, each user in the model is characterized by a taste and the true recommendation accuracy can be measured. We found that recommendations can be either random or very accurate depending on the frequency the users select a product without recommendations. For some algorithms, an abrupt increase in accuracy is observed when this frequency exceeds a threshold. On the other hand, we found that a high accuracy in-

dicated by common evaluation metrics does not necessarily imply to a high real accuracy. We tested our model using the MovieLens dataset [16] and observed similar behaviors. Finally, a recommendation approach based on our findings was suggested which outperforms conventional approaches.

Model

Specifically, we consider a group of N users selecting products from a group of M items. Each user i and item α is characterized by one of the G *tastes* or *genres*, denoted by g_i and g_α respectively. For instance, in terms of movies, these tastes may correspond to science fictions, romantic comedies or thrillers. The case where users have multiple tastes are described in Section C.

At each time step, a user i is randomly drawn. With a fraction f_sel of the times, user i chooses a product matching his/her own taste without using the recommender system. This is the conventional way to purchase a product and we call f_sel the *frequency of deliberate selection*. On the other hand, with a fraction $1 - f_\text{sel}$ of the times, user i buys a product following the recommender system. In both cases, a product in his/her collection is randomly removed since all products are assumed to be consumable and can be brought and consumed for more than once. In this case, the total number of products collected by user i remains constant at k_i, which simplifies our model as network growth is not required and N and M remain constant. The above procedures are repeated for a large number of times per user.

We remark that the recommender system has no direct knowledge of user taste and product genre, it can only infer user preferences through his/her purchase history. Since f_sel is the frequency a user makes purchases in the absence of recommender systems, on average at least f_sel of the purchases of user i must match his/her taste; f_sel is thus proportional to the amount of *available hints* the recommender systems can exploit. We further define *recommendation accuracy* A_rec to be the fraction of recommended products which match the taste of the user, and our goal is to examine A_rec to reveal the benefit of recommender systems to users.

For simplicity, we employ the common *Item-based Collaborative Filtering* (ICF) [17] to be the recommendation algorithm in our model. ICF provides personalized recommendations to users by computing similarity between their purchased products with other products.

We first denote the similarity between item α and β at time t to be $s_{\alpha\beta}(t)$. As shown by previous studies [17], the performance of the algorithm is strongly dependent on the definition of similarity. To shown that our results are relevant to different recommendation algorithms, we will employ two definitions of similarity, namely the *common neighbor* (CN) *similarity*, given by

$$s_{\alpha\beta}^{(\text{CN})}(t) = \sum_{i=0}^{N} a_{i\alpha}(t)a_{i\beta}(t),\tag{1}$$

and the *cosine similarity* [17], given by

$$s_{\alpha\beta}^{(\text{cosine})}(t) = \frac{1}{\sqrt{k_\alpha k_\beta}} \sum_{i=0}^{N} a_{i\alpha}(t)a_{i\beta}(t).\tag{2}$$

The adjacency variable $a_{i\alpha}(t) = 1$ if item α is collected by user i at time t, and otherwise $a_{i\alpha}(t) = 0$. The recommendation score $r_{i\alpha}(t)$ of product α for user i at time t is given by

$$r_{i\alpha}(t) = \sum_{\beta=1}^{M} a_{i\beta}(t)s_{\alpha\beta}(t) = \sum_{\beta \in C_i(t)} s_{\alpha\beta}(t),\tag{3}$$

where $C_i(t)$ is the set of products collected by user i at time t. Finally, the product with the highest score not yet collected by the user is recommended.

Results

A. Random versus accurate recommendations

To examine the benefit of recommender system to users, we first study the dependence of recommendation accuracy A_{rec} on the frequency f_{sel} of deliberate selection. The higher the value of f_{sel}, the more often the user chooses a product of a matching taste without recommendation, and the more the information for the recommender system to exploit. If recommender systems work perfectly, $A_{\text{rec}} = 100\% = 1$ whenever $f_{\text{sel}} > 0$ as there exists non-zero information about user tastes in the dataset; on the other hand, if recommender systems do not work at all, recommendations are always random, and $A_{\text{rec}} = 1/G$ independent of f_{sel}.

As shown in Fig. 1, the recommendation accuracy falls between the two extreme cases. The common neighbor similarity is employed in Fig. 1(a), and $A_{\text{rec}} \approx 1/G$ which corresponds

to the case of random recommendations when f_{sel} is less than a threshold. When f_{sel} increases beyond the threshold, recommendation accuracy increases abruptly to $A_{\text{rec}} = 1$, which corresponds to a case of perfect recommendation. As shown in Fig. 1(b), cosine similarity is employed and a similar dependence of A_{rec} on f_{sel} is observed, though the transition between the two phases is more gentle. We remark that $A_{\text{rec}} = 1$ is an artifact of the model since each user and product is categorized by only one taste, and after users and products of the same taste formed an isolated bipartite cluster, only products within the cluster are recommended and lead to a persistent perfect accuracy.

The accuracy A_{rec} is also dependent on the number of taste group G. Intuitively, the threshold value for perfect recommendation decreases with G, since it seems easier to identify an item with the correct taste out of a smaller number of taste groups. However, simulated results in both Fig. 1(a) and (b) show that the threshold value increases when G decreases. It is because users collect products of both relevant and irrelevant taste; when G is small, the irrelevant products belong to a small number of taste groups, and there exists a strong connection between users and each irrelevant taste group, making it difficult for the recommender system to identify these false connections. In short, the more diverse and distinct the users and products, the less amount of hints are required to provide correct recommendations.

Other than the number of taste group, recommendation accuracy also depends on the number of items collected by each user. For simplicity, all users collect the same number of items, i.e. $k_i = k$ for $\forall i$. As shown in Fig. 2(a) and (b), perfect recommendation is more difficult to be achieved for cases with larger k, where the stronger connection between users and irrelevant taste groups is again the reason. These results imply that when users collect a large number of products, false connections exist and may impact negatively on the recommender system. Hence, instead of drawing recommendations based on all the available data, an algorithm which effectively eliminates the false connections may lead to a high recommendation accuracy.

The above results suggest that recommender systems may provide irrelevant recommendations when users do not provide sufficient hints about their taste. On the other hand, given sufficient hints, recommender systems well utilize the information to match users with products. The amount of hints required for accurate recommendation is different for different algorithms and systems.

B. Estimated accuracy versus real accuracy

In real systems, since the real preference of users is unknown, there is no way to measure the real recommendation accuracy. Various metrics are thus introduced to evaluate recommendation accuracy. Nevertheless, whether these metrics correctly measure real accuracy is questionable. Since user taste and product genre are defined in our model, we can compare the accuracy measured by these metrics with the real accuracy.

One common metric to evaluate recommendation accuracy is AUC, i.e. the area under the *receiver operating curve* (ROC). When recommendations are made for user i, AUC is computed as the probability that a correct product α is ranked higher than an arbitrary product γ, given by

$$AUC_{i\alpha} = \frac{n(r_{i\gamma} < r_{i\alpha}) + 0.5n(r_{i\gamma} = r_{i\alpha})}{M - k_i} \tag{4}$$

where $n(r_{i\gamma} < r_{i\alpha})$ is the number of products with score $r_{i\gamma}$ lower than the score $r_{i\alpha}$ of the correct product, and $n(r_{i\gamma} = r_{i\alpha})$ is the number of items which tie with the correct item. Based on the definition of *correct predictions*, we compute two AUC measures - (i) the conventional estimated AUC_{est}, obtained by dividing the dataset into a training set and a probe set; links in the probe set are removed and are considered to be correct predictions if their existence are predicted; and (ii) the real AUC_{real} which quantifies the accuracy of the algorithm in recommending products of a matching taste.

The dependence of AUC_{est} and AUC_{real} on f_{sel} is shown in Fig. 3. As we can see, $AUC_{\text{real}} \approx 0.5$ when f_{sel} is small since recommendations are random (see Fig. 1) and the products of a matching taste are randomly ranked in the recommendation list. However, AUC_{est} is much higher and is not consistent with AUC_{real}. The reason for a large AUC_{est} at small f_{sel} is the frequent application of recommender systems, such that user purchases are strongly influenced by the algorithms regardless of their true preference. In this case, products which do not match their preference but are consistent with the algorithms are also collected by the users. This favors the evaluation by AUC_{est} using a random probe set, and lead to a high AUC_{est} even random recommendations are indeed provided.

When f_{sel} increases, AUC_{est} decreases since the user-product relations become less influenced by the recommender system. At the same time, AUC_{real} increases since more hints about the user tastes are present. We remark that although $A_{\text{rec}} \approx 1/G$ when f_{sel} is smaller

than the threshold (see Fig. 1(a)), the corresponding AUC_{real} is increasing in the same regime. Finally, AUC_{real} and AUC_{est} become consistent when f_{sel} further increases and the system achieves perfect recommendation.

The above results imply that the conventional evaluation of recommendation accuracy may not necessarily reflect the true accuracy. Indeed, AUC_{est} may over-estimate the accuracy of the algorithm, especially in cases where users rely frequently on the recommender system and do not reveal their own taste by deliberately selecting products. This serves as an alarm for researchers and operators of recommender systems. Alternative evaluations are therefore necessary to supplement conventional accuracy metrics to quantify the benefit of recommender systems for users.

C. Users with multiple tastes

Ordinary users usually have more than one interests, for instance, a user may be interested in both scientific fiction and action movies. To model this scenario, we assume that each user is characterized by two tastes, which we denote by taste 1 and taste 2. Similar to the previous case, with f_{sel} of the times, the user selects a product in the absence of recommender systems; otherwise, the recommendation algorithm is applied. When a user selects a product, f_1 of the selected products are in taste 1 and the rest are in taste 2. To simplify the model, we only study cases with large f_{sel}, with which perfect recommendation is achieved in the original single-taste system.

Since a fraction f_1 of the selected products of the user are in taste 1, the ratio $f_1/(1-f_1)$ corresponds to his/her preference between the two tastes. If optimal recommendations are achieved, f_1 of the recommended products should be in taste 1 and $1-f_1$ of them should be in taste 2. Nevertheless, as shown in Fig. 4, the fraction $A_{\text{rec}}^{(1)}$ of the recommended products in taste 1 does not coincide with the optmial line $A_{\text{rec}}^{(1)} = f_1$. For instance, when f_1 is small, the recommendations are mainly in taste 2. It leads to a sub-optimal state which under-represent the minority taste, i.e. taste 1 when $f_1 < 0.5$, among the recommended products. Similarly, taste 2 is under-represented when $f_1 > 0.5$. As we can see in Fig. 4, the difference between $A_{\text{rec}}^{(1)}$ and f_1 is larger when k is larger. This implies an increasing difficulty for the recommender system to identify a secondary taste if the user-product connections are denser. We remark that the results by employing the common neighbor similarity and the

cosine similarity are almost identical.

On the other hand, one may expect a perfect recommendation regime at $f_1 f_{\text{sel}} > f_{\text{sel}}^*$, where f_{sel}^* denotes the threshold value, or equivalently the smallest f_{sel} at which the system achieves perfect recommendation in the corresponding single-taste scenario. For the system parameters employed in Fig. 4, $f_{\text{sel}}^* \approx 0.73$, but perfect recommendations in taste 1 are not achieved with $f_1 > f_{\text{sel}}^*/f_{\text{sel}} = 0.77$ (indicated by the dotted line in Fig. 4) due to the presence of taste 2.

D.   Tests with empirical datasets

Finally, we incorporate our model with a real dataset obtained from MovieLens [16]. Since user taste and product genre are unknown in real systems, we again randomly divide the dataset into a training set and a probe set, and consider the recommended movie to be correct only if it was collected by the user and received a rating of 3 (in a scale from 1 to 5) from the user as recorded in the data. Similar to our model, with f_{sel} of the times, a user deliberately selects a correct movie and otherwise the recommendation algorithm is applied. For those users who rated at least two movies with a score of 3 or above, we set their degree to be $k_i - 1$ such that an un-collected correct movie always exists. As in the previous simulations, a user randomly removes one of his/her collected movies when he/she obtains a new movie; the system is then repeatedly updated.

As shown in Fig. 5(a), the accuracy A_{rec} obtained by both similarity definitions starts at a low value and increases with f_{sel}. Nevertheless, it does not show an abrupt jump to a high value similar to previous simulations but a plateau at small f_{sel} and a small jump at large f_{sel} are observed in the case with cosine similarity. These results again suggest that sufficient hints about user taste are essential for the system to obtain accurate recommendations. When f_{sel} approaches 1, A_{rec} decreases since users have collected most of the correct movies through deliberate selection and it becomes more difficult for the recommender system to identify the fewer correct items among all the other items.

As shown in Fig. 5(b), the dependence of AUC_{est} and AUC_{real} on f_{sel} is similar to that observed from the previous simulations. When f_{sel} is small, the conventional AUC metric over-estimates the accuracy of the recommender system. Especially, AUC_{est} is highest when AUC_{real} is lowest, and $AUC_{\text{est}} = AUC_{\text{real}}$ only when $f_{\text{sel}} = 1$. This suggests that conventional

metrics may again be over-estimate recommendation accuracy in real systems.

E. A recommendation algorithm with improvement

Based on the previous results, we slightly modify the ICF algorithm to improve the recommendation accuracy. The rationale is simple – since products deliberately selected by users usually match their taste, we simply give a higher weight to these products during the computation of recommendation scores, by modifying the adjacency variable $a_{i\alpha}(t)$ as follows:

$$
a_{i\alpha}(t) = \begin{cases} 0 & \text{if } \alpha \notin C_i(t), \\ 1 & \text{if } \alpha \in C_i(t) \text{ via recommendation,} \\ b & \text{if } \alpha \in C_i(t) \text{ via selection,} \end{cases} \tag{5}
$$

where $C_i(t)$ is again the set of products collected by user i at time t, and $b > 1$ is the *bias* on products collected via deliberate selection. The recommendation score of an item are then computed by the same formula Eq. (3). The recommendation accuracy obtained by the modified algorithm is compared to that of the original algorithm in Fig. 6. As we can see from Fig. 6(a) and (b), perfect recommendations are achieved at a smaller f_{sel} when selected products are weighed more in the algorithm. Similar results are observed with the MovieLens datasets as shown in Fig. 6(c) and (d). These results imply that products deliberately chosen by users are essential information to improve recommendation accuracy.

Discussion

To reveal the benefit of recommender systems for users, we studied a simple model where users either choose their own products or follow the recommendations from the system. Our results show that the recommendations may be equivalent to random draws if users rely too strongly on the recommender system and do not reveal their own taste by deliberately selecting products. On the other hand, if sufficient information about their taste is present, recommendation systems are able to achieve high accuracy in matching appropriate products to users. For some recommendation algorithms, the increase in accuracy is abrupt once the amount of available information exceeds a threshold. These results imply that recommender

systems can benefit users, but relying too strongly on the system may render the system ineffective.

On the other hand, our study reveals the difficulties to obtain a realistic and accurate evaluation of recommendation accuracy. Since real user preference is unknown, evaluation of recommender algorithms usually involves removing a set of existing data and quantifies their accuracy by their success to retrieve the removed set. Our results show that such metrics do not necessarily reflect and may over-estimate the true accuracy of the algorithm. This is because the choice of products collected by users was previously influenced by the recommendation algorithms; the presence of these products may not reflect their true preference and may favor the evaluation by the conventional accuracy metrics. The disagreement between the estimated and the real accuracy was observed in simulations with both generated network and a real dataset. These results imply that a high recommendation accuracy indicated by the conventional metrics may not necessarily imply a benefit for users. Alternative evaluations are necessary to supplement these metrics in order to quantify the effectiveness of the recommender systems.

Acknowledgments

[1] Resnick, P. & Varian, H. R. Recommender Systems. *Commun. ACM* 56–58 (1997).

[2] Schafer, J. B., Konstan, J. & Riedl, J. Recommender Systems in E-Commerce. *Proceedings of the 1st ACM conference on Electronic commerce* 158–166 (1999).

[3] Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B. *Recommender systems handbook* (Springer, New York, USA, 2011).

[4] Lu, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K. & Zhou, T. Recommender Systems. *Physics Reports* **519**, 1–49 (2012).

[5] Liu, J., Pedersen, E. & Dolan, P. Personalized News Recommendation Based on Click Behavior. *Proceedings of the 15th international conference on Intelligent user interfaces* 31–40

(2010).

[6] Chen, J., Geyer, W., Dugan, C., Muller, M. & Guy, I. Make New Friends, but Keep the Old: Recommending People on Social Networking Sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 201–210 (2009).

[7] Brynjolfsson, E., Hu, Y. & Smith, M. D. Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. *Management Science* **49**, 1580–1596 (2003).

[8] Chen, P.-Y., Wu, S.-Y. & Yoon, J. The impact of online recommendations and consumer feedback on sales. *Proceedings of the 25th International Conference on Information Systems* 711–724 (2004).

[9] Fleder, D. & Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* **55**, 697–712 (2009).

[10] Takács, G., Pilászy, I., Németh, B. & Tikk, D. On the Gravity Recommendation System. *Proc. KDD Cup Workshop at SIGKDD, San Jose, California* 22V30 (2007).

[11] Blei, D. L., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *The Journal of Machine Learning Research* **3**, 993–1022 (2003).

[12] Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. USA* **101**, 5228–5235 (2004).

[13] Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R. & Zhang, Y.-C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. USA* **107**, 4511–4515 (2010).

[14] Liang, T.-P., Lai, H.-J. & Ku, Y.-C. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings. *Journal of Management Information Systems* 45–70 (2006).

[15] Zeng, A., Yeung, C. H., Shang, M.-S. & Zhang, Y.-C. The reinforcing influence of recommendations on global diversification. *Europhys Lett.* **97**, 18005 (2012).

[16] The website of MovieLens:. `http://grouplens.org/datasets/movielens/` (2015).

[17] Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International Conference on World Wide Web* 285–295 (2001).
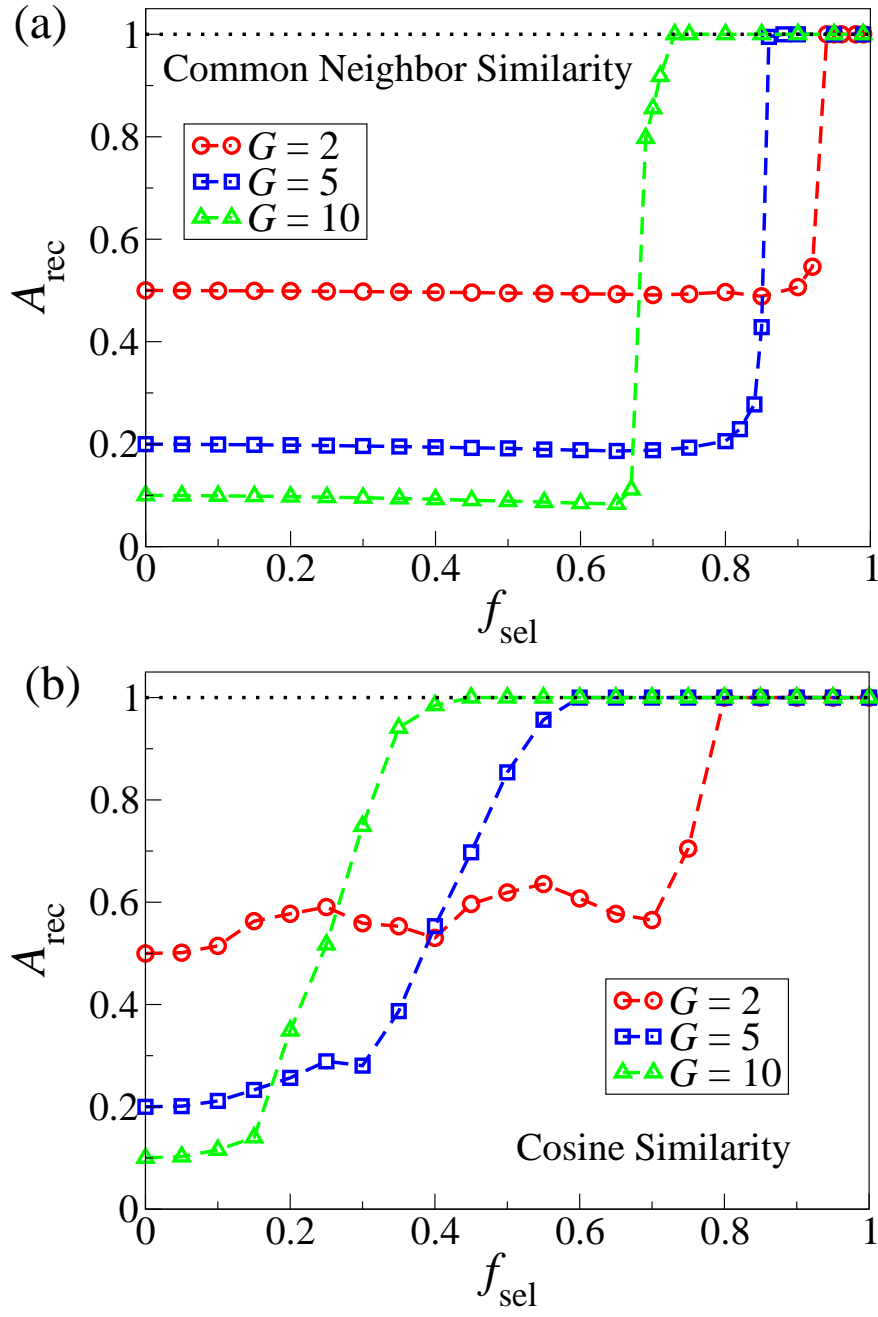
FIG. 1: The accuracy $A_{\rm rec}$ of the recommender system as a function of $f_{\rm sel}$ for different number of taste groups G. The simulation results were obtained with $N = 2000$ users and $M = 100$ products. Each user collects $k = 7$ products and is updated 1×10^5 times. Each data point was averaged over 50 instances. The common neighbor similarity Eq. (1) and the cosine similarity Eq. (2) were employed in (a) and (b) respectively.
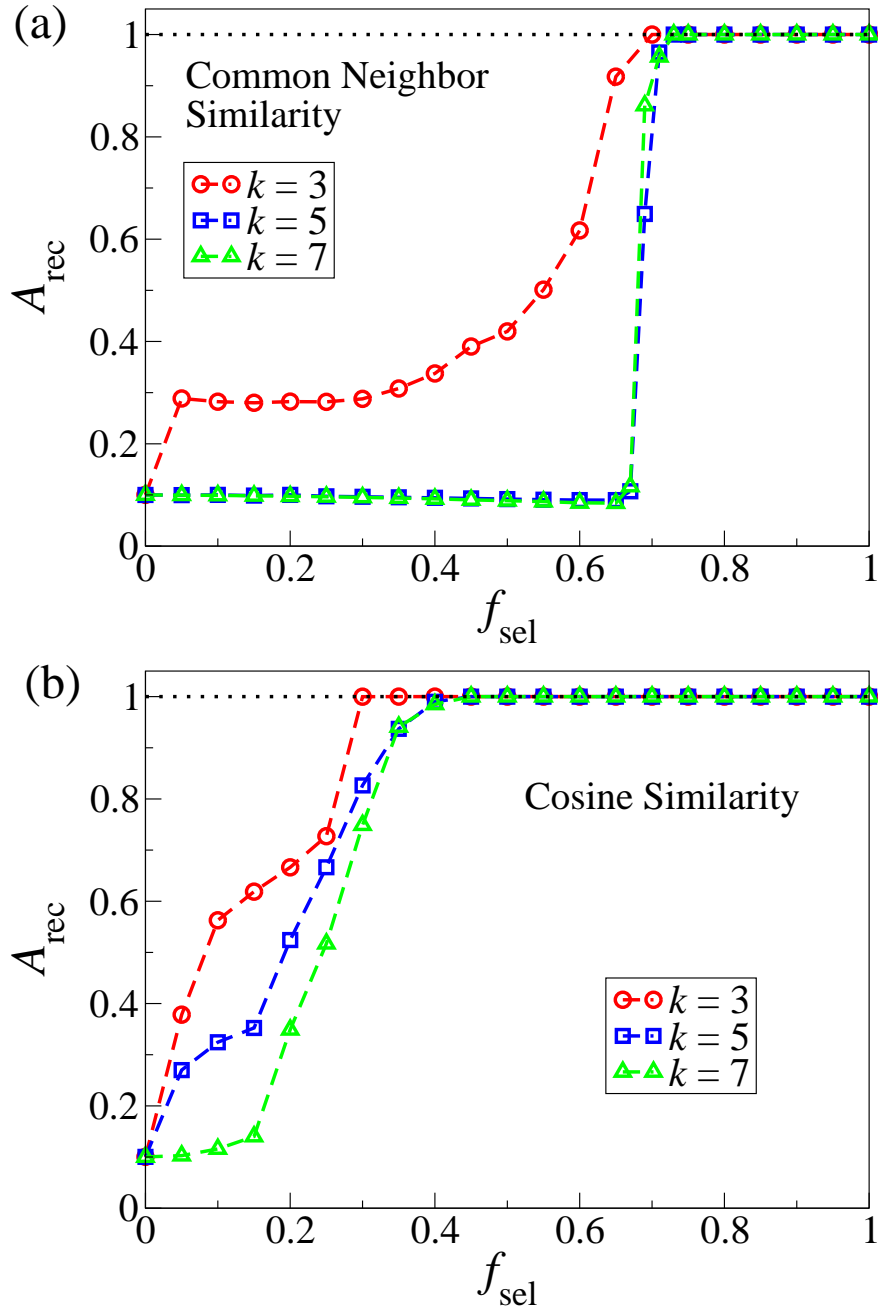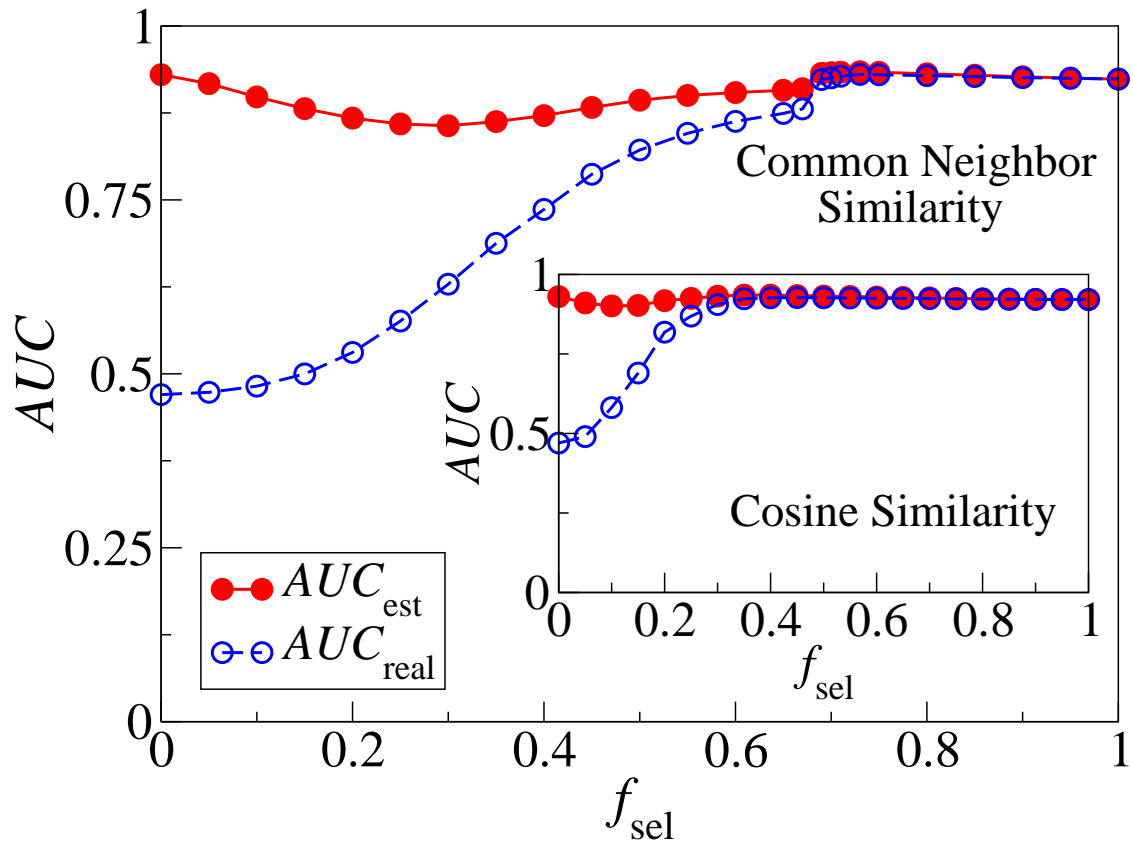
FIG. 2: The accuracy A_{rec} of the recommender system as a function of f_{sel} for different values of k, the number of products collected per user. The simulation results were obtained with $N = 2000$, $M = 100$ and $G = 10$. Each user was updated 1×10^5 times, and each data point was averaged over 50 instances. The common neighbor similarity Eq. (1) and the cosine similarity Eq. (2) were employed in (a) and (b) respectively.

FIG. 3: The two different AUC measures, AUC_{est} and AUC_{real}, as a function of f_{sel}, obtained by ICF with common neighbor similarity and cosine similarity (inset) on systems with $N = 2000$, $M = 100$, $k = 3$ and $G = 10$.
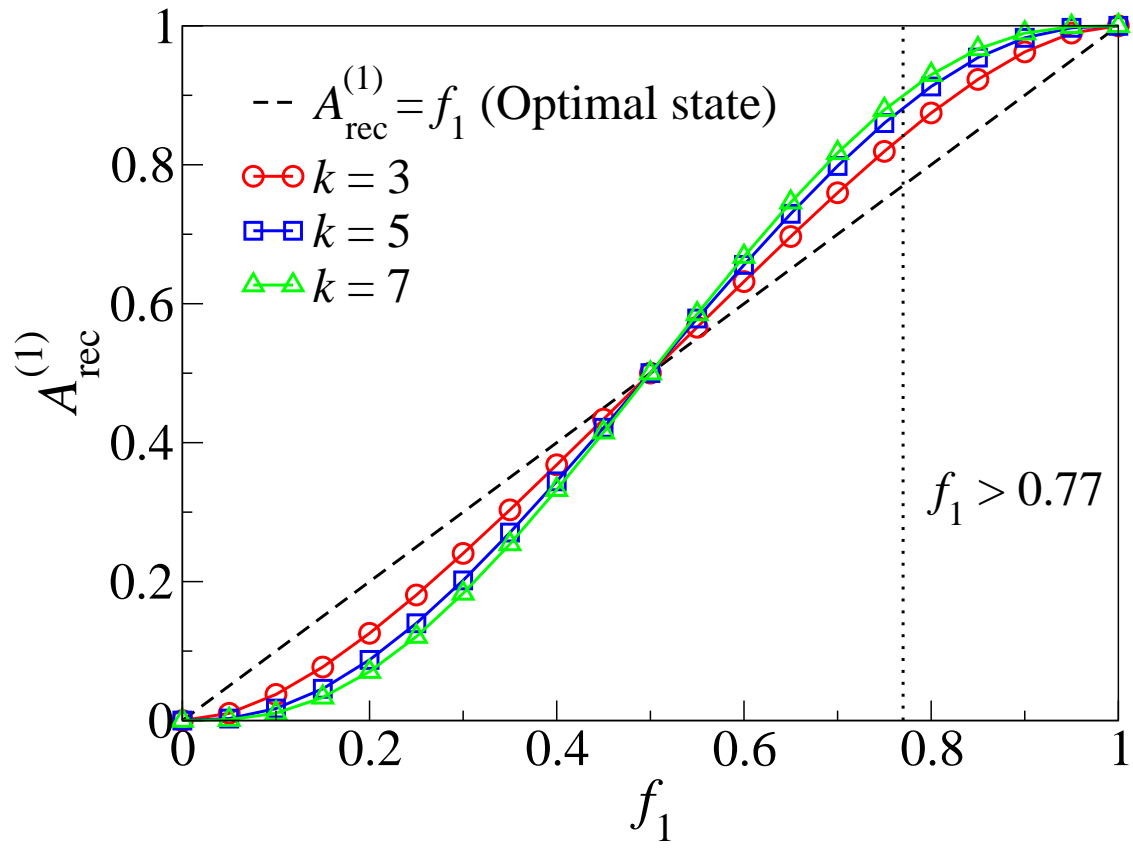
FIG. 4: The fraction $A_{rec}^{(1)}$ of recommended items in taste 1 as a function of f_1, the fraction of the selected products in taste 1. The simulations are obtained with $N = 2000$, $M = 100$, $G = 10$ and $f_{sel} = 0.95$ for $5 \times 10^4 N$ updates averaged over 50 instances. Only results obtained with common neighbor similarity are shown.
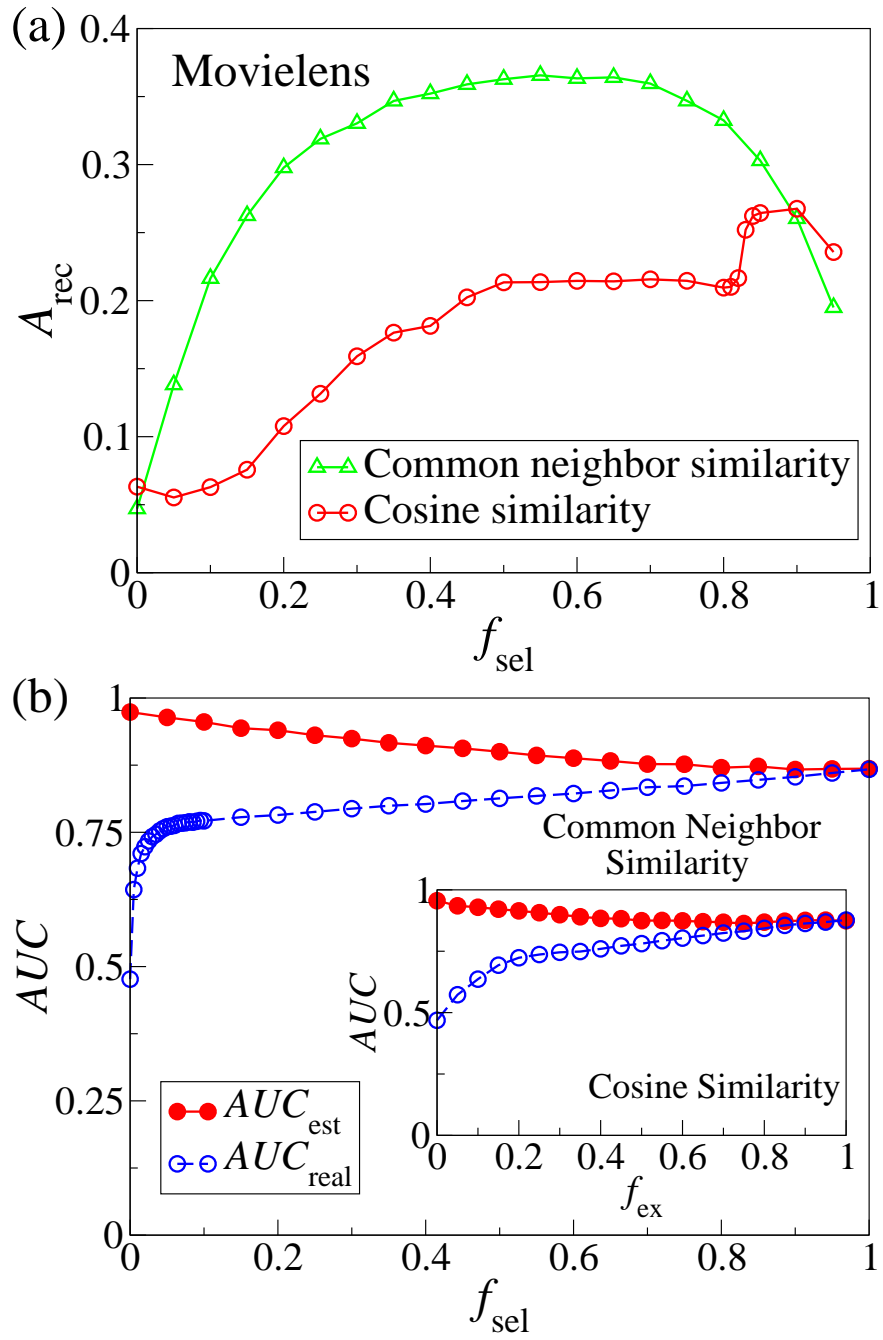
FIG. 5: (a) The recommendation accuracy A_{rec} as a function of f_{sel}, obtained by incorporating our model with the MovieLens dataset with 944 users and 1683 products, and 5000 updates per user. (b) The corresponding estimated AUC_{est} and the real AUC_{real} as a function of f_{sel}.
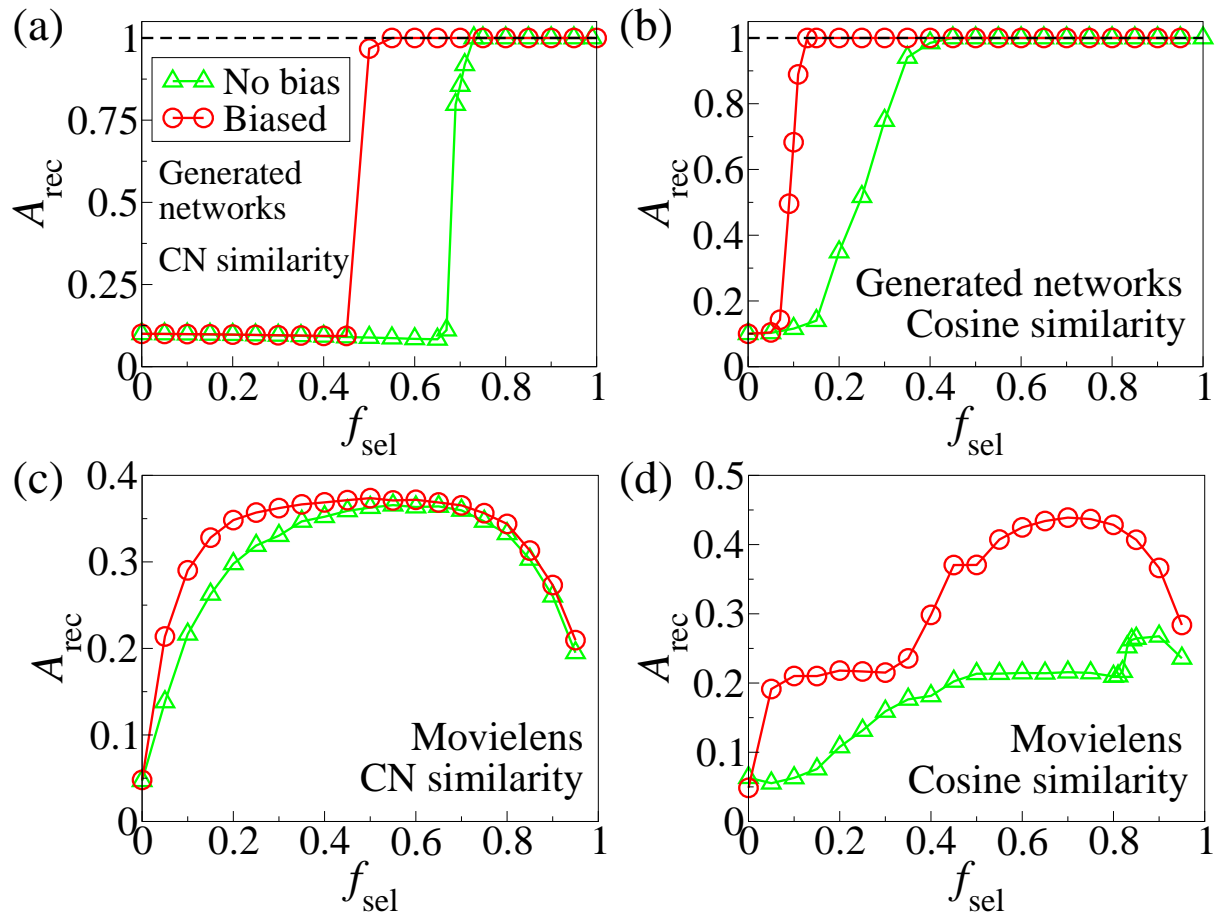
FIG. 6: The accuracy A_{rec} of the original ICF compared with ICF biased on products collected via deliberate selection (with $b = 2$ in Eq. (5)). The results are obtained by (a) the common neighbor (CN) similarity and (b) the cosine similarity on generated networks with $N = 2000$, $M = 100$, $k = 7$ and $G = 10$. The corresponding results on the MovieLens dataset are shown in (c) and (d).