Decision Support Systems 69 (2015) 92–106

Contents lists available at ScienceDirect

Decision Support Systems

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

# Alike people, alike interests? Inferring interest similarity in online social networks

Xiao Han [a,*], Leye Wang [a], Noel Crespi [a], Soochang Park [a], Ángel Cuevas [a,b]

[a] Institut-Mines Télécom, Télécom SudParis, 9 rue Charles Fourier, 91011 Evry Cedex France
[b] Universidad Carlos III de Madrid, Av de la Universidad, 30 28911 Legans, Madrid, Spain

ARTICLE INFO

*Article history:*
Received 16 April 2014
Received in revised form 22 September 2014
Accepted 30 November 2014
Available online 9 December 2014

*Keywords:*
Social networks
Interest similarity
Homophily
Prediction model

ABSTRACT

Understanding how much two individuals are alike in their interests (i.e., *interest similarity*) has become virtually essential for many applications and services in Online Social Networks (OSNs). Since users do not always explicitly elaborate their interests in OSNs like Facebook, how to determine users' interest similarity without fully knowing their interests is a practical problem. In this paper, we investigate how users' interest similarity relates to various social features (e.g. geographic distance); and accordingly infer whether the interests of two users are alike or unalike where one of the users' interests are unknown. Relying on a large Facebook dataset, which contains 479,048 users and 5,263,351 user-generated interests, we present comprehensive empirical studies and verify the *homophily* of interest similarity across three interest domains (movies, music and TV shows). The homophily reveals that people tend to exhibit more similar tastes if they have similar demographic information (e.g., age, location), or if they are friends. It also shows that the individuals with a higher interest entropy usually share more interests with others. Based on these results, we provide a practical *prediction model* under a real OSN environment. For a given user with no interest information, this model can select some individuals who not only exhibit many interests but also probably achieve high interest similarities with the given user. Eventually, we illustrate a use case to demonstrate that the proposed prediction model could facilitate decision-making for OSN applications and services.



## 1. Introduction

Online Social Networks (OSNs) have boomed and attracted a huge number of people to join them over the last decade. In OSNs, participants publish their profiles, make friends, and produce various contents (photos, answers/questions, videos, etc.). Unlike legacy web systems, OSNs are organized around both people and content, which provide us with unprecedented opportunities to understand human relationships, human communities, human behaviors and human preferences [13, 17, 27].

With the evolution of OSNs, understanding to what extent two individuals are alike in their interests (i.e., interest similarity) has become a basic requirement for the organization and maintenance of vibrant OSNs. On the one hand, such information about users' interest similarity could be leveraged to support friend recommendation and social circle maintenance. For instance, the decision to recommend users who share many interests with each other to be friends could increase users' approval rate of recommendation, because people usually aggregate by their mutual interests [14]. On the other hand, knowing interest similarity between users also facilitates social applications and advertising. For example, instead of randomly hunting for clients, exploring those users with a high interest similarity with existing clients could efficiently enlarge client groups for application providers and businesses.

However, estimating interest similarity between two users is not a straight-forward issue since users do not always explicitly elaborate their interests. In the Facebook data set prepared for this study, 51.6% of users do not present any interests in their profiles; and among nine interest domains in the dataset, except for movies, music and TV shows, less than a quarter of users reveal their interests in any of the other six interest domains (e.g., books, sports or games). Since such lack of users' interests occurs quite often in the real OSN environment, how to infer two users' interest similarity without complete information about their interests poses a challenge.

To deal with this problem, we investigate how two users' interest similarity relates to various social features in depth (e.g. profile overlap, geographic distance, and friend similarity) and further infer whether two users are alike/unalike in interest according to these learned relations. Existing studies have already demonstrated that friends share more interests than strangers [1] and verified that interest similarity strongly correlates to the trust between users [32]. However, the work to date has not address the issue of inferring users' interest similarity without complete information about users' interests. Furthermore, we carry out a comprehensive analysis on the correlations between users'

* Corresponding author. Tel.: +33 01 60 76 41 65.
*E-mail addresses:* han.xiao@telecom-sudparis.eu (X. Han), leye.wang@telecom-sudparis.eu (L. Wang), noel.crespi@telecom-sudparis.eu (N. Crespi), soochang.park@telecom-sudparis.eu (S. Park), acrumin@it.uc3m.es (Á. Cuevas).

http://dx.doi.org/10.1016/j.dss.2014.11.008
0167-9236/

interest similarity and diverse social features, and have unearthed additional relative factors that could enhance interest similarity prediction.

Particularly, we quantify interest similarity over an aggregation of user pairs by two metrics: *probability of sharing interest*, defined as the likelihood that two users have any mutual interests; and *degree of interest similarity*, which captures interest overlaps between two users based on the weighted cosine similarity. In addition, we extract social features (e.g. profile overlap, geographic distance, and friend similarity) from users' social information regarding three aspects: demographic information (age, gender, location, etc.), social relations (i.e., friendship), and obtainable users' interests. Specifically, we conduct the study in three interest domains, namely movies, music and TV shows, over a large dataset of 479,048 users and 5,263,351 user-generated interests crawled from Facebook.

We highlight our key findings captured from the wide variety of analysis — the homophily of interest similarity. Generally, homophily shows the level of homogeneity in people's social networks in relation to multiple sociodemographic, behavioral and intrapersonal characteristics [16]. Specifically, in this paper, homophily

- reveals that people tend to be interested in the same movies, music and TV shows when they are similar in their demographic information, such as age, gender and location;
- implies that friends have higher interest similarity than strangers. Furthermore, the interest similarity increases if two users share more common friends;
- indicates that the individuals with a larger interest entropy are likely to share more interests with others. Note that we exploit interest entropy to quantify the characteristics of one user's interests. A user's interest entropy is influenced by two factors: the total number of a user's interests and the popularity of these interests. The more interests a user presents, and the less popular the interests are, the more the user gains in interest entropy.

Based on the empirical studies, we propose a prediction model with a number of features (e.g. geographic distance, friend similarity and interest entropy). This prediction model can determine whether two users are similar or not in interest when one of the users does not provide his interests. The prediction result can be properly applied to various interest similarity based applications (e.g., recommendation system [3, 5], friend prediction [1, 10] and user evaluation system [4]). For instance, the model can help to address the *new user problem* in the typical collaborative recommendations. Normally, a collaborative recommendation system recommends a user some items that are liked by the others with similar interests. Whereas, the recommendation may fail when it comes to a *new user u* not revealing his interests, as the system cannot determine which of its existing users may share interests with *u*. In this case, even without *u*'s interests, the proposed prediction model is able to find some existing users who are predicted being similar to *u* and recommend *u* some items according to their interests.

In summary, the main contributions of this paper include:

- To the best of our knowledge, this is the first work to infer the interest similarity of two users where we do not know one of the user's interests. Owing to the frequent lack of users' interest in OSNs and the common requirement for applications of knowing the interest similarity between users, this research problem has a practical significance.
- We capture various social features depending on users' social information and investigate how interest similarity relates to these social features through a comprehensive perspective at a collective level. We uncover the homophily between these social features and users' interest similarity. Relying on a large dataset crawled from Facebook, the analytical results can advance the collective knowledge of OSNs.
- We devise a practical interest similarity prediction model based on the learned social features, namely *InterestSim* model. We also introduce two baselines referred to *Friend* model and *DemoSim* model. These two baselines depends on users' friendships [12, 29] and demographic similarity [7, 15, 20] respectively. The experiments show that *InterestSim* model outperforms *Friend* and *DemoSim* model by 12%–16% and 3%–4% respectively in terms of AUCs in different interest domains.
- We illustrate a use case where we leverage the proposed *InterestSim* model to practically address the *new user recommendation problem*. Compared with several state-of-the-art approaches, it turns out that our proposed *InterestSim* model can facilitate the *new user recommendation* with a higher precision.

## 2. Literature review

### 2.1. Studies on OSNs

Understanding social characters from large-scale OSNs is a hot research topic in recent years. Jure et al. conduct a comprehensive analysis on the MSN message network [13], and Alan et al. examine and compare four social networks (Flickr, YouTube, LiveJournal, Orkut) simultaneously [17]. These early studies mainly shed light on the high-level characteristics and verified many relationship based properties in OSNs, such as power law and small world [27]. Complementary to these studies on basic relationship social graph, some other work aims at users' interactions, such as posts, comments and mentions, and analyzes features on the user interaction graph [28, 30]. Different from the above work, we concentrate on a more specific question — how various social features would affect two users' interest similarity.

In fact, many ways are proposed to model users' similarity. Six similarity measurements are compared in [26] where the authors conclude that cosine distance performs the best for recommending online communities to users. Additionally, users' similarity can be measured by various information, such as profile similarity, connection similarity and interest similarity. Users' similarityis proved to be related to their friendship to some extent. This relation is usually leveraged to estimate the relationship strength between users [1, 31]. Relying on this relation, some other work infers users' missing profile properties, such as age [9] and school [18], via their social relations. In this work, we discuss the users' interest similarity.

Users' interests are normally desirable to know for many applications. When a user's interests cannot be obtained, it is common to infer his interests from the interests of other users who probably are similar to him. For instance, authors deduce a user's interests by considering this user's social neighbors' interests [29]. Also interests are proposed to be inferred from the users who share more demographic attributes [7, 15, 20]. Although [12] evaluates the interest similarity between pairs on CiteUlike and concluded that social connected users exhibit significantly higher interest similarity than the disconnected ones. Unfortunately, to our knowledge, how various social features relate to users' interest similarity has not been discussed in detail in any previous studies. This paper evaluates the interest similarity with multiple social features including demographic characteristics, friend relations as well as interest entropy.

Entropy is widely leveraged in the analysis of OSNs, besides demographic information and friendship. As a lower entropy generally implies a higher predictability, entropy is employed to study the mobility patterns and to infer the predictability of mobile phone users' behavior [21, 25]. Entropy is also used over users' interests and measures to what extent those users focus on topic categories [2, 11]. Our work tries to capture the patterns of users' interests by using interest entropy, where the initial intention is to investigate whether the interest entropy relates to the interest similarity. If the interest entropy does correlate to the interest similarity, then we can introduce it into the prediction as a social feature with other demographic and friendship features.

### 2.2. Applications of interest similarity

Much existing work either explicitly or implicitly leverage users' interest similarity into various research problems and applications, such as item recommendation system [3, 5], friend prediction [1, 10] and user evaluation system [4]. In order to recommend items to a given user, the collaborative recommendation systems require capturing users who are similar in interest to him [3, 5]. Friend prediction can also be improved based on the observation that users sharing more interests are more likely to be friends with each other [1, 10, 14]. Besides, interest similarity can affect the evaluations that one user provides to another (e.g., whether one user trusts another user's reviews on a product) [4].

The above-mentioned studies assume that both users' interests are known, then their interest similarity can be easily computed. However, they have a limitation when one of the users does not expose his interests, like the new user problem in recommendation system. For a new user, the current recommendation systems recommend items based on the interests of his friends or the users with similar demographic information, since the researchers indicate that two friends [12, 29] or two users who are similar in their demographic information [7, 15, 20] may have high interest similarity with each other.

In this paper, we set up an interest similarity prediction model assuming that one of the given two users does not expose his interests. First, our prediction model can be applied to many applications that require capturing similar users in interests for a given user but not knowing the given user's interests (e.g., item recommendation system [3, 5], user evaluation system [4]). Second, compared to the existing work, our prediction model is constructed according to comprehensive empirical studies and considers more social features (i.e., demographic features, friendship features and interest entropy).

## 3. Data description

In this work, we will study users' interest similarity based on real social network data from Facebook, as Facebook leaves open-ended spaces for users to present their interests in several domains such as movies, music, TV shows, and books. We crawled Facebook from March to June 2012 and collected profile data from 479,048 users, involving 5,263,351 user-generated interest items. To our knowledge, these data represent one of the largest and most comprehensive online social information databases to date. The analyzed data can be split into three parts:

- *User Interests:* Nine interest domains are collected: movies, music, TV shows, books, games, athletes, teams, sports, and activities. We find that 51.6 % of users do not publicly reveal any interests, while 41.0 %, 31.8 % and 28.3 % of users describe interests of movies, TV shows and music respectively — the top three interest domains with most users. Our focus in this paper is thus on music, movies and TV shows.
- *Demographic Information:* Refers to seven specific profile attributes[1] : age, gender, current city, hometown, high school, college and employer. We use these attributes to compute profile overlap between users so as to examine its influence on interest similarity; Besides, gender, current city, and age are further discussed separately. In the data set, 256, 163 users (53.5 %) report their gender; 173, 027 users (36.1 %) publish their current city; and only 14, 055 users (2.9 %) reveal their age.
- *Social Relationships:* We captured users' friend lists, thus here we define social relationship as user-claimed friendship. Note that friendship in Facebook is bidirectional, i.e., A is B's friend and B is a friend of A. In our dataset, 300, 204 (62.7 %) users make their social relationships public.

[1] In this paper, profile attribute is different to social feature. Profile attributes are the information that users claim on their Facebook page (e.g., age, hometown, gender); social feature indicates the quantitative values, like age distance, location distance, friend similarity and etc., which are derived from attributes;.

Note that we construct the dataset exclusively with users' public information and anonymize all the data during the analysis.

### 3.1. Characters of data set

We reveal some users' characters in our dataset. Among the 256, 163 gender reporters, 124, 677 of them are self-reported as female and 134, 486 are male; among all the age reporters, 4196 are male and 4096 are female.

Fig. 1(a) plots the number of users at each age. We observe that the numbers of users are skewed by age. The proportion of the users older than 40 or younger than 20 is rather small (less than 10 %). Therefore, in the age related studies, only the users whose age falls into the range of 20–40 years are taken into account. Moreover, we group the reporters in the age between 20 and 40 into generations by an interval of 3 years. Fig. 1(b) presents the average numbers of interests by generation. We notice that the young people report more interests than the elders.

Fig. 1(c) displays location distribution of current city reporters over the globe. We observe that people from North America and Europe are the dominant users on Facebook (indicated by the red dots). Fig. 1(d) illustrates the distribution of geographic distance and shows that the percentages of pairs fluctuate by distances with a gradual downward trend. The peaks and drops at some specific distances may reveal geographic characters. For instance, the peaks at distances of 5000 km and 6500 km may respectively indicate the width of America and the width of Atlantic.

## 4. Overview

We provide a brief overview to state the research problem, present an outline of a potential solution and introduce the empirical analysis framework, visualized in Fig. 2.

The goal of this paper is *to estimate the interest similarity between two users without knowing one user's interest information*. To achieve this goal, we first distinguish two kinds of users, *Active Users* and *Passive Users*:

- *Active users:* (i.e., $u_a$) explicitly present their demographic information (D), friendships (F) and interests (I), which can be denoted by a tuple of $u_a : < D_a, F_a, I_a >$;
- *Passive users:* (i.e., $u_p$) only report partial demographic information and/or friendships, but hide interests from the public; we denote a passive user as $u_p : < D_p, F_p >$.

On this basis, the fundamental problem becomes, given an active user $u_a$ and a passive user $u_p$, to infer whether $u_a$ and $u_p$ are similar or dissimilar in interest. The problem also could be extended to select a subset of active users who probably share many interests with $u_p$, given a $u_p$ and a set of active users (i.e., $C_{u_a} = \{u_a : <D_a, F_a, I_a>\}$).

Our solution for this problem is to train a prediction model which can infer the interest similarity between users relying on their obtainable social information. For instance, it might speculate that two users are more likely to share interests if they are friends. Consequently, we attempt to achieve the interest similarity prediction by two steps: (1) based on users' social information, we can capture several social features that may reflect users' interest similarity to some extent; and (2) based on the learned social features, we construct an interest similarity prediction model.

According to the proposed solution, the primary issue is to determine what specific social features correlate to the users' interest similarity. Therefore, we conduct extensive empirical analysis on interest similarity with respect to various social features derived from the

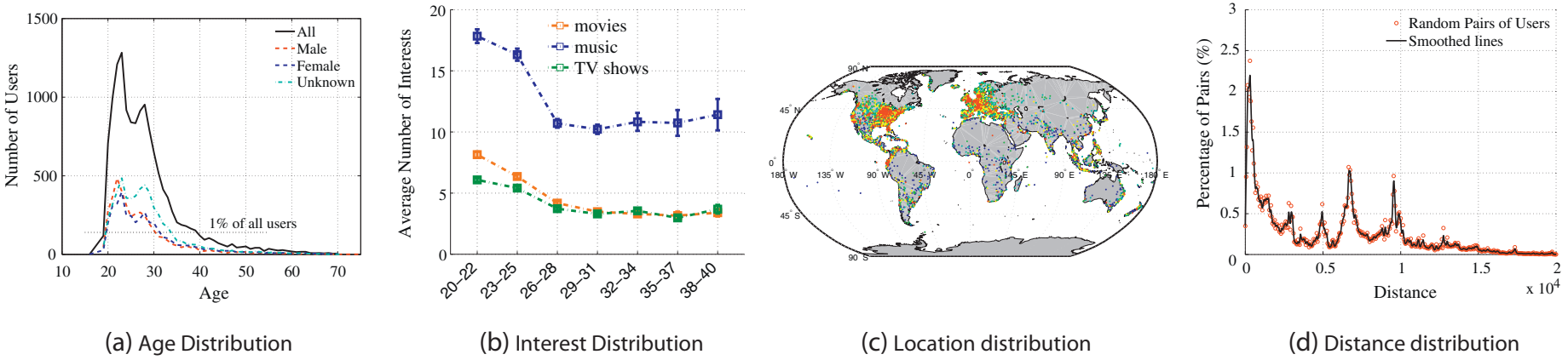


**Fig. 1.** Users' characters. The color of each dot in Fig. 1(c) corresponds to the number of users in a city, applying a spectrum of colors ranging from blue (low), green, yellow to red (high).

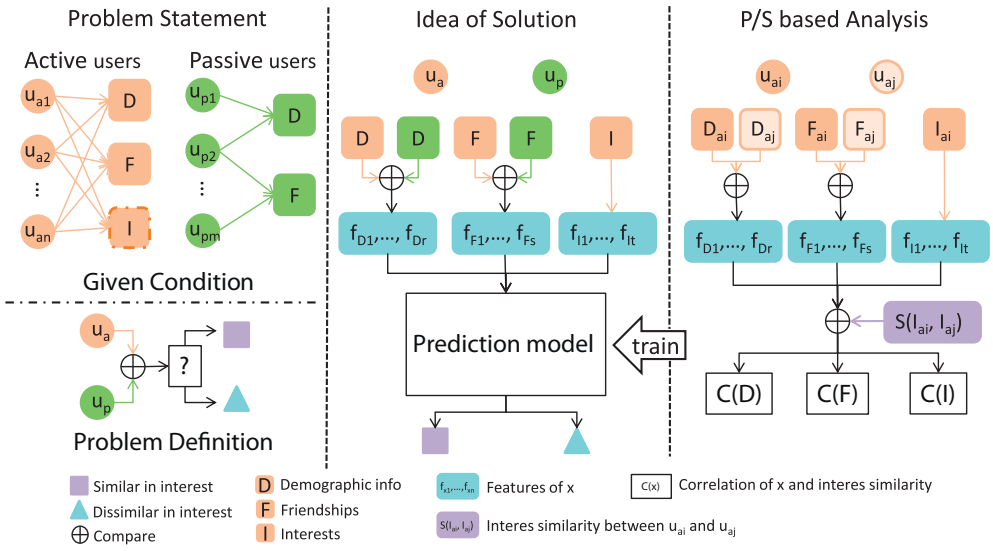


**Fig. 2.** Overview of problem, proposed solution and the problem and solution (P/S) based analysis framework.

users' social information. In particular, we perform the analysis through three perspectives:

- *Demographic-related features ($f_D$):* We extract the demographic features by comparing two users' demographic information (D) and investigate how they correlate to interest similarity. For example, we measure the geographic distance between users and examine how users' interest similarity varies regarding their geographic distance.
- *Friendship-related features ($f_F$):* We generate friendship features based on the friendships (F) of two users. For example, we define a feature of friend similarity by counting the mutual friends of two users and study its influence on interest similarity.
- *Interest-related feature ($f_I$):* Since we do not know the passive user's interests in the prediction problem, we tend to explore interest-related feature by capturing the interest characteristics from the active user side ($I$). We expect that the users who exhibit certain characteristics on his interests would generally achieve a higher/lower interest similarity with others. In this paper, we specially employ entropy to quantify a user's interests as the interest-related feature.

Furthermore, based on the learned social features, we exploit Support Vector Machines (SVM) [6, 8] method to train the interest similarity prediction model.

## 5. Measurements for interest similarity

To study the properties of interest similarity among users, we define the measurement of interest similarity by two steps: (1) we first limit the computation of interest similarity between two users (i.e., a user pair); (2) we extend the computation to an aggregation of user pairs and obtain a measurement of collective interest similarity. The analysis regarding interest similarity in the following sections depends on the collective interest similarity. Consequently, we first introduce two ways to measure interest similarity between two users: **binary similarity** and **weighted cosine similarity**. Then, based on these two measurements, we define two metrics to evaluate interest similarity at an aggregated level, namely the **probability of sharing interests** and the **degree of interest similarity**.

### 5.1. Interest similarity of two users

Binary similarity and weighted cosine similarity are the two measurements used to calculate interest similarity between two users. Note that user $u$'s interests are denoted by an interest set $I_u$ instead of a binary interest vector to avoid a very sparse interest vector.

**Binary similarity** measures whether or not two users are similar in terms of their interests. We assume that two users are similar in interest, as long as they have any mutual interests; otherwise, they are dissimilar, denoted as:

$$s_b(u,v) = \begin{cases} 1 & \text{if } I_{uv} \neq \varnothing \\ 0 & \text{if } I_{uv} = \varnothing \end{cases} \quad (1)$$

where $I_{uv}$ represents the intersection of interests between user $u$ and $v$. Binary similarity is defined to evaluate the probability of sharing interests.

**Weighted cosine similarity** estimates the extent to which two users are similar in interest. It is introduced by two steps. First, drawing on the general calculation of cosine similarity, the interest similarity between users $u$ and $v$ is then defined as the cosine distance between their interest sets: $s_c(u,v) = \frac{\|I_{uv}\|_1}{\|I_u\|_2 \cdot \|I_v\|_2}$ where $\|I_u\|_2 = \sqrt{l_u}$ ($l_u$ is the number of interests of $u$) and $\|I_{uv}\|_1$ is the number of mutual interests of $u$ and $v$. If either $l_u = 0$ or $l_v = 0$, $s_c(u,v)$ is undefined.

Moreover, as it seems easier for two users to share a very popular interest (e.g., the movie 'Harry Potter') than a rare one (e.g., the documentary 'La Dany'), we consider the interest similarity to be more significant if two users share a less popular interest. So, we introduce interest popularity into the calculation of cosine similarity. Specifically, we count the number of users who like an interest as its popularity and weight the cosine similarity according to the popularity of two users' mutual interests. The more an interest occurs, the less weight it is assigned. Thus we formulate the weighted cosine interest similarity as:

$$s_w(u,v) = \frac{\sum_{i \in I_{uv}} w(i)}{\|I_u\|_2 \cdot \|I_v\|_2} \quad (2)$$

in which $w(i)$ equals the inverse log $N$ where N stands for the number of users who are interested in interest $i$, i.e., $w(i) = \frac{1}{\log N}$. Weighted cosine similarity is applied to compute the degree of interest similarity.

### 5.2. Collective interest similarity

Based on the above-introduced interest similarity metrics regarding two users, we further estimate the collective interest similarity over an aggregation of user pairs. We denote the aggregation of user pairs as $C$

and average the interest similarities of the user pairs in $C$ as its collective interest similarity.

In particular, we define **probability of sharing interests** (i.e., $p$) of user pairs in $C$ as the mean binary similarity of the collective pairs as follows:

$$p = \frac{\sum_{(u,v)\in C} s_b(u,v)}{\|C\|}. \quad (3)$$

In addition, we calculate the **degree of interest similarity** (i.e., $s$) of $C$ as the average weighted cosine similarity of all the user pairs in $C$, denoted as:

$$s = \frac{\sum_{(u,v)\in C} s_w(u,v)}{\|C\|} \quad (4)$$

where $\|C\|$ stands for the number of pairs that are included in the pair set $C$. In the rest of this paper, we use these two collective measurements to study how interest similarity varies depending on various social features.

## 6. Homophily of interest similarity

In this section, we examine the relations between interest similarity and various social features that emerge from the collective users. We investigate the changes of interest similarity with respect to demographic-related features, social relationships and interest-related feature subsequently.

Note that, each empirical study is carried out on a specific social feature and a particular interest domain (i.e., movies, music and TV shows). Therefore, for each study, the pair set $C$ is generated by considering two factors: (1) the related profile attribute and (2) the focused interest domain. For instance, to test the relation between gender and interest similarity in terms of movies, we construct a gender/movie set of pairs by coupling users who present both gender and movies. Note that we only consider the users who exhibit more than three items in the focused interest domain.

### 6.1. Interest similarity by demographics

We study how demographic information affects interest similarity from four perspectives, profile overlap, gender, location (geographic distance and country) and age (age distance and generation).

#### 6.1.1. Interest similarity by profile overlap

Profile overlap measures the number of the profile attributes where two users exhibit the same value. In particular, for each user, we generate a profile vector with 16 cells which corresponds to nine interest domains and seven demographic attributes (refer Section 3). Concerning a particular interest domain cell, if a user $u$ presents any items in the interest domain, we say $u$ is interested in this domain and denote the cell as 1; otherwise, it is set to 0. We directly put the users' demographic attributes into the corresponding cells.

We separately generate profile/interest sets for the three interest domains (i.e., movies, music and TV shows) with 1, 000, 000 user pairs where the users present more than three interest items and at least one demographic attribute. Let $C_{d_q} = \{(q_u, q_v, p_{uv}, s_{uv}) : |q_u \cap q_v| = d_q\}$ denote a collection of user pairs where the profile overlap between the user pair ($u$ and $v$) is $d_q$; $q_u$ and $q_v$ represent $u$ and $v'$ profile vectors; $p_{uv}$ and $s_{uv}$ are the probability of sharing interest and the degree of interest similarity between $u$ and $v$ respectively.

Fig. 3 plots the interest similarity over profile overlap in movies, music, and TV shows respectively. As the number of user pairs with profile overlap beyond 11 is very small, we concentrate on the user pairs whose profile overlap falls between 1 and 11. The results reveal that both of the probability of sharing interests and the degree of interest similarity go up with the increase of profile overlap regardless of interest domains. This observation demonstrates that two users are more similar in their tastes if they share more common attributes in their profiles.

#### 6.1.2. Interest similarity by gender

We produce gender/interest sets with 1,000,000 randomly coupled user pairs where the users present their gender and more than three interest items (movies, music or TV shows). Let $C_{g_c} = \{(g_u, g_v, p_{uv}, s_{uv}) : g_u \cup g_v = g_c\}$ denote an aggregation of user pairs where two users are of gender combination $g_c$. Here, the gender combination of a user pair takes three possible values (i.e., $g_c$) as male–male, female–female and male–female.

Table 1 shows the probability of sharing interests and the degree of interest similarity according to the different gender combinations. We observe the homophily for gender that the pairs present higher interest similarities when they are in the same sex (i.e., male–male or female–female). In addition, we find that males are more similar on the interests of movies and music whereas females present higher interest similarity in TV shows.

This observation of homophily for gender here is different from the heterophily for gender in communication network reported in the previous work [13]. It demonstrates that people communicate more with the ones in the opposite gender. In other words, although people like to make connection with others of different sex, the pairs of cross-gender do not share interests highly. This suggests that we should exploit the gender property of the homophily or heterophily properly according to the specific applications. For instances, for some specific communication/dating applications, users in the opposite gender might take the priority to be considered; while the users of the same gender are supposed to be thought at the first place when it comes to enhancing the recommendation for interests.

#### 6.1.3. Interest similarity by location

We study how location affects interest similarity by geographic distance and country.

**Interest similarity by geographic distance**: denote a set of user pairsb where the two users of a pair are apart of $d_{uv}$ in the span of $[d_l, d_l + \nabla)$ by $C_{d_l} = \{(l_u, l_v, p_{uv}, s_{uv}) : distance(l_u, l_v) = d_{uv} \wedge d_{uv} \in [d_l, d_l + \nabla)\}$. $l_u$ is the location of user $u$ represented by its latitude and longitude and $\nabla$ stands for an interval of distance.

Fig. 4(a) reports the degree of interest similarity by a full view of distance range from 0 to 15,000 km with an interval of 100 km. Although the results fluctuate at some points when the distances are larger than 3000 km, we see a decreasing trend of the degree of interest similarity by the distance. Furthermore, we zoom in the x-axes and show the interest similarity with distances in the range of 0 and 3000 km in Fig. 4(b), (c) and (d). We observe that the interest similarity decreases quickly when the distance is small, and it gets steadily when the distance continuous increasing. This implies that the interest similarity correlates to the distance very sensitively only in a limited range of distance.

In addition, we look into a number of pair samples which might lead to the fluctuations at distances larger than 3000 km. Taking the peak at 3500 km as an example, we find that the two users at this distance are mostly from the east and west of the USA. Therefore, we speculate that such peaks may reveal some implicit connections (e.g., nationality, language, culture) between the specific geographic regions. Therefore, we further examine how interest similarity varies depending on the geographic region in terms of country.

**Interest similarity by country**: let $C_{t_{hk}} = \{(t_u, t_v, p_{uv}, s_{uv}) : t_u = h \wedge t_v = k\}$ denote the set of pairs in which the two users come from the countries (denoted by $t_u$ and $t_v$) of $h$ and $k$. We select users from 20 representative countries over six continents and randomly generate

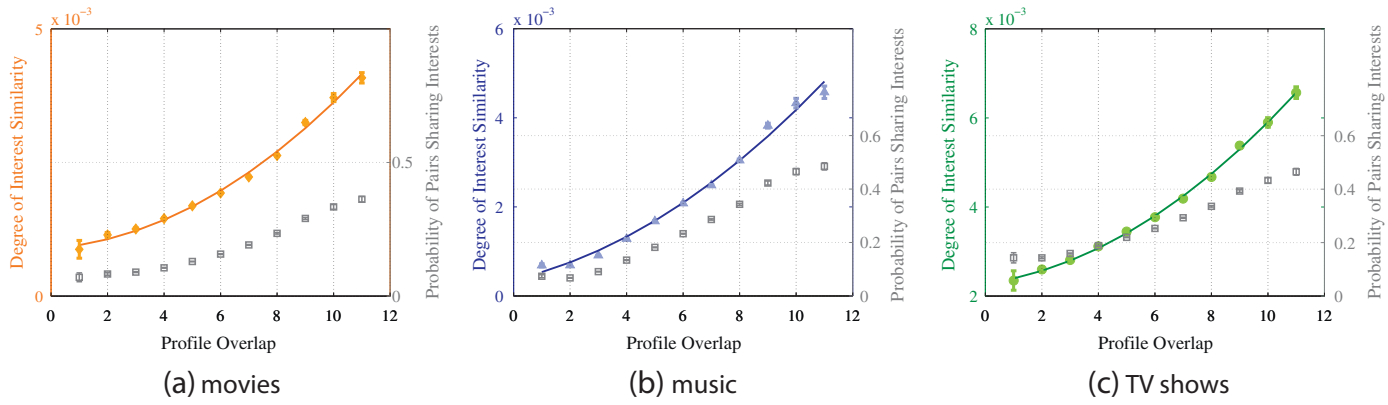


**Fig. 3.** Interest similarity with profile overlap. Standard error is estimated by bootstrap re-sampling throughout this paper. The colorful left Y-axes stand for the degree of interest similarity and the right gray Y-axes indicate the probability of sharing interests.

200, 000 pairs for each country combination (cross-country or same-country).

Fig. 5 displays the heatmaps of the degree of interest similarity by country combination, where a brighter cell indicate that users from the corresponding countries (represented by the row and column) share more interests. Note that the cells on the secondary diagonal represents the interest similarity of pairs from the same country (i.e., native pairs).

We observe that the cells on the secondary diagonal is brighter than the other cells in the same row or column. This demonstrates that, compared to the pairs from two diverse nations (i.e., alien pairs), native pairs share more interests. Besides, we notice Chinese share less movies with Philippine and Indonesian, but report a high movie similarity with American. We also notice that users from South America countries share a lot of interest. This observation might imply that the different countries share interests with distinctions.

#### 6.1.4. Interest similarity by age

How age distance and generation affect interest similarity are learned in this section.

**Interest similarity by age distance**: age distance measures the gap of two users in terms of age. Let $C_{d_a} = \{(a_u, a_v, p_{uv}, s_{uv}) : |a_u - a_v| = d_a\}$ denote a set of pairs whose ages differ at $d_a$. Note that the discussed age distance (i.e., $d_a$) varies from 0 to 20 years.

Fig. 6 shows that the interest similarity decline as the age distance goes up. This observation demonstrates that users share more interests if they are closer at age. Moreover, we observe that the interest similarity drops fast when the age distance is small; and it gets to decline gradually as the age distance continues increasing.

**Interest similarity by generation**: Let $C_{g_a} = \{(a_u, a_v, p_{uv}, s_{uv}) : a_u \in g \wedge a_v \in g\}$ denote a set of user pairs where the two users are in the same generation $g$. Remind that we select 3 years as an age interval of one generation.

Fig. 7 reveals that the younger generations present higher interest similarity than the middle-age generations. And comparing the interest similarity by age distance inside a generation, the results basically hold the rule that the interest similarity decrease with the increase of the age distance although several exceptions exist (e.g., 38–40 for movie).

**Table 1**
Interest similarity by gender.

| | Probability of sharing interests | | | Degree of interest similarity | | |
|---|---|---|---|---|---|---|
| | Movies | Music | TV shows | Movies | Music | TV shows |
| Male & male | 0.164 | 0.179 | 0.209 | 0.0022 | 0.0019 | 0.0035 |
| Female & female | 0.145 | 0.157 | 0.245 | 0.0020 | 0.0015 | 0.0042 |
| Female & male | 0.118 | 0.151 | 0.176 | 0.0015 | 0.0014 | 0.0027 |

### 6.2. Effects of friendship

We examine interest similarity according to friendship through two perspectives: friend distance and friend similarity. Friend distance is computed by the connected hops between two users; friend similarity measures the common friends of two users.

#### 6.2.1. Interest similarity by friend distance

Let $C_{d_f} = \{(f_u, f_v, p_{uv}, s_{uv}) : D(f_u, f_v) = d_f\}$ denote a set of pairs where the friend distance of the two users $u$ and $v$ is $d_f$ hops. Particularly, we take into account friendship in two-hop with three users pair groups: *direct-friend* pair — $u$ and $v$ connect to each other directly ($d_f = 1$); *indirect-friend* pair — $u$ is a friend of $v$'s friends but $u$ and $v$ are not direct-friend ($d_f = 2$); *stranger* pair — $u$ and $v$'s friend distance is larger than 2 ($d_f > 2$).

Fig. 8(a) and (b) report the probability of sharing interests and degree of interest similarity by friend distance respectively. These results reveal that the users with less friend distance share more interests: *direct-friend* pairs exhibit the highest interest similarity; and the *indirect-friend* pairs share more interests than the *stranger* pairs do.

#### 6.2.2. Interest similarity by friend similarity

Friend similarity measures two users' common friends by cosine similarity, i.e., $f_{uv} = \frac{\|f_u f_v\|}{\|f_u\|\|f_v\|}$. Note that we only consider the user pairs who present at least one mutual friend where 95% of them show a friend similarity less than 0.02. So the studied friend similarity is in the range of (0, 0.02]. Let $C_{s_f} = \left\{(f_u, f_v, p_{uv}, s_{uv}) : \frac{\|f_u f_v\|}{\|f_u\|\|f_v\|} = f_{uv} \wedge f_{uv} \in [f_s, f_s + \nabla)\right\}$ denote a set of user pairs in which the two users exhibit a friend similarity in the range of $[f_s, f_s + \nabla)$. $\nabla$ represents an interval of friend similarity.

Fig. 8(c) shows the change of the probability of sharing interests with friend similarity; Fig. 8(d), (e) and (f) display the relation between the degree of interest similarity and friend similarity with respect to movies, music and TV shows respectively. All these figures reveal that the user pairs generally share more interests if they obtain a higher friend similarity. In particular, we observe that the interest similarity goes up steeply when the friend similarity is less than 0.001, and hereafter it becomes steady with rise of friend similarity.

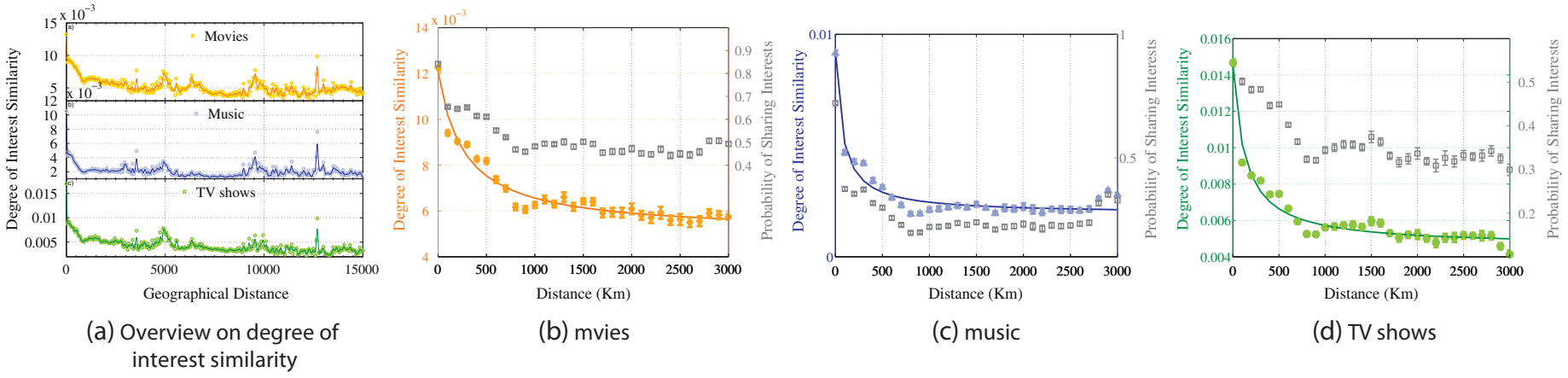


**Fig. 4.** Interest similarity by geographical distance.

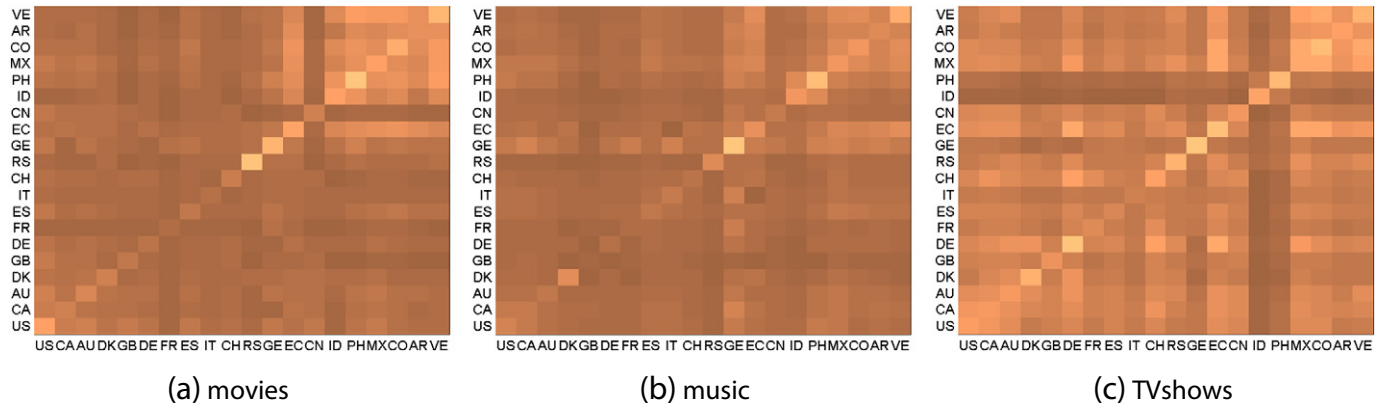


**Fig. 5.** Degree of interest similarity by country.

### 6.3. Effects of interest entropy

In this section, we are interested in looking at interest related feature. We employ entropy to capture a user's interest feature. Entropy quantifies the information amount of the user's interests by two elements of the interests: the number of interests and the weight of interests. Generally speaking, a user with many high weighted interests should be assigned with a large entropy. Using the natural log, we define interest entropy $H(I_u)$ as:

$$H(I_u) = -\sum_{x_i \in I_u} w(x_i) log\, w(x_i)$$

where $w(x_i)$ represents the weight of interests $x_i$ (defined in Section 5). As 95% of users' interest entropy is less than 8, we discuss the interest similarity by entropy in [0, 8].

Let $C_{e_i} = \{(I_u, I_v, p_{uv}, s_{uv}) : H(I_u) = e_i \vee H(I_v) = e_i\}$ denote a set of pairs by users' interest entropy of $e_i$. Note that, in this set, only one user in a user pair is required to have an interest entropy of $e_i$ because we tend to study whether the interest similarity would be influenced by one user' interest entropy in a pair.

Fig. 9(a) displays the probability of sharing interest; Fig. 9(b), (c) and (d) show degree of interest similarity. We observe that the interest similarity grow as the increase of interest entropy. And it particularly rises very quick as the interest entropy is small.

## 7. Inferring interest similarity

In the previous section, we conducted extensive analysis of how various social features correlate to interest similarity of two users. The goal of this section is to design a prediction model for inferring whether two users are similar in interest (namely interest similarity between users) relying on these new learned correlations.

Let us consider many applications which directly exploit interest similarity between users to improve the performance [1, 4, 10]. Obviously, the interest similarity can be easily computed if both of two users' interests are known. However, as there are always some users not revealing their interests, for such applications, missing users' interests is indeed a practical obstacle to computing interest similarity directly (e.g., new user problem in recommendation system [7, 12, 15, 20, 29]). Therefore, it is appealing to infer two users' interest similarity for this case.

Besides, users' interests are normally desirable for personalized recommending or advertising [3, 5]. For a number of passive users who do not explicitly reveal their interests (51% of users in our Facebook data set), if it is possible to capture some active users who not only expose their own interests but also are predicted to have similar interests as a given passive user, then we can infer the passive user's interests according to the similar active users' interests. In this case, how to predict users' interest similarity (i.e., to determine whether two users are similar or not in their interests) without knowing interests from one of the users becomes a meaningful problem.

Specifically, in this prediction, we consider two users: a passive user $u$ who only presents some demographic information and social relationships with limited friends but does not reveal his interests (i.e., $u_p : \langle D_p, F_p \rangle$); and an active user $v$ who has complete information including demographic attributes, friends as well as interests (i.e., $u_a : \langle D_a, F_a, I_a \rangle$). Then, *the prediction task is to determine whether the passive user u and the active user v are similar or dissimilar regarding their interests.*

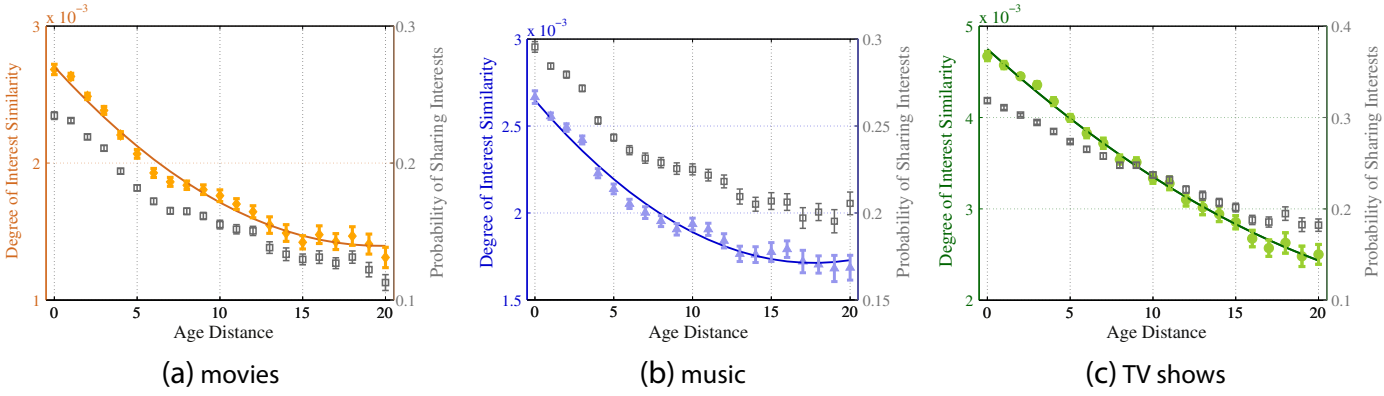


**Fig. 6.** Interest similarity by age distance.

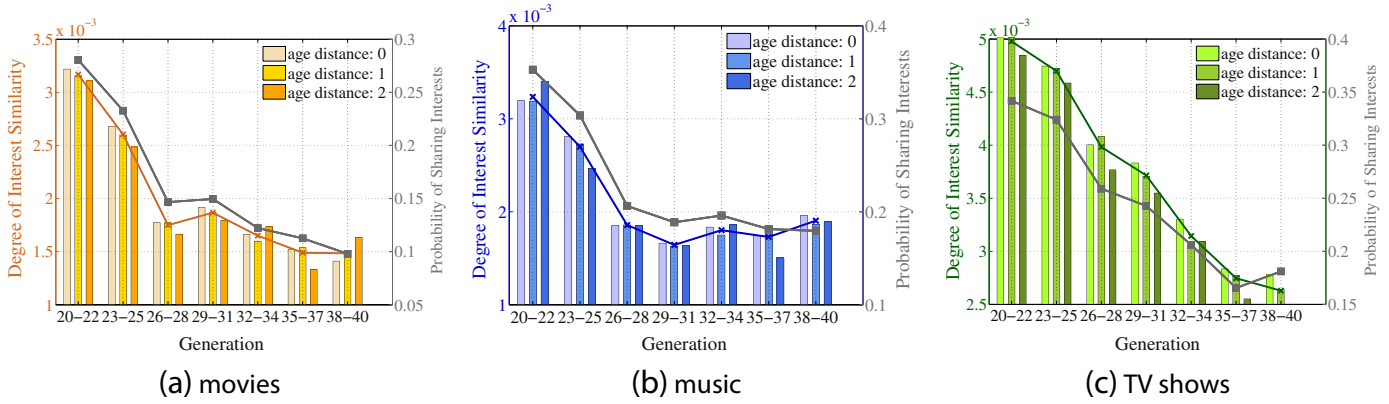


**Fig. 7.** Interest similarity by generation. The lines represent the interest similarity of each generation. Inside each generation, the grouped three histograms display the degree of interest similarity with age distance at 0, 1 and 2 respectively.

### 7.1. Interest similarity prediction model

According to the prediction task itself, two possible results are expected: *i*) the given passive user $u$ and active user $v$ are similar regarding their interests (i.e., labeled as *interest-similar*); *ii*) $u$ and $v$ are not similar (i.e., labeled as *interest-dissimilar*). To achieve the task, the basic idea is to train a prediction model to label $u$ and $v$ as either *interest-similar* or *interest-dissimilar* by learning their *social features*. Therefore, in this section, we introduce our prediction model in details from three aspects: (1) we clarify the criterion to determine whether two users are *interest-similar* or *interest-dissimilar*; (2) we illustrate the social features that are leveraged to train the prediction model; (3) by exploiting Support Vector Machines (SVM) method [8], we establish our interest similarity prediction model, namely *InterestSim* model.

**Criterion**: Given a pair of users $u$ and $v$, whether they are similar or dissimilar is determined by their interest similarity and an established threshold. We compute $u$ and $v$'s interest similarity by the degree of interest similarity (i.e., $s_w(u, v)$) and compare the value to the established threshold (i.e., $\varepsilon$). We use $z_{uv}$ to label the interest similarity between $u$ and $v$. If the interest similarity is larger than $\varepsilon$, $z_{uv}$ is labeled to 1, representing

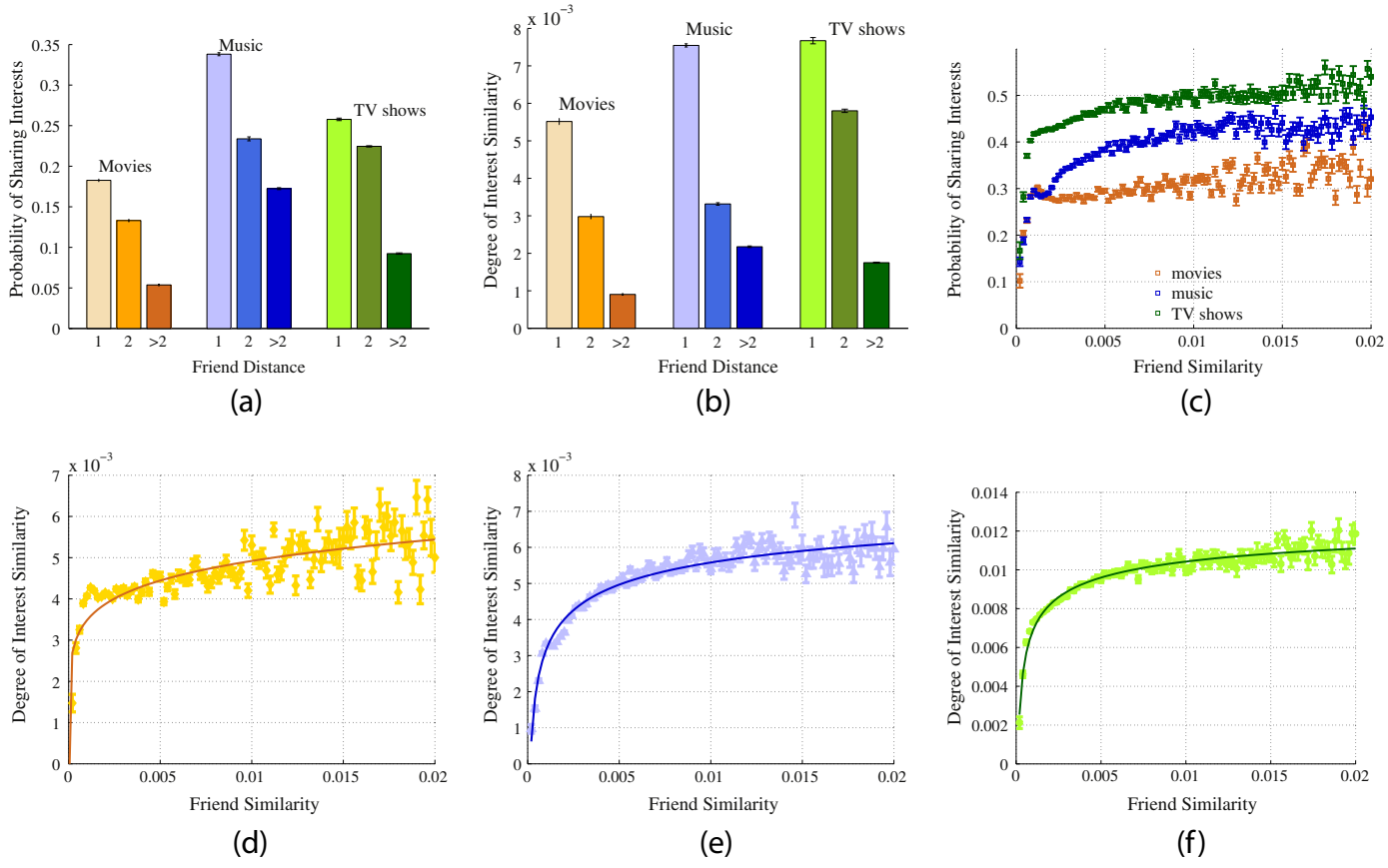


**Fig. 8.** Effects of friendship. Fig. 8(a) and (b) plot the interest similarity by friend distance; Fig. 8(c), (d), (e) and (f) display the results by friend similarity.

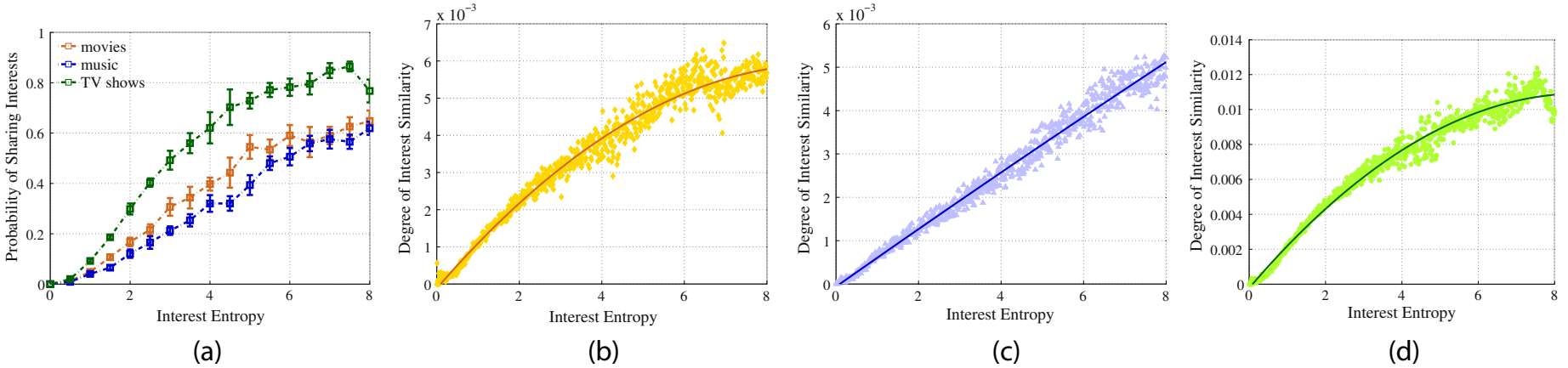


**Fig. 9.** Interest similarity by interest entropy.

$u$ and $v$ are *interest-similar*; otherwise, $z_{uv}$ is labeled to $-1$, indicating $u$ and $v$ are *interest-dissimilar*:

$$z_{uv} = \begin{cases} 1 & s_w(u,v) \geqslant \varepsilon \\ -1 & s_w(u,v) < \varepsilon \end{cases}. \quad (5)$$

### 7.2. Social features

Moreover, given a passive user $u$, an active user $v$ and all of their obtainable social information (i.e., demographic information, friends and $v$'s interests), we extract the following social features drawing on the studies in the previous section:

- *Profile Overlap* ($PO_{uv}$) computes the percentage of the same attributes that $u$ and $v$ share among the seven demographic attributes: age, gender, current city, hometown, high school, employer, and college.
- *Gender Combination* ($GC_{uv}$) takes three possibilities: 1 (male–male), −1 (female–female), and 0 (male–female).
- *Geographic Distance* ($GD_{uv}$) measures the distance between $u$ and $v$'s current city (refer to Section 6.1.3).
- *Binary Country* ($BC_{uv}$) is set to 1 if $u$ and $v$ come from the same country; otherwise it equals 0.
- *Age Distance* ($AD_{uv}$) calculates the absolute difference of $u$ and $v$'s ages.
- *Friendship Distance* ($FD_{uv}$) is set to 1 if two users are friends; otherwise, it equals 0.
- *Friend Similarity* ($FS_{uv}$) is calculated by cosine similarity (refer to Section 6.2.2).
- *Interest Entropy* ($IE_v$) is computed by the active user $v$'s interests (refer to Section 4 and 6.3).

Note that we normalize *Geographic Distance*, *Age Distance*, *Friendship Distance* and *Interest Entropy* to ensure all the features belonging to [−1,1]. Thus, for the user pair $u$ and $v$, we obtain a social feature vector: $\mathbf{x}_{uv} = \langle PO_{uv}, GC_{uv}, GD_{uv}, BC_{uv}, AD_{uv}, FD_{uv}, FS_{uv}, IE_v \rangle$.

### 7.3. SVM-based InterestSim model

So far, from each user pair $(u, v)$ where $u$ is a passive user and $v$ is an active user, we can generate a tuple $\langle \mathbf{x}_{uv}, z_{uv} \rangle$. $\mathbf{x}_{uv}$ is the social features extracted from $u$ and $v$'s social information; $z_{uv}$ is the label which stands for whether $u$ and $v$ are *interest-similar* or *interest-dissimilar*. To train the *InterestSim* model, we aggregate a number of user pairs where all the pairs are made of a passive user and an active user. Similarly, from all these user pairs, we can generate a tuple collection where each tuple corresponds to a pair of users, denoted as $\mathcal{C}\{pair_i : (\mathbf{x}_i, z_i)\}$. Assume $q$ stands for the total number of the user pairs and $i$ denote the $i$ th pair. Then constructing the SVM-based prediction model is solving the following optimization problem:

$$\begin{aligned} & minL(w) = \frac{1}{2}\|w\|^2 + \delta \textstyle\sum_{i=1}^{q} \xi_i \\ & \text{subject to}: \begin{cases} \xi_i \geqslant 0 \\ z_i \langle w, \mathbf{x}_i \rangle \geqslant 1 - \xi_i \end{cases} \end{aligned} \quad (6)$$

where $\delta$ is a constant and $\xi_i$, $(i = 1,..., q)$ are slack variables for optimization. Note that, for training the prediction model, we assume that $u$'s interests are known to calculate $u$ and $v$'s interest similarity so as to determine the label (*interest-similar* or *interest-dissimilar*). However, when computing the social features, we think of $u$'s interests as unavailable information in keeping with the prediction problem's pre-condition that $u$ is a passive user.

Specifically, to train the proposed *InterestSim* model, we generate 150,000 user pairs by randomly coupling two users ($u$ and $v$) where both $u$ and $v$ exhibit all the demographic information, friend lists as well as more than three interests in movies, music, or TV shows. Afterward, we split the whole 150,000 user pairs into ten subsets (i.e., 15,000 user pairs per subset) and do a ten-fold cross validation.

### 7.4. Evaluation of prediction

In this section, we are going to evaluate the *InterestSim* model through two ways: (1) we leverage the 'leave-one-feature-out' approach to investigate the effects of various social features on the interest similarity predictions; (2) we evaluate the performance of *InterestSim* model and compare it with other two baseline approaches.

#### 7.4.1. Leave-one-feature-out evaluation

We carry out 'leave-one-feature-out' comparisons and train prediction models by excluding one of overall features. For instance, we train a *No Profile Overlap* model by taking out *Profile Overlap* from the social feature vector $\mathbf{x}_{uv}$. In addition, for some features originated from one attribute, we remove them as one integrated feature to train the 'leave-one feature-out' model. For example, we view Friendship Distance and Friend Similarity (both originated from friend lists) as an integrated feature, namely *Social Relation*; and also regard *Geographic Distance* and *Binary Country* as *Location*. In particular, we generated models without any one out of the six features of *Profile Overlap*, *Gender Combination*, *Age Distance*, *Location*, *Social Relation*, and *Interest Entropy*. In total, we obtain 18 'leave-one-feature-out' models with respect to the three interest domains of movies, music and TV shows ($6 \times 3$).

Table 2 compares the 'leave-one-feature-out'models with the *InterestSim* model in terms of the areas under ROC curves (AUCs). From the table, we can see that our proposed *InterestSim* model, which infers interest similarity according to all the learned social features, outperforms the other models which miss one type of social features. It demonstrates that all the used social features are beneficial for the prediction. Note that a social feature (e.g. *Gender Combination*) would be more important if the AUC of a model trained without the feature (e.g., *No Gender Combination* model) is smaller. Therefore, from the results, we can say that *Profile Overlap*, *Gender Combination* and *Social Relation* are less sensitive in the predictions of interest similarity compared to the other attributes, such as *Interest Entropy*, *Age Distance*, and *Location*. In addition, we observe that the impacts of the social features on the predictions in different interest domains exhibit their own properties. For instance, *Location* is more sensitive to music similarity prediction than movie similarity prediction, while *Social Relation* plays a more important role in movie similarity prediction than music similarity prediction.

#### 7.4.2. Prediction performance comparison

To the best of our knowledge, this is the first work aiming at inferring whether two users are similar or not in terms of their interests, without knowing one user's interests. Some existing work has pointed out several good features that can indicate similar interests between users. The friendship between two users is one of the most acknowledged feature that are used to infer a user's interests from the other's [12, 24, 29]. Additionally, in order to make accurate recommendations for new users without rating any items, demographic information is

**Table 2**
Comparison of effects on interest similarity prediction by different social features.

| | AUC | | |
|---|---|---|---|
| Type of model | Music | Movies | TV shows |
| *No profile overlap* | 0.6201 | 0.6388 | 0.6825 |
| *No gender combination* | 0.6521 | 0.6410 | 0.6889 |
| *No age distance* | 0.5831 | 0.5943 | 0.6061 |
| *No location* | 0.5490 | 0.5880 | 0.6550 |
| *No social relation* | 0.6491 | 0.6206 | 0.6727 |
| *No interest entropy* | 0.5171 | 0.5236 | 0.6047 |
| *Interestsim* model | 0.6720 | 0.6644 | 0.7027 |

also explored to indicate that users with more common demographic information might share more interests [7, 15, 20]. Therefore, we draw on their main ideas on interest similarity indications and train two baseline prediction models respectively exploiting users' friendships and demographic information, namely *Friend* model and *DemoSim* model. In particular, we train *Friend* model by using two features: *Friend Distance* and *Friend Similarity*; and we construct the *DemoSim* model by applying *Profile Overlap*, *Age Distance*, *Gender Combination* and *Geographic Distance*.

Fig. 10 plots the ROC curves for the three interest domains of movies, music, and TV shows, comparing the proposed *InterestSim* model to the *Friend* model and *DemoSim* model in the aspect of prediction capacity. Table 3 compares AUCs between the three sets of models. The ROC curves of *Friend* model almost approach to the secondary diagonal which represents the capability of random prediction. It indicates that we can hardly infer users' interest similarity merely with respect to their friendships. By considering four demographic features which involves in seven profile attributes, *DemoSim* model generates larger AUCs and performs better than *Friend* model. Even though, much of the area improvement under the ROC curves of *InterestSim* model has been shown in Fig. 10. From Table 3, for movies, music and TV shows, we gain more than 3%–4% of improvement compared with *DemoSim* in terms of AUC.

**Table 3**
AUC comparisons among *Friend* model, *Demo* model and *InterestSim* model.

| | Friend | Demo | InterestSim |
|---|---|---|---|
| Music | 0.5487 | 0.6411 | 0.6720 |
| Movies | 0.5335 | 0.6142 | 0.6644 |
| TV shows | 0.5478 | 0.6593 | 0.7027 |

## 8. Case study: recommendation for new users

Recommendation system recommends items to a user if these items are presumably preferred by the user. In order to make efficient recommendations, many existing approaches, which are categorized as *content-based recommendations*, *collaborative recommendations* and *hybrid recommendations*, need to acquire the users' interests. These approaches encounter a common and difficult problem — *new user problem* — when the recommendations are required for the new users who have no or very little information about their interests [3, 23]. Fortunately, our proposed *InterestSim* model just can make a bridge between the new users and their interests via some existing active users who present interests in the recommendation system: we can recommend the interests of the existing active users who are predicted being similar in interest with the new users. For this reason, we leverage our proposed *InterestSim* model to address the *new user problem*. With this case study, we aim at demonstrating the practical use of our proposed prediction model.

### 8.1. Approaches

In this subsection, we briefly describe how to recommend items to a new user based on our proposed *InterestSim* model — namely *InterestSimPop* recommendation; we also introduce several state-of-the-art new user recommendation approaches to compare with:

- *InterestSimPop*: exploits *InterestSim* model to infer a number of users who are similar with the new user in interest; and then it recommends the new user the most popular items that liked by those similar users;
- *OverallPop*: For a new user without claiming his interests, a straightforward way is recommending the overall most popular items among all the existing users. Such a method, called *OverallPop* here, is often used as an intuitive baseline in the existing research about the new user problem [22];
- *FriendPop*: In [12, 29], the authors indicate that using the friends' interests may facilitate the recommendation performance for a new user. We thus borrow the basic idea from these works to implement the *FriendPop* baseline method, which selects the most popular items among a new user's friends;
- *DemoSimPop*: Demographic information, such as age, location, gender, is another useful source to tackle the new user problem [7, 15, 20]. Following the idea in [7], *DemoSimPop* first finds the users whose demographic attributes (e.g., gender, location, and age) are similar to the new user, and then selects the most popular items from those demographic-similar users;
- *DemoComAgree*: Based on $\alpha$-community spaces model and 'level of agreement' of the community, the authors propose another way to use demographic information to improve the item recommendation for a new user [19]. Here, we also implement this method and call it as *DemoComAgree*.

### 8.2. Experiment setup and results

According to our data set, we randomly select 200 users who present demographic information (including age, gender, current city, hometown, high school, college and employer), friends and interests respectively in terms of movies, music and TV. We hide these users' interests and collect them into a new users set (i.e., $U_{new}$) to recommend items. In addition, we use the rest of users who present more than 3 movies, music or TV shows as the *existing active users*. By using the above-

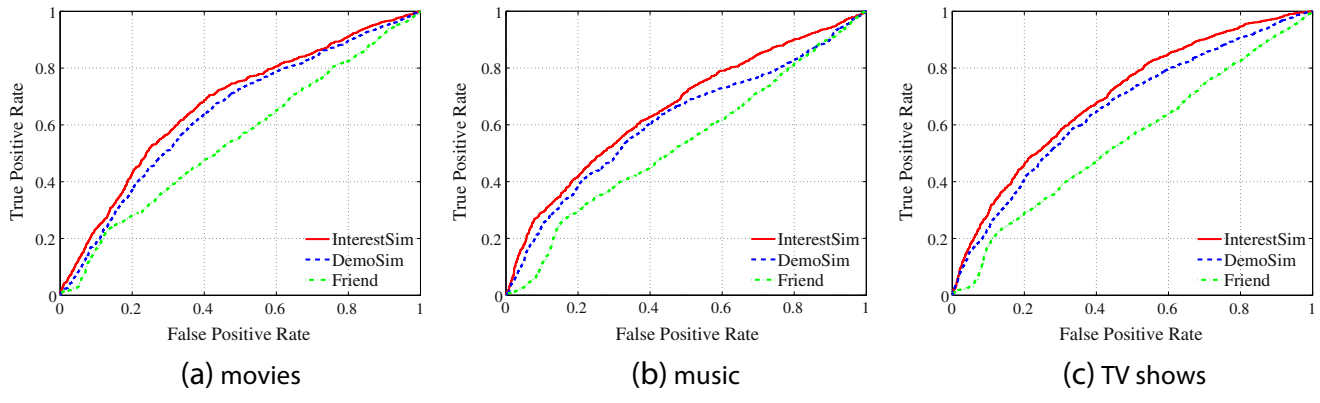


**Fig. 10.** ROC curves of prediction.

mentioned recommendation approaches, we generate recommendation item lists for the new users from the preferences of the *existing active users*, and eventually we compare the recommended items with the new users' real preferences.

To evaluate and compare the performance of the above-mentioned approaches, we respectively select the top 5, top 10, top 20 and top 100 items to generate the recommendation lists. We estimate the effectiveness of the recommendations by a quite commonly used metric — precision [3, 5, 19, 22]. In fact, *precision* estimates how many percentage of recommendations are the users' real interests. Assume that a new user $u \in U_{new}$ has $p_u$ specific preferences; we recommend $q_u$ items to the $u$ where $r_u$ among these $q_u$ items are $u$'s real interests. Then, we have $precision = \frac{1}{N}\sum_{u \in U_{new}} r_u/q_u$, where $N$ is the number of new users in $U_{new}$. By the definition of *precision*, a good recommendation approach should exhibit a large *precision*.

Fig. 11 compares the *precision* of our proposed *InterestSimPop* recommendation to the other four baselines. We observe that our proposed *InterestSimPop* approach achieves the largest *precision* no matter what the interest domain refers to. This indicates that our proposed approach can improve effectiveness of recommendations for a new user. For instance, in Fig. 11(a), the *precision* of *InterestSimPop* is around 0.45 for the top 5 recommendations, which means we can correctly recommend 2–3 movies out of the top 5 recommendations to the new users on average; however the other approaches cannot ensure 1 correct movie recommendation.

## 9. Discussion

In this section, we further discuss two concerns: 1) social feature selection; 2) the practical use of the proposed interest similarity prediction model.

**Social feature selection:** To fully exploit the obtainable information in the prediction, besides demographic information and friendships, we handily use interest entropy to characterize the active user's interests and luckily find that two users' interest similarity correlates to interest entropy. Thus, we leverage the active user's interest entropy with other demographic and friendship features into the prediction model. The 'leave-one-feature-out' evaluation reveals the positive effects of interest entropy and all other social features. This just indicates all the studied social features can improve the prediction. For the future work, we may improve the prediction model if more social features could be obtained.

**Use of the proposed prediction model:** We have illustrated how to use our prediction model to enhance the recommendation for new users. We also believe that our proposed model can be easily used to other applications, like friend recommendation. Although several existing approaches may rely on mutual friends, colleague or classmate, we propose to recommend friends according to interest similarity for the following reasons: 1) as our proposed interest similarity prediction model exhaustively exploits the users' obtainable information, the interest similarity based friend recommendation may substitute for the existing approaches once their requisite information (e.g., friend, job or school) is missing; 2) The promising of the interest-based OSNs like Pinterest, CircleMe and Yaamo reveals that people like to connect other people with similar interests. It has also been proved that users who share certain interests are more likely to be friends [1, 10, 14]. Thus, a mixed solution, which includes all the approaches based on mutual friend, colleague, classmate and interest similarity, may be an alternative.

## 10. Conclusion

As users do not always explicitly elaborate their interests in OSNs, in this paper, we address a practical problem for OSNs: How to infer two users' interest similarity when we cannot fully know their interests.

To solve this problem, from users' demographic information, friendships and their interests, we first attempt to identify some users' social features (e.g. geographic distance, friend similarity) that are strongly correlated to their interest similarity. In particular, we conduct a comprehensive empirical study on how users' interest similarity relates to various social features in a large Facebook dataset including 479, 048 users and 5, 263, 351 user-generated interests. We conduct the study in three largest interest domains (i.e. movies, music, and TV shows). The result reveals that people tend to exhibit more similar tastes if they have similar demographic information (e.g., age, location) or share more common friends. In addition, we also find that the individuals with a higher interest entropy would generally share more interests with the others. Finally, we identify several effective social features that are strongly correlated to users' interest similarity, including geographic distance, gender combination, age distance, friend similarity, interest entropy, etc.

Based on the above identified social features, we propose a user interest similarity prediction model that can determine whether two users are similar or not in an interest domain while interests cannot be obtained from one of them. The evaluation demonstrates that the prediction model integrating all the learned social features outperforms other models that lack some of those features.

## Acknowledgment

This work has been funded by the European Union under the project eCOUSIN (EU-FP7-318398) and the project SITAC (ITEA2-11020). This work has also been partially funded by the Ministerio de Economía y Competitividad of SPAIN through the project BigDatAAM (FIS2013-47532-C3-3-P).

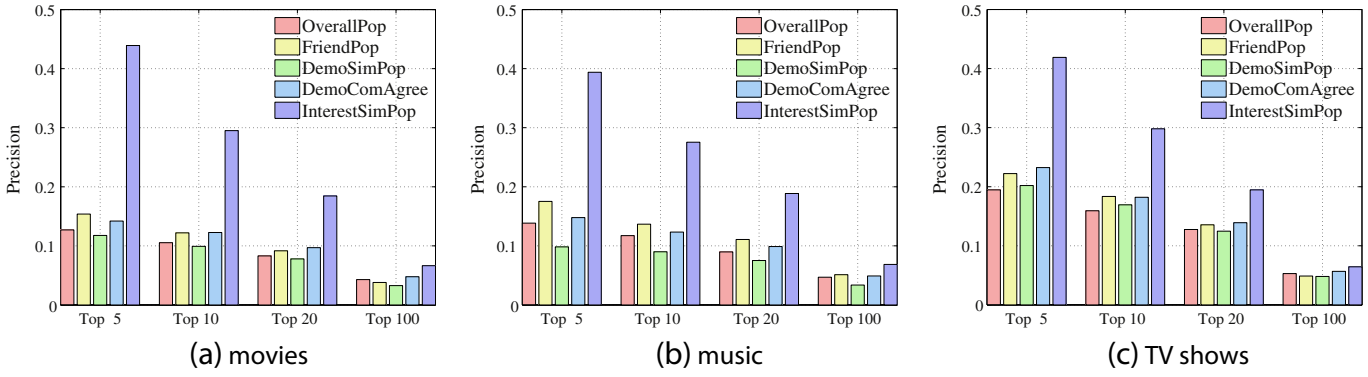


**Fig. 11.** Evaluation on recommendation precision.

## References


[1] L. Adamic, E. Adar, Friends and neighbors on the Web, Social Networks 25 (2003) 211–230.
[2] L.A. Adamic, J. Zhang, E. Bakshy, M.S. Ackerman, Knowledge sharing and yahoo answers: everyone knows something, Proceedings of the 17th International Conference on, World Wide Web, 2008, pp. 665–674.
[3] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 734–749.
[4] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, Effects of user similarity in social media, Proceedings of the fifth ACM international conference on Web search and data mining, ACM, New York, NY, USA, 2012, pp. 703–712.
[5] J. Bobadilla, F. Ortega, A. Hernando, J. Bernal, A collaborative filtering approach to mitigate the new user cold start problem, Knowledge-Based Systems 26 (2012) 225–238.
[6] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27.
[7] T. Chen, L. He, Collaborative filtering based on demographic attribute vector, International Conference on Future Computer and, Communication, 2009, 2009, pp. 225–229.
[8] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[9] R. Dey, C. Tang, K. Ross, N. Saxena, Estimating age privacy leakage in online social networks, INFOCOM, 2012, Proceedings IEEE, 2012, pp. 2836–2840.
[10] L. Gou, F. You, J. Guo, L. Wu, X.L. Zhang, Sfviz: interest-based friends exploration and recommendation in social networks, Proceedings of the 2011 Visual Information Communication-International Symposium, ACM, 2011, p. 15.
[11] S. Jamali, H. Rangwala, Digging digg: comment mining, popularity prediction, and social network analysis, International Conference on Web Information Systems and Mining (WISM), 2009, 2009, pp. 32–38.
[12] D.H. Lee, P. Brusilovsky, Social networks and interest similarity: the case of citeulike, Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 2010, pp. 151–156.
[13] J. Leskovec, E. Horvitz, Planetary-scale views on a large instant-messaging network, Proceedings of the 17th international conference on, World Wide Web, 2008, pp. 915–924.
[14] K. Lewis, M. Gonzalez, J. Kaufman, Social selection and peer influence in an online social network, Proceedings of the National Academy of Sciences 109 (2012) 68–72.
[15] Loh, S., Lorenzi, F., Granada, R., Lichtnow, D., Wives, L.K., de Oliveira, J.P.M., Identifying similar users by their scientific publications to reduce cold start in recommender systems, in: WEBIST'09, 2009, pp. 593–600.
[16] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annual Review of Sociology 27 (2001) 415–444.
[17] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, Proceedings of the 7th ACM SIGCOMM conference on Internet, measurement, 2007, pp. 29–42.
[18] A. Mislove, B. Viswanath, K.P. Gummadi, P. Druschel, You are who you know: inferring user profiles in online social networks, Proceedings of the third ACM international conference on Web search and data mining, 2010, pp. 251–260.
[19] A.T. Nguyen, N. Denos, C. Berrut, Improving new user recommendations with rule-based induction on cold user data, Proceedings of the 2007 ACM Conference on Recommender Systems, 2007, pp. 121–128.
[20] M.J. Pazzani, A framework for collaborative, content-based and demographic filtering, Artificial Intelligence Review 13 (1999) 393–408.
[21] S. Phithakkitnukoon, H. Husna, R. Dantu, Behavioral Entropy of a Cellular Phone User, Proceedings of the First International Workshop on Social Computing, Behavioral Modeling, and Prediction, 2008, pp. 160–167.
[22] A.M. Rashid, I. Albert, D. Cosley, S.K. Lam, S.M. McNee, J.A. Konstan, J. Riedl, Getting to know you: learning new user preferences in recommender systems, Proceedings of the 7th international conference on Intelligent user, interfaces, 2002, pp. 127–134.
[23] A.I. Schein, A. Popescul, L.H. Ungar, D.M. Pennock, Methods and metrics for cold-start recommendations, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in, information retrieval, 2002, pp. 253–260.
[24] P. Singla, M. Richardson, Yes, there is a correlation: from social networks to personal behavior on the web, Proceedings of the 17th international conference on, World Wide Web, 2008, pp. 655–664.
[25] C. Song, Z. Qu, N. Blumm, A.L. Barabási, Limits of predictability in human mobility, Science 327 (2010) 1018–1021.
[26] E. Spertus, M. Sahami, O. Buyukkokten, Evaluating similarity measures: a large-scale study in the orkut social network, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, New York, NY, USA, 2005, pp. 678–684.
[27] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the facebook social graph, 2011. (CoRR abs/1111.4503).
[28] B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, On the evolution of user interaction in facebook, Proceedings of the 2nd ACM workshop on Online, social networks, 2009, pp. 37–42.
[29] Z. Wen, C.Y. Lin, On the quality of inferring interests from social neighbors, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 373–382.
[30] C. Wilson, A. Sala, K.P.N. Puttaswamy, B.Y. Zhao, Beyond social graphs: user interactions in online social networks and their implications, ACM Transactions on the Web 6 (2012) 17:1–17:31.
[31] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, ACM, New York, NY, USA, 2010, pp. 981–990.
[32] C.N. Ziegler, J. Golbeck, Investigating interactions of trust and interest similarity, Decision Support System 43 (2007) 460–475.



**Xiao HAN** studied in College of Automation of Northwestern Polytechnical University, China, and received her BSc and Msc there respectively in 2008 and 2011. Since Sep. 2011, she is pursuing her PH.D. at Service Architecture Lab in Télécom SudParis, Institut Mines-Télécom. Her research interests include Analysis on Social Networks, Social-based Applications, Peer-to-Peer Networks and P2P networks.

**Leye WANG** is currently a full-time PhD candidate in TELECOM SudParis. He received his M.S. and B.S. in Computer Science in Peking University. His current research interests lies in ubiquitous computing, particularly in mobile crowd sensing.

**Noel CRESPI** holds a Master's from the Universities of Orsay and Kent, a diplome d'ingenieur from Telecom ParisTech, and a Ph.D. and a Habilitation from Paris VI University. He worked from 1993 in CLIP, Bouygues Telecom, France Telecom R&D in 1995, and Nortel Networks in 1999. He joined Institut Mines-Telecom in 2002 and is currently professor and program director, leading the Service Architecture Laboratory. He is appointed as coordinator for the standardization activities in ETSI and 3GPP. He is also a visiting professor at the Asian Institute of Technology and is on the four-person Scientific Advisory Board of FTW, Austria. His current research interests are in service architectures, P2P service overlays, future Internet, and Web-NGN convergence. He is the author/coauthor of more than 230 papers and contributions in standardization.

**Soochang PARK** received the Ph.D. degree in Computer Science and Engineering from Chungnam National University in Aug. 2011. He worked at RUCOR in Rutgers University in USA as a post-doctoral researcher in 2012, and currently he is working as a research fellow at Wireless Networks and Multimedia Services Department of Telecom SudParis, Institut Mines-Telecom. His research interests are the areas of computer communication and networking. He is mainly interested in IP routing with AS-level configuration; mobility management in both IPv4 and IPv6 including host-based and network-based; and routing, mobility management, and QoS in MANETs and WSNs.

**Angel CUEVAS** received his MSc in Telecomunication Engineering, MSc in Telematics Engineering, and Ph.D. in Telematics Engineering from the Universidad Carlos III de Madrid in 2006, 2007, and 2011 respectively. He is currently a tenure-track Visiting Professor at the Department of Telematic Engineering at Unviersidad Carlos III de Madrid. Pior to that he was Postdoc researcher at Institut Mines-Telecom, Telecom Sudparis from March 2011 until Jan 2013. His research interests focus on On-line Social Networks, P2P Networks, Wireless Sensor Networks and Internet measurements.