Social Networks 55 (2018) 160–169

Contents lists available at ScienceDirect

Social Networks

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

Homophily of music listening in online social networks of China

Zhenkun Zhou[a], Ke Xu[a], Jichang Zhao[b,c,*]

[a] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[b] School of Economics and Management, Beihang University, Beijing, China
[c] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China

ARTICLE INFO

Keywords:
Homophily
Online social networks
Music listening
Music genres

ABSTRACT

Homophily, ranging from demographics to sentiments, breeds connections in social networks, either offline or online. However, with the prosperous growth of music streaming services, whether homophily exists in online music listening remains unclear. In this study, two online social networks of the same group of active users who listened to complete songs over 1000 times and posted over 100 tweets are established, separately, in Netease Music and Weibo. Through presented multiple similarity measures, it is evidently demonstrated that homophily does exist in music listening for both online social networks. The unexpected listening similarity in Weibo also implies that knowledge from generic social networks can be confidently transferred to domain-oriented networks for context enrichment and algorithm enhancement. Comprehensive factors that might function in the formation of homophily are further probed, and many interesting patterns are profoundly revealed. It is found that female friends are more homogeneous in music listening and that positive and energetic songs significantly pull users close. Our methodology and findings shed light on realistic applications in online music services.

1. Introduction

One of the best-established findings in social networks is that people who are friends exhibit plenty of similarities in human behaviors (Mcpherson et al., 2001; De Klepper et al., 2010). Friendship relationships, either offline (Reagans, 2005) or online (Buote et al., 2009; Antheunis et al., 2012), in which individuals socially interact involve a need for shared mutual understandings. Tremendous efforts have been devoted to the homophily of social networks from many aspects, ranging from demographics (Chmiel et al., 2011) to mental states (Bollen et al., 2011; Fan et al., 2014). It is even revealed from a recent study that personality similarity exists among close relationships (Youyou et al., 2017). Particularly, with the flourishing of online social media, previous studies have extensively investigated the homophily of behaviors in online social networks (Brown et al., 2007; Mislove et al., 2010; Centola, 2010), and the friend similarity that exists in face-to-face offline networks is identically revealed. However, as a prominent element of daily life that possesses cultural universality (Blacking, 1995; North, 2004; Wright, 2013), music listening is rarely explored in the context of social networks, and little knowledge is established regarding the behavior referring to homophily, particularly in the circumstance of online social networks.

In fact, individuals embedded in social networks come across music of varying categories, including vast types of genre, language and mood, and continually judge whether they like the music (Frith, 2002). In addition, music is always shared with families, friends and other people socially around us (Håkansson et al., 2007; Voida et al., 2005). Before the era of the Internet, CDs and cassettes were the main media for recording music, and music communication was thus limited (Boström et al., 1999; Miell et al., 2005). However, in the last twenty years, with the prosperous development of the Internet, portable music players have exploded in popularity, essentially promoting music communication (Holmquist, 2005; Rondeau, 2005). Friends have been willing to exchange their iPods with each other, and now, music streaming platforms offer low-latency access to large-scale databases, such as Spotify (Kreitz and Niemela, 2010), Last.fm (Henning and Reichelt, 2008), QQ Music (Priest, 2006) and Netease Music (Fung, 2007). Since then, people have exchanged music with each other online and shared the amazing music they like, and music even creates interpersonal bonds between different individuals in turn (Boer et al., 2011). Although evidence of listening similarity in offline friendships has been demonstrated (Selfhout et al., 2009), the listening similarity of friends in online social networks has not been comprehensively explored and understood yet. Specifically, questions such as whether we enjoy the music with which online friends are enchanted, in other words, whether the homophily of music listening exists in online social

* Corresponding author at: School of Economics and Management, Beihang University, Beijing, China.
E-mail address: jichang@buaa.edu.cn (J. Zhao).

https://doi.org/10.1016/j.socnet.2018.07.001
Received 19 November 2017; Received in revised form 4 July 2018; Accepted 5 July 2018

networks, still deserve a systemic investigation.

Until recently, empirical research willing to answer the questions about music listening had to depend on interviews and surveys in controlled environments (Sloboda, 1999; Greenberg et al., 2016) with inevitable limitations in both data scale and granularity. However, the technological and societal evolutions that sustain the emergence of online music listening indeed provide unparalleled opportunities for human behavior understanding. Detailed footprints, including where, when and how numerous individuals listen to music, can be regarded as a big-data window through which the homophily in music listening can be collectively or individually probed and studied thoroughly. For instance, Netease Music, one of the most popular online music providers in China, provides a high-quality music streaming service to millions of users and accordingly accumulates the detailed behavior records of these users continuously. According to the official report of Netease Music in 2016 News (2016), it had over 200 million users. With music playlist creation being the core listening pattern, each day, users establish approximately 420 thousand playlists, and user-generated playlists total 800 million. In the first half of 2016, users played songs 1.82 billion times, and the duration amounted to 7.2 billion minutes, implying the impressive vitality of users in online music listening. Even more inspiring, Netease Music has developed one extraordinary trait of socializing its users. Specifically, it first provides a domain-specialized social network through which users engage in sharing music interests. Similar to generic online social networks, users can be networked by following others, not only ordinary users but also artists. Indeed, the social network sourced in music listening profoundly facilitates the acquiring and sharing of music interests, implying an ideal entanglement between social ties and music listening for the present study. Thus, anonymous digital traces of numerous users are collected to quantitatively support the investigation of homophily in music listening.

However, domain-oriented social networks such as the one established by Netease Music cannot be a typical representation of online social networks, which generally result from comprehensive factors. Specifically, the Netease Music social network relies predominately on the musical interest, and its digital traces are insufficient to describe other individual traits. In the meantime, evidence of musical preferences being linked to individual traits, such as personalities (Greenberg et al., 2016), cognitive styles (Greenberg et al., 2015) and even socioeconomic statuses (Park et al., 2015), has been extensively demonstrated, implying the consideration of more generic online social networks. We argue that aiming at a systemic understanding, it is necessary to study listening practices based on other more typical social networks in which users are joined sophisticatedly but realistically by psychological traits, extensive interests or other individual characteristics. Considering the development of social media in recent decades, prosperous networks such as Twitter or its variant Weibo of China, which aggressively replicate offline social networks to online counterparts, can be ideal targets. These online social networks are natural, long-term and diverse, and thus, the objective footprints of massive individuals can be promising proxies for the present study. Nonetheless, it is still extremely difficult to correctly match each individual of the music social network to the identical one in networks such as Weibo, which is the essence of embedding music listening in a generic online social network. Very fortunately, users can log into Netease Music through their Weibo accounts, and along this line, we can obtain the digital traces of Weibo for these users. Therefore, it is possible to further study the similarity of music listening for friends who are linked in Weibo. In addition, demographics and tweets in Weibo are excellent supplements for enriching individual characteristics. We can even explore which key factors influence the similarity of music listening. Hence, we investigate the following research questions in this study:

- RQ1. Is there homophily of music listening behavior in online social networks of China?
- RQ2. Which factors influence the listening similarity of friends in online social networks of China?

Starting from the above motivations and assumptions, in this study, digital footprints of over seventy thousand individuals from both Netease Music and Weibo are thoroughly collected through a crawler. Then, 25,953 active users who listened to songs 1000 times and posted at least 100 tweets are sampled as the subjects for further exploration. Two online social networks, the Netease network and Weibo network, are constructed through user followings in Netease Music and Weibo, respectively. To examine whether music listening is homogeneous for users joined by online social networks, listening similarities from six perspectives are defined and measured, which can represent the strength of ties of music between online friends. To investigate the crucial factors influencing the music listening homophily, subjects are clustered into different categories from multiple perspectives such as social attributes and musical preferences.

Our results demonstrate that friends linked by each of online social networks (Weibo and Netease Music) indeed appreciate identical songs and possess similar music preferences. As for gender, patterns for music listening between female friends are closer than those between male friends, implying that women are more sensitive to emotional expression through music (Wells and Hakanen, 1991; Robazza et al., 1994). The listening practices of friends with common music preferences in the current culture are similar, particularly for music languages (Chinese) and genres (pop and folk). With regard to musical mood, the users who enjoy exciting, wild and happy music share more similarities in music listening. It is also difficult for users possessing high musical diversity to find friends with similar musical tastes. The present study confirms the existence of homophily in music listening of online social networks and elaborately clarifies the roles of human demographics and music traits in influencing the homophily. Our findings shed insight on music recommendations and friend suggestions in online applications. We merge different social circles of an individual and surprisingly reveal that generic social networks, such as Weibo, still significantly demonstrate the homophily that intuitively only exists in domain-oriented networks. This result indeed implies that rich information in general social media can be confidently introduced into the study of specialized social networks (Carmagnola and Cena, 2009).

Although most previous studies explored the homophily by questionnaires (Baym and Ledbetter, 2009; Hagen and Lüders, 2017), there have been some effective data-driven methods to investigate homophily of music listening in online social networks. Lambiotte and Ausloos analyzed correlations between online music groups of different genres in Last.fm and constructed a music genre cartography, with a tree representation (Lambiotte and Ausloos, 2006). Some researchers detected the communities from the social network and investigated whether the creation of these ties was influenced by the similarity of interest (Bisgin et al., 2010). Aiello et al. defined the topical similarity among users who were close to each other in the social network of three systems to study the presence of homophily that combined tagging (Aiello et al., 2012). However, they lacked connections between users in these different social networks. To the best of our knowledge, we are the first to build different social networks with the same group of active users from Netease Music and Weibo for probing six measures of homophily in music listening.

2. Dataset

2.1. Netease music dataset

Digital service providers began to amass large user bases, increasingly offering the primary sources of digital music streaming through the Internet. New innovations, including digitalization and the Internet, have transformed the existing landscape over the past decade and attracted new artists and listeners into the fields, and digital music

streaming services have also profoundly reshaped user behaviors. China has followed this global trend and become a leading digital country in terms of music services. Netease Music is one of the most trending music streaming providers in China, whose special trait is the music social network. Users can create many own playlists to collect the songs. Users are also allowed to follow others in the platform, such as friends and artists. Therefore, this music platform has become a new and prosperous domain-oriented social network.

The website of Netease Music provides abundant online information about users, playlists and music. In Netease Music, the individual information is publicly open and users can view others' historical listening records (only the top 100 songs), like records and lists of followers and followees. We develop an crawler aimed at the Netease Music platform first to perform the data collection. Our crawler adapts numerous agent servers, and each agent will simulate a real user to visit pages and click links from the browser. We choose one of the authors who has a Netease Music account as the seed user. Applying the method of snowball sampling (Heckathorn, 1997; Baltar and Brunet, 2012), by traversing both the following and followed links from August 2016 to May 2017, we obtain a dataset of over 200 thousand users, 1.5 million playlists and 3.2 million songs. We employ the snowball sampling from one seed to collect the structure of the Netease Music network, since crawling the entire Netease Music is not practically feasible. In previous study of online social networks, it is pervasively found that the snowball sampling, which usually starts from a random seed and gradually collects the structure, can produce a complete picture of a dense core of the entire network (Lee et al., 2006; Mislove et al., 2007) and the obtained local network can sufficiently reflect the characteristics of the entire network (Heckathorn, 1997). Referring to the seed node, we randomly select one user to collect the network with more than 200,000 nodes. Considering that the snowball sampling always drills into the dense core of the network (Mislove et al., 2007), replacing the current seed with other nodes will not essentially influence the following experiments.

The collected data referring to users include the following lists, historical listening and like records. For each user, the following list contains the complete ID numbers of users (UID) that she/he followed. The top 100 songs listened to by one user are recorded in her/his historical listening. The platform provides the 'like' button to label their favorite playlists or songs, which are restored in like records. According to the historical listening and like records, we obtain the corresponding detailed playlists and attributes of songs. In each playlist, the ID numbers of songs (SID) it contains and style tags labeled by its creator are both collected. The attributes of a song include the album, artists and several acoustic features.

2.2. Weibo dataset

Tremendous social relationships are forged in popular online services such as Facebook, Twitter and Weibo with the explosive development of online social networks. Twitter, as one of the most popular social networking and micro-blog services, enables registered users to read and post short messages, so-called tweets. At the beginning of 2016, Twitter had reached 310 million monthly active users (MAU). As of the third quarter of 2017, it averaged 330 million MAU. However, more people are using Weibo, the Chinese variant of Twitter, now. According to the Chinese company's first quarter reports, it has 340 million MAU, a 30% increase from the previous year, implying that the social relationships and online behaviors of numerous individuals from Weibo can be sensed and profiled with unparalleled richness and granularity. Moreover, replicating offline social ties to online and replacing face-to-face communication with interactions in cyberspace have essentially digitized daily life and reshaped social networks, suggesting that the online social network is of fundamental significance for explorations in both social theories and applications. The individual information in Weibo is publicly open. For example, users in Weibo can other's view the profiles (gender, etc.), tweets (texts, pictures and videos) and lists of followers and followees (incomplete and last 1000 names) of each other.

In the preceding section that discussed the Netease Music dataset, 74,056 users connect their Weibo accounts to Netease Music, offering us an opportunity to establish a perfect match between these two different platforms. Employing the crawler agents from May to June 2017, their (74,056 users) corresponding ID numbers in Weibo are obtained. With the help of Weibo's open APIs, for each ID, the public profiles, following relationships and historical tweets can be accordingly collected. In addition, to speed up the collection, cloud servers are also extensively utilized. The profiles of Weibo users include demographics (e.g., gender), social attributes (e.g., the number of followers) and other individual information. For each Weibo user, the following relationships contain the Weibo ID numbers (WID) of users they follow. Unfortunately, due to the official limit on viewing following lists, we can obtain only the latest 1000 following relationships. We also investigate the mentioning behavior (@ behavior) of users to examine the strength of ties between users and their last one thousand friends. Mentioning behavior is regarded as one of the most popular forms of online interactions between friends (Gao et al., 2012). In our study, 72% of mentioning behavior appears between users and their last one thousand friends and particularly, 35% of Weibo users only mention these friends. These numbers imply that the last one thousand followings could sufficiently reflect recently active social connections the user possesses. Besides, there are 1463 users who follow more than one thousand friends, taking up only 5% of active users. Considering the number of following links for all the 25,953 active users is up to 10,540,290, the number of missed following links in our dataset only occupies 6.5%. It is also found that in Twitter, users who follow over one thousand users account for less one percent (Kwak et al., 2010). Therefore, this limitation of collecting the last 1000 friends will not significantly influence our following experiments.

3. Methods

3.1. Two online social networks

We refine the subjects of the present study by selecting active users from both datasets. Specifically, active users are defined as those have listened to complete songs at least 1000 times and posted over 100 tweets. The complementary cumulative distribution function (CCDF) of the number of songs listened to and tweets by users is depicted in Fig. 1. As seen in Fig. 1a, for the majority of users, the number of songs listened to is less than 10^4. The percentage of users who have listened to music over 1000 times is less than 55%. As shown in Fig. 1b, approximately 70% of users have tweeted at least 100 times. Note that owing to the limit of Weibo, the maximum number of tweets is 1000. After the refinement, a total of 25,953 active users are filtered out to be subjects of further exploration.

Regarding the same group of active users, two social networks are then separately established from the following relationships in Netease Music and Weibo. The example of network topology is shown in Fig. 2. In both networks, nodes represent the users, and edges represent that there is a following relationship in the network between a pair of users. The first one, named the Netease network, contains 89,988 edges. The second one, named the Weibo network, contains 112,753 edges, and these two networks share 25,953 identical nodes, implying that we construct an ideal scenario to reveal the homophily of music listening in both domain-oriented social networks and generic social networks. It is worth noting that here, we regard social networks as undirected graphs along this study. When user A follows B, user B will know the following link from A and can access user A's historical listening records, profiles and tweets. Accordingly it can be conjectured that a following relationship will influence the behavior of both its ends mutually and simultaneously, and thus, from the perspective of listening behavior, it

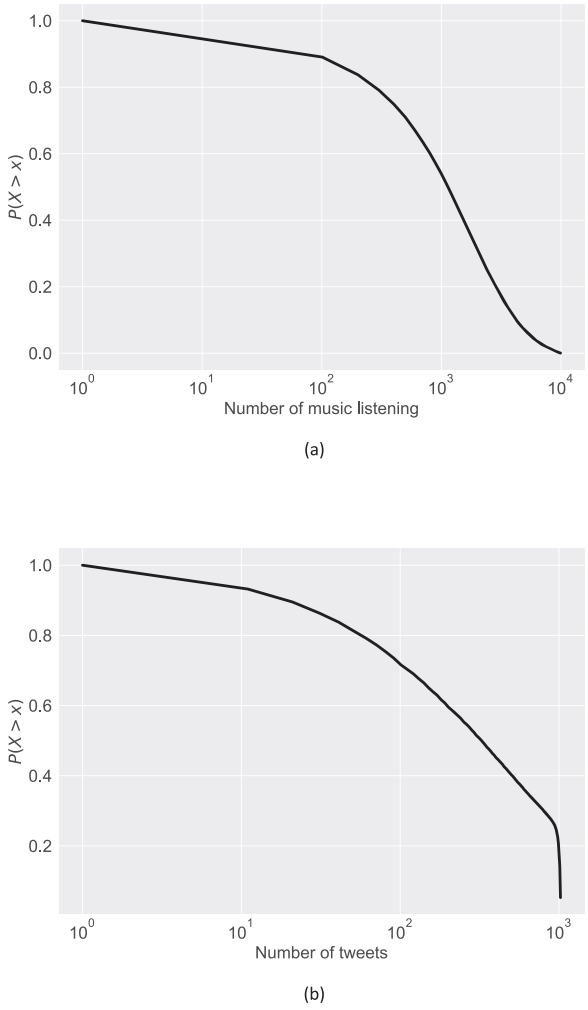

Fig. 1. CCDF of the number of songs listened to and tweets.

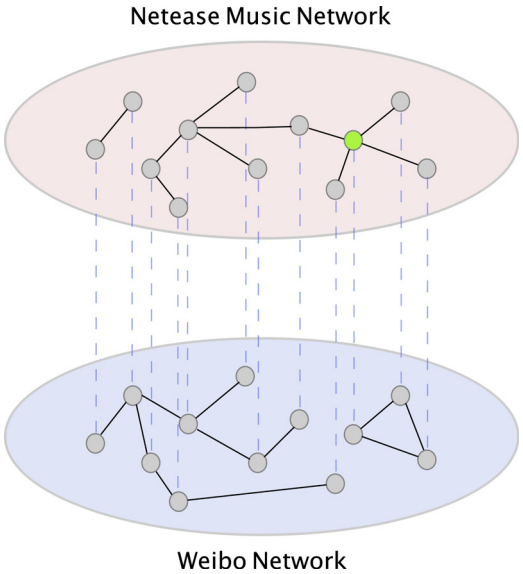

Fig. 2. An example of the network topology. A single set of nodes (users) with two distinct edge sets in Network Music and Weibo. A dashed line implies that the nodes it connects are identical, i.e., both of its ends are the same user. The green node represents the seed node in crawling.

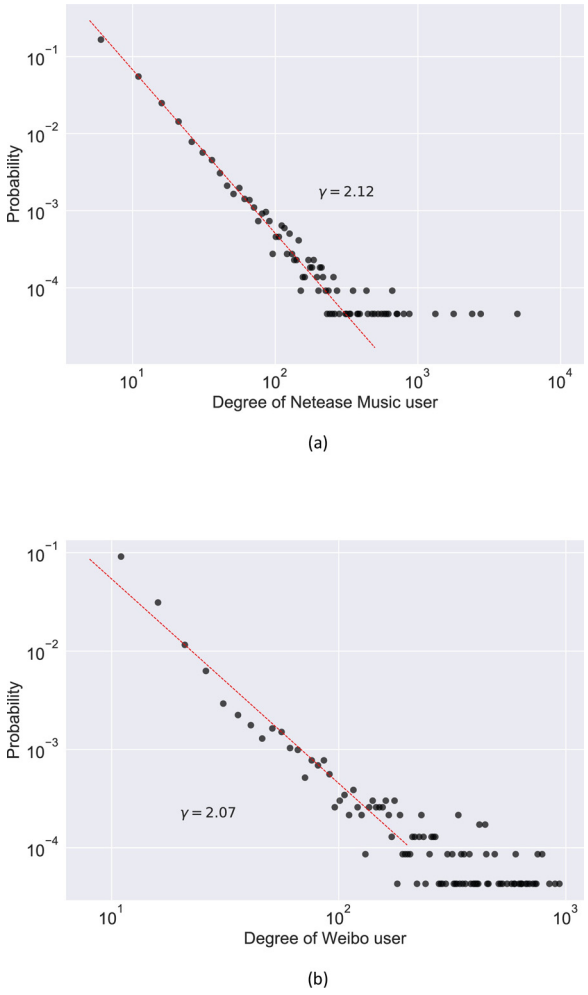

Fig. 3. Probability distribution of the degree of the Weibo network and Netease Music network. The degree distribution exponent γ of the networks is, respectively, 2.12 and 2.07 ($2 < \gamma < 3$).

can be viewed as a undirected connection. The degree distributions of the Netease network and Weibo network are demonstrated in Fig. 3, and the power-law trend implies consistence with an existing understanding of typical social networks (Newman, 2010). As for the Netease network ($\gamma = 2.12$), the distribution evidently shows a long tail, indicating that there are few users with a large number of relationships in music-oriented social networks. As for the Weibo network ($\gamma = 2.07$), a power-law like distribution is observed of degree lower than 500; however, an exponential cutoff then emerges, and the maximum degree is 100, both due to the API limit of Weibo.

3.2. Listening similarities

Based on the personal historical listening and like records, we define six types of listening similarity with respect to music listening for any pair of users A and B in both social networks. The similarities between the friends of a user and the user itself discussed in this study exactly follow the six types of listening similarity. Additionally, higher similarities suggest more significant homophily in social networks.

The historical listening records for A and B are denoted as $\boldsymbol{H}_A$ and $\boldsymbol{H}_B$, respectively, in which the top 100 songs of high frequency are included. The historical similarity sim_{song}, indicating the number of co-occurrence of songs in top records, can be thus defined as

$$\mathrm{sim}_{\mathrm{song}} = |\boldsymbol{H}_A \cap \boldsymbol{H}_B|. \quad (1)$$

For presenting measures of similarity from the aspects of music

traits, we first develop ways to label songs and users in terms of trait vectors. Given the lack of tags for songs in the Netease dataset, the tags of playlists can be fused to infer the trait vectors of songs. Specifically, Netease officially provides 6 tags of languages, 24 tags of genres, 13 tags of moods and 12 tags of scenes, and among these 55 tags, at most three are selected to label a playlist by its creator. Assuming that a song can be labeled confidently by tags of the playlists it belongs to, a 55-dimension trait vector (initialized to zero) can be accordingly derived for each song by examining 1.5 million playlists, i.e., a tag occurrence will result in the addition of 1 to the corresponding dimension. However, considering the fact that values of the trait vector of a song can be significantly influenced by the number of playlists in which it appears, we split the 55-dimension trait vector into four sub-vectors including S_c (S refers to **s**ongs, $c \in \{language, genre, mood, scene\}$ refers to traits) and then perform normalization on each sub-vector to avoid bias. To be specific, we calculate the proportion of each tag in the four vectors, and the sum of values in each sub-vector should be 1. In fact, S_c can be features to appropriately represent the category distribution of songs. Following the same idea, supposing that the music preferences of a user can be well reflected by the top songs listened to, the trait vectors of users can be inferred through

$$U_c = \sum_{i=1}^{100} S_{c,i}. \tag{2}$$

Finally, based on the feature vectors of 3.2 million songs, trait vectors for 25,953 active users are derived and normalized to z-scores. Then, the similarity of user pair A and B can be intuitively measured through the cosine distance between their trait vectors, which is defined as

$$\mathrm{sim}_c = \frac{U_{c,A} \cdot U_{c,B}}{|U_{c,A}||U_{c,B}|}. \tag{3}$$

Note that $-1 \leqslant sim_c \leqslant 1$ and that values closer to 1 indicate more similar music preferences, and as c varies over different music traits, we accordingly obtain four music preference similarities from the perspectives of languages, genres, moods and scenes.

In addition, the similarity of favorite songs can also reflect the strength of homophily in music preference. Different from historical listening records, the number of songs in like records is not limited. Therefore, we define the Jaccard distance as the similarity for like records between users A and B as follows:

$$\mathrm{sim}_{\mathrm{like}} = \frac{|\boldsymbol{L}_A \cap \boldsymbol{L}_B|}{|\boldsymbol{L}_A \cup \boldsymbol{L}_B|}, \tag{4}$$

in which $\boldsymbol{L}$ stands for the set of one's favorite songs.

To sum up, from the above, we obtain six similarities referring to music listening, and higher similarities indicate more significant homophily in music listening for online social networks. These different measures pave the way for quantitative investigations of homophily in this study from various views.

Homophily can also be measured through structural equivalence (Burt, 1982). Due to the limitation of the dataset, here, we define an index called the conditional probability of sharing ties between Netease Music and Weibo to represent structural similarities. This probability is defined to reflect to what extent a tie between user A and user B in Netease Music will also exist in Weibo, or vice versa. The results show that the structural similarities between Weibo and Netease Music networks are significant, specifically, P(tie in Netease | tie in Weibo) = 0.23 and P(tie in Weibo | tie in Netease) = 0.34. This result suggests that the similarity, or homophily from the perspective of structural equivalence, is consistent with our definitions from the view of listening behavior, and understanding the structural similarity of the two social networks will offer a promising direction for future work.

3.3. User classifications

Many factors in social networks might influence the homophily between friends. To reveal the crucial factors, the demographics, social attributes and listening behaviors of active users are further probed for a detailed understanding of their roles in homophily of music listening. We argue that by regarding these factors as features, active users can be thus clustered into groups from multiple perspectives, and discriminations of inter-groups offer windows into homophily investigation and factor weighting.

Gender is one of the most significant demographics in understanding human behaviors (Gefen and Straub, 1997; Costa et al., 2001; Schwartz et al., 2013). We extract gender from Weibo profiles and split active users into 10,388 female ones and 15,565 male ones. Whether a user is officially verified by Weibo can be an indicator of influence, and through the 'verified' labels, active users are split into 24,368 non-verified ones and 1585 verified ones. In addition, social attributes that reflect users' popularity in the social network can be well modeled through the number of follow**er**s (*NER*) as well as *RFF*, defined as the **r**ate of **f**ollowed numbers to **f**ollowing numbers ($log(NER + 1/NEE + 1)$, in which *NEE* refers to the number of the users' followees). The logarithmic rate of followed numbers to following numbers of users is measured since we attempt to normalize this rate for following analysis. If the followed number of one user is larger (or less) than the following number, the logarithmic rate will be positive (or negative). And then the logarithmic normalization of rates benefits K-means to cluster users. The Pearson correlation between *NER* and *RFF* is very low (0.22, p-value < 0.001 ***), suggesting the independence between these two attributes is significant. Because *RFF* contains the information of following numbers, but *NER* does not. Hence, we classify users in terms of both of *NER* and *RFF*. Given that both *NER* and *RFF* are continuous variables, the discretization based on K-means clustering is employed to categorize active users into groups of **low** rank and **high** rank based on these two attributes.

Patterns of online music consumption could be sufficiently reflected in terms of users' music preferences that directly result in behavioral differences. Hence, active users can also be clustered into groups from the perspective of music preference. The trait vector U_c can be features representing users in clustering, and the preference feature matrix of c is accordingly constructed for all users. The rows of feature matrix represent traits of each users and the columns represent different dimensions of traits, including "pop" (genre), "Chinese" (language), "happy" (mood) and etc. The matrix is decorrelated by principal component analysis (PCA) with full covariance, and then, the approach of K-means is employed to cluster active users into groups. Clusterings based on all traits are denoted separately as $C_{language}$, C_{genre} and C_{mood}, indicating that for each trait, active users are grouped into three clusters with divergent patterns of music preferences. To better interpret the clusterings, for the groups of each clustering, the group feature, defined as the mean user feature within the group, is calculated. Then, the semantic of each group is explained by the top 2 ~ 3 attributes of

Table 1
User classifications based on perspectives about music preferences.

Traits	Groups	Interpretations
Language	0	Chinese (including Cantonese)
	1	Japanese, Korean
	2	English, minority
Genre	0	rap, dance, alternative
	1	pop, folk
	2	light, new age, classical,
Mood	0	exciting, wild, happy
	1	sad, missing, lonely
	2	fresh, sanative, easy

the group feature. Table 1 reports the clusterings and group interpretations. In the music cultural background of China, genre schema is straightforward and reasonable. For example, in terms of rhythm, users in group 0 (rap, dance and alternative) prefer music with a quick rhythm. Users in group 2 (light, new age and classical) always listen to slow and soothing music. Hence, the groups of each clustering could be well explained by the top attributes, suggesting that our methods indeed capture the music preferences and that behavioral patterns in music listening can be effectively detected. People who love pop and folk music do not seemly have obvious preferences as to rhythm. Moreover, we also attempt to classify the users by other solutions such as DBSCAN, a popular algorithm that does not need the number of clusters in advance. However, the sizes of most groups are very small and even less than 10. Hence, to obtain a relatively stable and even distribution of cluster sizes, DBSCAN is not appropriate in our study.

In the meantime, as exploited in previous work (Van Eijck, 2001; Park et al., 2015), diversity of musical preferences is also an indicator of great significance in reflecting listening behaviors. We draw on the concept of entropy (Alexander, 1996) to define users' diversity in music preferences as

$$\text{div} = -\frac{1}{3}\sum_{c=1}^{3}\sum_{i=1}^{n} p(u_{c,i})\log(p(u_{c,i})), \tag{5}$$

in which $p(u_{c,i})$ refers to the proportion of the i-th attribute in the user vector of trait c. Higher diversity implies that the user has more appreciation for a variety of experience in music listening. We then employ K-means to cluster active users into high-diversity and low-diversity groups.

In summary, Table 2 shows the complete description of user classification, including the size of each group, centroids and performance of clustering. We use K-means clustering to discretize the continuous variables and vectors, hence the users could be grouped into different types with labels. All the variables can be converted into discrete value (factor value). We could further analyze the relationship between factors and listening similarities since individuals obtain their factor value. As can be seen in Table 2, the average distance of samples to closest centroid is significantly smaller than distances between centroids.

There are a total of 19 factors[1] in eight traits. As a dummy variable, the value of the factor is 1 when two linked users are classified into the same group; otherwise, the value is 0. For example, factors *female* is 1 and *male* is 0 when user A (female) is linked with user B (female). If *female* and *male* are both 0, this indicates that A and B possess different genders. We could further analyze the relationship between the above factors and listening similarities in online social networks.

4. Results

4.1. Existence of homophily

Being direct indicators of homophily, similarities from different perspectives are first investigated in the two social networks constructed from Netease and Weibo. Meanwhile, to testify to the significance of the similarity distribution, in terms of shuffling nodes 10 times, we build 10 random Weibo networks and 10 random Netease Music networks as baselines.

The CCDF of sim_{song}, as seen in Fig. 4, demonstrates that the similarities of actual networks are evidently higher than those of random networks, particularly when $sim_{song} > 5$, indicating that friends in online social network are inclined to listen to the same songs and homophily in music listening significantly exists. In particular, compared to the Weibo network, friendships in the Netease network are

[1] The factors range from demographics to music traits and include *female*, *male*, *V*0, *V*1, *NER*0, *NER*1, *RFF*0, *RFF*1, *lang*0, *lang*1, *lang*2, *genre*0, *genre*1, *genre*2, *mood*0, *mood*1, *mood*2, *div*0 and *div*1.

Table 2
User classification: size of groups and the centroids of K-means.

	Group 0		Group 1		Group 2		Average distances of samples to closest centroid	Average distances of centroids
	size	cluster center	size	cluster center	size	cluster center		
Gender (female and male)	15,565	–	10,388	–	–			
Verified (non- and verified)	24,368	–	1585	–	–			
NER (low and high)	18,754	142.52	7199	2025.04	–		431.55	1882.52
RFF (low and high)	22,304	0.62	3649	14.27	–		3.05	13.65
Language	10,198	(0.90 −0.64 −0.35 −0.20 0.54 −0.20)	4023	(−0.90 −0.70 1.63 1.03 −0.51 −0.24)	11,732	(−0.68 1.01 −0.24 −0.17 −0.41 0.33)	1.49	2.84
Genre	7624	(−0.68 0.50 −0.30 0.70 0.28 −0.23 0.48 0.21 0.31 −0.12 −0.13 0.68 0.40 0.48 0.48 0.37 −0.04 0.23 0.70 −0.12 −0.34 0.32 0.34 0.59)	14,476	(0.55 −0.13 0.30 −0.39 −0.04 −0.31 −0.23 −0.04 −0.03 −0.20 −0.13 −0.25 −0.20 −0.17 −0.19 −0.16 −0.31 −0.15 −0.31 −0.32 −0.03 −0.17 −0.19 −0.27)	3853	(−0.73 −0.48 −0.51 0.07 −0.38 1.62 −0.08 −0.24 −0.51 1.00 0.75 −0.41 −0.04 −0.29 −0.22 −0.12 1.28 0.11 −0.20 1.46 0.81 0.01 0.05 −0.15)	2.56	3.78
Mood	5746	(−0.73 −0.16 −0.17 1.11 −1.01 −0.89 0.72 −0.64 −0.79 1.27 0.82 −0.95 −1.07)	11,008	(0.69 −0.37 −0.14 −0.49 0.77 0.01 −0.73 0.61 0.25 −0.59 −0.64 0.43 0.74)	9199	(−0.37 0.55 0.28 −0.10 −0.29 0.54 0.42 −0.33 0.18 −0.09 0.25 0.07 −0.22)	2.42	3.63
Diversity (low and high)	18,835	7.64	7118	51.01	–		19.65	43.37

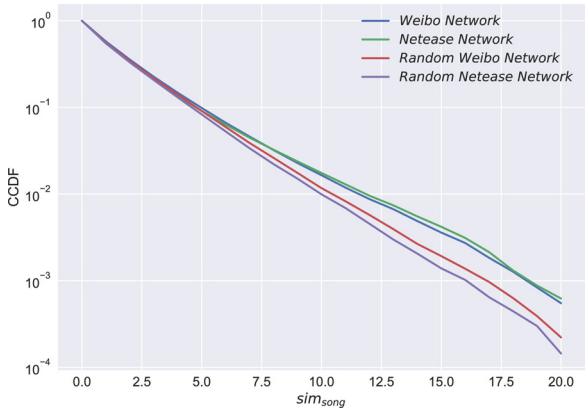

Fig. 4. Distribution of similarities of historical listening records. Through the two-sample Kolmogorov–Smirnov test, except for the two random networks ($p = 0.99$), the distributions of other pairs have significant differences ($p < 0.001$, ***).

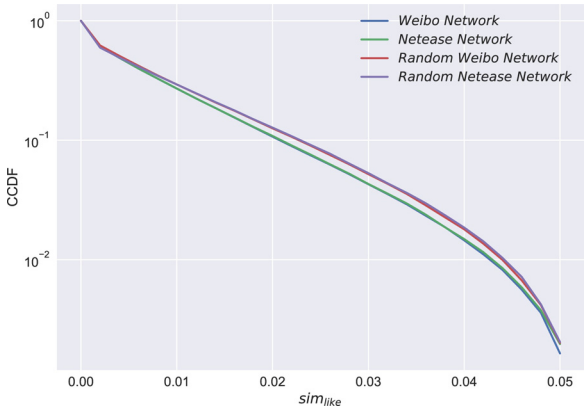

Fig. 6. Distribution of similarities of favorite songs between users. Through the two-sample Kolmogorov–Smirnov test, except for the two random networks ($p = 0.163$), the distributions of other pairs have significant differences ($p < 0.001$, ***).

slightly closer, implying that from the perspective of listening to the same songs, homophily is enhanced in the music-oriented social network.

The CCDFs of similarities in perspectives of music preferences are further investigated, as seen in Fig. 5. Similar to observations of sim_{song}, it is consistently demonstrated that the friend similarity in music preferences is significant, suggesting the existence of homophily in music listening for online social networks. Even more interesting, homophily in the Netease network is more evident than that in the Weibo network, implying again that friends are more homogeneous in domain-oriented social networks than their generic counterparts. However, as for the measure of sim_{like} shown in Fig. 6, there is no evidence for the existence of homophily because the similarity of friends in the 'like' behavior of online networks is close to or even lower than that in random networks. We also testify significance of the similarity distribution (CCDF) between two actual networks and two random networks by two-sample

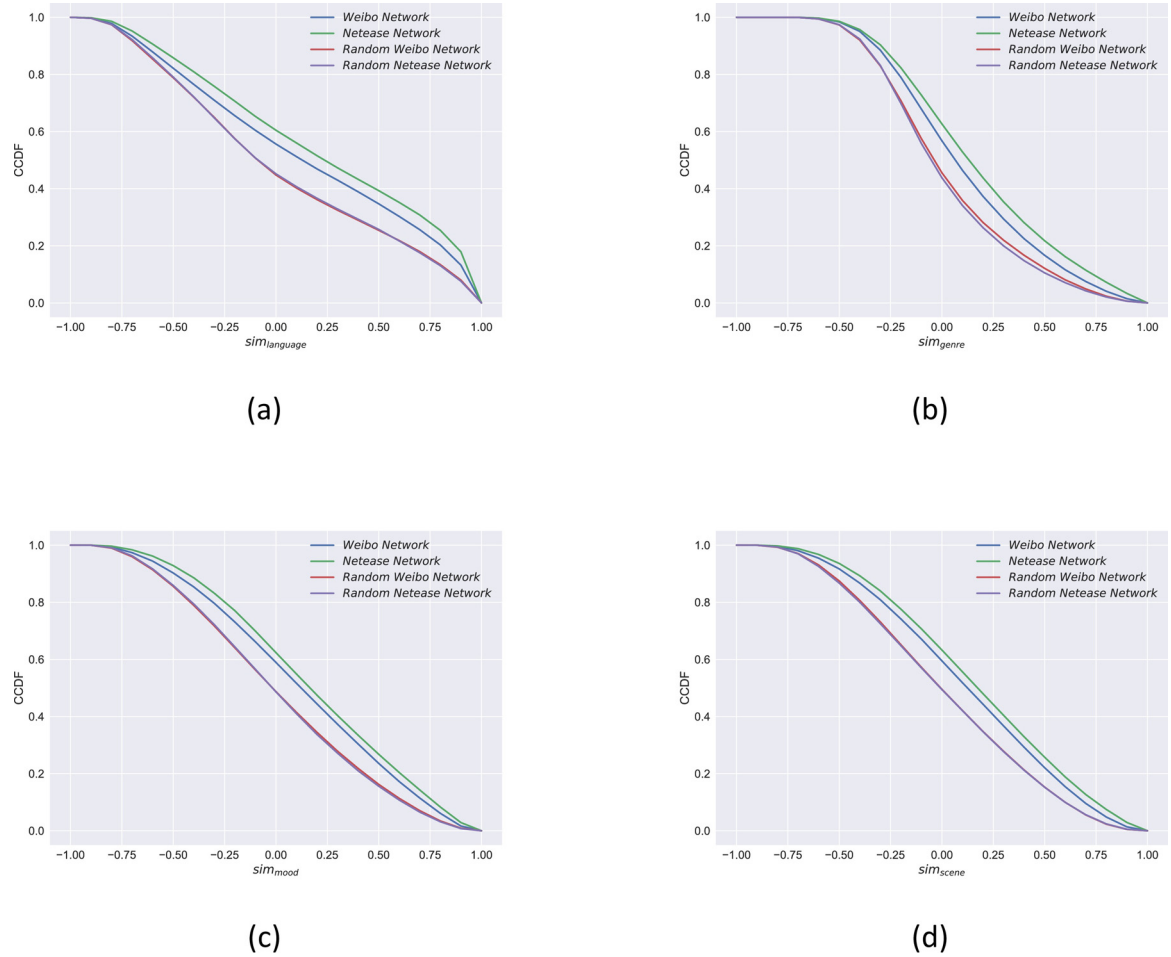

Fig. 5. Distribution of similarities of musical preferences. Through the two-sample Kolmogorov–Smirnov test, except for the two random networks ($p > 0.05$), the distributions of other pairs have significant differences ($p < 0.001$, ***).

Kolmogorov–Smirnov test, which are shown in captions of Figs. 4–6.

To summarize, in terms of similarity metrics, except the one regarding the 'like' behavior, the homophily of music listening in online social networks, particularly the music-oriented one, is significantly demonstrated and confirmed. Friends connected in online social networks indeed enjoy identical music and share similar music preferences. The absence of homophily in the 'like' behavior suggests that even in homogeneous online social networks, individual differences still exist across friends. Although the homophily in specialized networks is slightly more significant, we indeed find the same behavior pattern in generic networks such as Weibo, where the relationships are bonded with more common interests. This result further implies that social media, such as Weibo, can be generic but typical instances of online social networks. We further explore the important factors that affect the listening similarity in the Weibo network.

4.2. Critical factors for homophily

Given the rich backgrounds of active users carried by Weibo, we further investigate how characteristics of social ties, i.e., factors, influence the homophily disclosed in music listening. Except for sim_{song}, other similarities referring to music preference are measured with U_c (trait vectors of users); however, the vectors are also used to cluster the individuals. Errors could occur when the dependent variable and some independent variables are calculated from the same source. Therefore, we consider only sim_{song} as the dependent variable to predict in regression analysis. As seen, Table 3 presents the multiple regression results of the nineteen factors' function on the homophily.

As for gender, the coefficients of female are higher than those of male, indicating that relationships in online social networks between female users are more homogeneous in music listening, but the correlations between listening similarities and the verification statuses are not significant due to trivial coefficients. For social attributes *NER* and *RFF*, which reflect individual extraversion and openness (Amichai-Hamburger and Vinitzky, 2010; Bachrach et al., 2012), the coefficient of *NER*0 (0.444) is significantly larger than that of *NER*1 (−0.149).

Table 3
Results of multiple regression to predict sim_{song} based on user classifications. The column of "Coef." (coefficient) shows the influence of factors on homophily.

Traits	Factors	Coef.	Std. Error	*t*-Value	$Pr(>\|t\|)$
gender	female	0.369	0.021	17.25	***
	male	0.254	0.015	16.91	***
verified	V0	0.119	0.017	6.91	***
	V1	0.014	0.035	0.41	
NER	NER0	0.444	0.042	10.58	***
	NER1	-0.149	0.018	-8.49	***
RFF	RFF0	0.517	0.031	16.72	***
	RFF1	0.036	0.026	1.35	
language	lang0	1.047	0.026	40.43	***
	lang1	0.569	0.036	15.79	***
	lang2	0.228	0.020	11.28	***
genre	genre0	-0.023	0.022	-1.05	
	genre1	1.043	0.020	50.90	***
	genre2	0.061	0.041	1.51	
mood	mood0	0.668	0.024	27.71	***
	mood1	0.359	0.027	13.32	***
	mood2	0.126	0.023	5.35	***
diversity	div0	0.770	0.015	50.59	***
	div1	0.208	0.023	8.89	***
	Intercept	0.396	0.019	20.50	***
	Observations	112,753			
	Residual standard error	2.305			
	Multiple R^2	0.1701			
	Adjusted R^2	0.1699			
	F-Statistic	1100			***

Note: $^{**}p < .01$, $^{***}p < .001$.

Since *RFF*1 is not significant in the regression analysis, we could not directly compare the coefficient of *RFF*0 with that of *RFF*1. However, we find that the coefficient of *RFF*0 (0.517) is positive and even larger than *NER*0. These results consistently imply that users of low extroversion and openness demonstrate more homophily in music listening.

For factors from the perspective of music preferences, the positive coefficients shown in Table 3 indicate that users within the same group of clusterings referring to language and mood preferences are similar. As for the language preference, users in the Chinese group are more similar than those in the other groups. Pop music and folk music are known as the mainstream genres in China. As for genre, it can be found that the users following the pop and folk genres are significantly more similar. According to the results for languages and genres, we suggest that friends in online social networks whose music preferences accord with the mass taste always listen to similar music. One of the possible origins is the popularity of music. For example, the BILLBOARD HOT 100, an indicator of the mass taste (Mauch et al., 2015), publishes 100 popular songs weekly and these users might be more likely to listen to these same top songs, resulting higher sim_{song} than that of other groups. From the perspective of moods, the coefficients of the factor 'exciting, wild and happy' are the highest for sim_{song}. This result demonstrates that people who like positive and energetic songs are more homogeneous in music listening. In addition, we investigate whether the diversity (*div*0 and *div*1) of music consumption is an influential factor for listening similarity. The results show that users with low musical diversity are more similar. However, for users who enjoy multiple types of music and always have uncommon and unique preferences, it is difficult for them to find friends sharing similar musical preferences.

5. Discussion

The flourishing of social media greatly facilitates exploitations of social networks by offering an unprecedented big-data window. In the present study, aiming at understanding the homophily of music listening in online social networks, two networks with the same group of active users from Netease Music and Weibo are separately built for probing six measures of homophily. It is significantly confirmed that homophily exists in music listening for online social networks, even for the generic one from Weibo. Factors of social ties ranging from demographics to music preferences are further investigated in terms of their influence on homophily, and many interesting patterns are revealed. To the best of our knowledge, for the first time, a systematic and comprehensive study of music homophily in China is performed in a data-driven solution manner.

Meanwhile, we are also the first to construct big music data by matching active users in both a music streaming platform and Weibo. The unexpected existing homophily in Weibo suggests that the knowledge from generic social networks can be confidently introduced and transferred to the study (Malhotra et al., 2012) and application (Carmagnola and Cena, 2009) of domain-oriented social networks, which will essentially enrich the context of subjects. Previous investigation shows that humans spend on average 17% of their lives listening to music (Rentfrow, 2012). Nevertheless, abundant behavioral data are still missing in existing studies. By combing behavioral and demographical data from different online social networks, our study demonstrates that explorations of music listening can be profoundly boosted.

In fact, music steaming platforms have been experiencing fast growth with increasing influence through the establishment of musical environments tailored to individual preferences. Plenty of music service providers use the relationships between online friends to estimate users' musical tastes and then recommend music (Zhong and Sastry, 2017). In our study, we provide solid evidence for the homophily of music listening both in specialized online social networks and generic networks, suggesting that regardless of whether networks are specialized or

generic social ones, they can serve as high quality sources of music recommendation (Simşek and Jensen, 2008; Bu et al., 2010). We argue that the homophily should be exploited in building and enhancing music services. Specifically, the influential factors that function in listening similarities can be directly introduced into the design of music recommendation algorithms, and the coefficients we obtained can be the principal weights for these factors. Taking gender as an example, music platforms could put more emphasis on female friends when recommending music to female users.

This study has inevitable limitations. For example, both social networks investigated here are just sampled in China, and the findings might not be directly extendable to other countries. Meanwhile, measures such as structural or equivalent similarities could also be indicators for reflecting homophily. Nevertheless, these limitations will be promising directions in our future work, and understanding the cultural influence on music listening homophily from more perspectives, including structural ones, will be our next step.

Acknowledgements

This work was supported by NSFC (Grant nos. 71501005, 71531001 and 61421003) and the fund of the State Key Lab of Software Development Environment (Grant no. SKLSDE-2017ZX-05).

References

Aiello, L.M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., Menczer, F., 2012. Friendship prediction and homophily in social media. ACM Trans. Web (TWEB) 6 (2), 9.
Alexander, P.J., 1996. Entropy and popular culture: product diversity in the popular music recording industry. Am. Sociol. Rev. 61 (1), 171–174.
Amichai-Hamburger, Y., Vinitzky, G., 2010. Social network use and personality. Comput. Hum. Behav. 26 (6), 1289–1295.
Antheunis, M.L., Valkenburg, P.M., Peter, J., 2012. The quality of online, offline, and mixed-mode friendships among users of a social networking site. Cyberpsychol.: J. Psychosoc. Res. Cybersp. 6 (3).
Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D., 2012. Personality and patterns of Facebook usage. ACM Web Science Conference 24–32.
Baltar, F., Brunet, I., 2012. Social research 2.0: virtual snowball sampling method using Facebook. Internet Res. 22 (1), 57–74.
Baym, N.K., Ledbetter, A., 2009. Tunes that bind? Predicting friendship strength in a music-based social network, Information. Commun. Soc. 12 (3), 408–427.
Bisgin, H., Agarwal, N., Xu, X., 2010. Investigating homophily in online social networks. In: International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM, vol. 1. IEEE. pp. 533–536.
Blacking, J., 1995. Music, Culture, and Experience: Selected Papers of John Blacking. University of Chicago Press.
Boer, D., Fischer, R., Strack, M., Bond, M.H., Lo, E., Lam, J., 2011. How shared preferences in music create bonds between people: values as the missing link. Personal. Soc. Psychol. Bull. 37 (9), 1159–1171.
Bollen, J., Gonçalves, B., Ruan, G., Mao, H., 2011. Happiness is assortative in online social networks. Artif. Life 17 (3), 237–251.
Boström, T., Eliasson, S., Lindtorp, P., Moioli, F., Nyström, M., 1999. Mobile audio distribution. Pers. Technol. 3 (4), 166–172.
Brown, J., Broderick, A.J., Lee, N., 2007. Word of mouth communication within online communities: conceptualizing the online social network. J. Interact. Market. 21 (3), 2–20.
Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., He, X., 2010. Music recommendation by unified hypergraph: combining social media information and music content. ACM International Conference on Multimedia 391–400.
Buote, V.M., Wood, E., Pratt, M., 2009. Exploring similarities and differences between online and offline friendships: the role of attachment style. Comput. Hum. Behav. 25 (2), 560–567.
Burt, R.S., 1982. Toward a Structural Theory of Action: Network Models of Social Structure, Perception, and Action.
Carmagnola, F., Cena, F., 2009. User identification for cross-system personalisation. Inf. Sci. 179 (1), 16–32.
Centola, D., 2010. The spread of behavior in an online social network experiment. Science 329 (5996), 1194–1197.
Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., Hołyst, J.A., 2011. Negative emotions boost user activity at bbc forum. Phys. A: Stat. Mech. Appl. 390 (16), 2936–2944.
Costa Jr., P.T., Terracciano, A., McCrae, R.R., 2001. Gender differences in personality traits across cultures: robust and surprising findings. J. Personal. Soc. Psychol. 81 (2), 322.
De Klepper, M., Sleebos, E., Van de Bunt, G., Agneessens, F., 2010. Similarity in friendship networks: selection or influence? The effect of constraining contexts and non-visible individual attributes. Soc. Netw. 32 (1), 82–90.
Fan, R., Zhao, J., Chen, Y., Xu, K., 2014. Anger is more influential than joy: sentiment correlation in Weibo. PLOS ONE 9 (10), e110184.
Frith, S., 2002. Music and everyday life. Crit. Q. 44 (1), 35–48.
Fung, A.Y., 2007. The emerging (national) popular music culture in china. Inter-Asia Cult. Stud. 8 (3), 425–437.
Gao, Q., Abel, F., Houben, G.-J., Yu, Y., 2012. A comparative study of users' microblogging behavior on Sina Weibo and twitter. International Conference on User Modeling, Adaptation, and Personalization. Springer, pp. 88–101.
Gefen, D., Straub, D.W., 1997. Gender differences in the perception and use of e-mail: an extension to the technology acceptance model. MIS Q. 389–400.
Greenberg, D.M., Baron-Cohen, S., Stillwell, D.J., Kosinski, M., Rentfrow, P.J., 2015. Musical preferences are linked to cognitive styles. PLOS ONE 10 (7), e0131151.
Greenberg, D.M., Kosinski, M., Stillwell, D.J., Monteiro, B.L., Levitin, D.J., Rentfrow, P.J., 2016. The song is you: preferences for musical attribute dimensions reflect personality. Soc. Psychol. Personal. Sci. 7 (6), 597–605.
Håkansson, M., Rost, M., Holmquist, L.E., 2007. Gifts from friends and strangers: a study of mobile music sharing. In: ECSCW 2007. Springer. pp. 311–330.
Hagen, A.N., Lüders, M., 2017. Social streaming? Navigating music as personal and social. Convergence 23 (6), 643–659.
Heckathorn, D.D., 1997. Respondent-driven sampling: a new approach to the study of hidden populations. Soc. Prob. 44 (2), 174–199.
Henning, V., Reichelt, J., 2008. Mendeley-a last. fm for research? In: IEEE Fourth International Conference on eScience'08. IEEE. pp. 327–328.
Holmquist, L.E., 2005. Ubiquitous music. Interactions 12 (4), 71.
Kreitz, G., Niemela, F., 2010. Spotify-large scale, low latency, p2p music-on-demand streaming. In: IEEE Tenth International Conference on Peer-to-Peer Computing (P2P). IEEE. pp. 1–10.
Kwak, H., Lee, C., Park, H., Moon, S., 2010. What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. ACM. pp. 591–600.
Lambiotte, R., Ausloos, M., 2006. On the genre-fication of music: a percolation approach. Eur. Phys. J. B Condens. Matter Comp. Syst. 50 (1-2), 183–188.
Lee, S.H., Kim, P.-J., Jeong, H., 2006. Statistical properties of sampled networks. Phys. Rev. E 73 (1), 016102.
Malhotra, A., Totti, L., Meira Jr., W., Kumaraguru, P., Almeida, V., 2012. Studying user footprints in different online social networks. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM. IEEE. pp. 1065–1070.
Mauch, M., MacCallum, R.M., Levy, M., Leroi, A.M., 2015. The evolution of popular music: USA 1960–2010. R. Soc. Open Sci. 2 (5), 150081.
Mcpherson, M., Smithlovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27 (1), 415–444.
Miell, D., MacDonald, R.A., Hargreaves, D.J., 2005. Musical Communication. Oxford University Press on Demand.
Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B., 2007. Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM. pp. 29–42.
Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P., 2010. You are who you know: inferring user profiles in online social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM. pp. 251–260.
Newman, M., 2010. Networks: An Introduction. Oxford University Press.
News, N., 2016. Analysis Report of Netease Music in 2016. (accessed August, 2016). http://www.cbdio.com/BigData/2016-08/25/content_5213870.htm.
North, A.C., 2004. Uses of music in everyday life. Music Percept. Interdiscip. J. 22 (1), 41–77.
Park, M., Weber, I., Naaman, M., Vieweg, S., 2015. Understanding musical diversity via online social media. ICWSM 308–317.
Priest, E., 2006. The future of music and film piracy in china. Berkeley Technol. Law J. 795–871.
Reagans, R., 2005. Preferences, identity, and competition: predicting tie strength from demographic data. Manag. Sci. 51 (9), 1374–1383.
Rentfrow, P.J., 2012. The role of music in everyday life: Current directions in the social psychology of music. Soci. Personal. Psychol. Compass 6 (5), 402–416.
Robazza, C., Macaluso, C., D'Urso, V., 1994. Emotional reactions to music by gender, age, and expertise. Percept. Motor skills 79 (2), 939–944.
Rondeau, D.B., 2005. For mobile applications, branding is experience. Commun. ACM 48 (7), 61–66.
Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLOS ONE 8 (9), e73791.
Selfhout, M.H., Branje, S.J., ter Bogt, T.F., Meeus, W.H., 2009. The role of music preferences in early adolescents' friendship formation and stability. J. Adolesc. 32 (1), 95–107.
Simşek, O., Jensen, D., 2008. Navigating networks by using homophily and degree. Proc. Natl. Acad. Sci. U. S. A. 105 (35), 12758.
Sloboda, J.A., 1999. Everyday uses of music listening: a preliminary study. Music, Mind Sci. 354–369.
Van Eijck, K., 2001. Social differentiation in musical taste patterns. Soc. Forces 79 (3), 1163–1185.
Voida, A., Grinter, R.E., Ducheneaut, N., Edwards, W.K., Newman, M.W., 2005. Listening in: practices surrounding itunes music sharing. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems. ACM. pp. 191–200.
Wells, A., Hakanen, E.A., 1991. The emotional use of popular music by adolescents. J. Q. 68 (3), 445–454.
Wright, C., 2013. Listening to Music. Nelson Education.

Youyou, W., Stillwell, D., Schwartz, H.A., Kosinski, M., 2017. Birds of a feather do flock together: behavior-based personality-assessment method reveals personality similarity among couples and friends. Psychol. Sci. 28 (3), 276–284.

Zhong, C., Sastry, N., 2017. Systems applications of social networks. ACM Comput. Surv. 50 (5) 63:1–63:42.