"Know Thyself" How Personal Music Tastes Shape the Last.Fm Online Social Network

Riccardo Guidotti[1] and Giulio Rossetti[1]

ISTI-CNR, Via G. Moruzzi, Pisa, Italy
name.surname@isti.cnr.it

Abstract. As Nietzsche once wrote *"Without music, life would be a mistake"*[1]. The music we listen to reflects our personality, our way to approach life. In order to enforce self-awareness, we devised a Personal Listening Data Model that allows for capturing individual music preferences and patterns of music consumption. We applied our model to 30k users of Last.Fm for which we collected both friendship ties and multiple listening. Starting from such rich data we performed an analysis whose final aim was twofold: *(i)* capture, and characterize, the individual dimension of music consumption in order to identify clusters of like-minded Last.Fm users; *(ii)* analyze if, and how, such clusters relate to the social structure expressed by the users in the service. Do there exist individuals having *similar* Personal Listening Data Models? If so, are they directly connected in the social graph or belong to the same community?

Keywords: Personal Data Model · Online Social Network · Music

1 Introduction

Music consumption is one of the activities that better reflects human personality: each one of us has her own tastes and habits when talking about music. In recent history, the World Wide Web revolution has deeply changed the way music enters in our daily routine. Online giants like Spotify, iTunes, SoundCloud have made huge accessible catalogs of music products to everybody everywhere.

We propose a *Personal Listening Data Model* (*PLDM*) able to capture the characteristics and systematic patterns describing music listening behavior. PLDM is built on a set of personal listening: a listening is formed by the song listened, the author of the song, the album, the genre and by the listening time. PLDM summarizes each listener behavior, explains her music tastes and pursues the goal of providing *self-awareness* so as to fulfill the Delphic maxim *"know thyself"*.

However, listening music is not only an individual act but also a social one. This second nature of music consumption is the stone on which several online services pose their grounds. Among them, one of the most famous is *Last.Fm*. On such platform, users can build social ties by following peer listeners. The social

[1] Twilight of the Idols, 1889.

network that arises from such a process represents highly valuable information. On such structure, artists/tracks/album adoptions give birth to a social-based recommender system in which each user is exposed to the listening of her friends.

It has been widely observed how homophily [19, 26] often drives implicitly the rising of social structures encouraging individuals to establish ties with like-minded ones. Does music taste play the role of social glue in the online world? To answer such a question, we combine individual and group analysis, and we propose a way to characterize communities of music listeners by their preferences and behaviors.

The paper is organized as follows. Section 2 surveys works related to personal data model and Last.Fm online social network. Section 3 describes our model for analyzing musical listening and the relationship with friends. In Section 4 are presented the individual and social analysis performed on a dataset of 30k Last.Fm users. Finally, Section 5 summarizes conclusion and future works.

2 Related Work

The analysis of music listening is becoming increasingly valuable due to the increasing attention the music world is receiving from the scientific community.

Several works have analyzed data regarding online listening in order to model diffusion of new music genres/artists, as well as to analyze the behaviors and tastes of users. In [24] the authors identified through factor analysis three patterns of preference associated with liking for most types of *Rock Music*, general *Breadth of Musical Preference*, and liking for *Popular Music*. Also [25] examined individual differences in music preferences, and preferences for distinct music dimensions were related to various personality dimensions. In [6] was proposed a music recommendation algorithm by using multiple social media information and music acoustic-based content. In [4], the authors, studied the topology of the Last.Fm social graph asking for similarities in taste as well as on demographic attributes and local network structure. Their results suggest that users connect to "online" friends, but also indicate the presence of strong "real-life" friendship ties identifiable by the multiple co-attendance of the same concerts. The authors of [22] measured different dimensions of social prominence on a social graph built upon 70k Last.Fm users whose listening were observed for 2 years. In [23] was analyzed the cross-cultural gender differences in the adoption and usage of Last.Fm. Using social media data, the authors of [21] designed a measure describing the diversity of musical tastes and explored its relationship with variables that capture socioeconomic, demographics, and traits such as openness and degree of interest in music. In [32] is shown how to define statistical models to describe patterns of song listening in an online music community. In [13] is shown how the usage of a personal listening data model (also exploited in this work) can provide a high level of self-awareness and to enable the development of a wide range of analysis exploited here with social network analysis measures.

The access to this huge amount of data generates novel challenges. Among them, the need to handle efficiently individual data is leading to the development of personal models able to deal and summarize human behavior.

These data models can be generic or specific with respect to the type of data. In [20] is described *openPDS*, a personal metadata management framework that allows individuals to collect, store, and give fine-grained access to their metadata to third parties. [31] described *My Data Store* a tool that enables people to control and share their personal data. My Data Store has been integrated in [30] into a framework enabling the development of trusted and transparent services. Finally, in [1] is proposed that each user can select which applications have to be run on which data enabling in this way diversified services on a personal server. The majority of works in the literature focus their attention on the architecture of the personal data store and on how to treat data sharing and privacy issues. The main difference between the personal data model proposed and those present in the literature is that our model focuses in obtaining an added value from the personal data through the application of data mining techniques.

In this work, we propose to apply the methodological framework introduced in [16] for mobility data to analyze personal musical preferences. An application of this approach in mobility data and transactional data can be found in [29, 11, 17]. Moreover, in [14] is shown haw the network component becomes fundamental to leverage the power of the analysis from the personal level to the collective ones. User experience in online social media services, however, is composed not only of individual activities but also of interactions with other peers. The role of social communities and friendship ties is, for sure important to understand the factors that drive the users' engagement toward an artist/product. In order to assess the strength of social influence measures based on *homphily* [19] and on common interests have long been applied in social networks. For instance, the structure of ego-networks and homophily on Twitter was studied in [3] where the authors investigated the relations between homophily and topological features discovering a high homophily w.r.t. topics of interest. The authors of [2] exploited homophily in latent attributes to augment the users' features with information derived from Twitter profiles and from friends' posts. Their results suggest that the neighborhood context carries a substantial improvement to the information describing a user. To the best of our knowledge, this work is the first attempt to define a data model able to capture musical listening behavior and to use it to analyze the relationships in the social network.

3 Personal Listening Data Model

In this section, we formally describe the *Personal Listening Data Model*. By applying the following definitions and functions, it is possible to build for each user a listening profile giving a picture of her habits in terms of listening.

Definition 1 (Listening). *Given a user u we define $L_u = \{\langle$time-stamp, song, artist, album, genre$\rangle\}$ as the set of listening performed by u.*

Fig. 1: A listening $l = \{\langle time\text{-}stamp, song, artist, album, genre\rangle\}$ is a tuple formed by the *time-stamp* indicating when the listening was performed, the *song* listened, the *artist* which played the song, the *album* the song belongs to and the *genre* of the artist.

Each listening l (see Fig. 1) is an abstraction of a real listening since a song can belong to more than a genre and can be played by more than an artist[2]. However, we can assume this abstraction without losing in generality.

From the set of listening L_u we can extract the set of songs S_u, artists A_u, albums B_u and genres G_u for each user. More formally we have:

- $S_u = \{song | \langle \cdot, song, \cdot, \cdot, \cdot \rangle \in L_u\}$
- $A_u = \{artist | \langle \cdot, \cdot, artist, \cdot, \cdot \rangle \in L_u\}$
- $B_u = \{album | \langle \cdot, \cdot, \cdot, album, \cdot \rangle \in L_u\}$
- $G_u = \{genre | \langle \cdot, \cdot, \cdot, \cdot, genre \rangle \in L_u\}$

Besides the sizes of these sets, a valuable summary of the user behavior can be realized through frequencies dictionary indicating the support (i.e. the relative number of occurrences) of each feature of the listening.

Definition 2 (Support). *The* support *function returns the frequency dictionary of (item, support) where the support of an item is obtained as the number of occurring items on the number of listening.*

$$sup(X, L) = \{(x, |Y|/|L|) | x \in X \wedge Y \subseteq L s.t. \forall l \in Y, x \in l\}$$

We define the following frequency dictionaries: $s_u = sup(S_u, L_u)$, $a_u = sup(A_u, L_u)$, $b_u = sup(B_u, L_u)$, $g_u = sup(G_u, L_u)$, $d_u = sup(D, L_u)$ and $t_u = sup(T, L_u)$ where $D = \{mon, tue, wed, thu, fri, sat, sun\}$ contains the days of weeks, and $T = \{(2\text{-}8], (8\text{-}12], (12\text{-}15], (15\text{-}18], (18\text{-}22], (22\text{-}2]\}$ contains the time slots of the day (i.e. early and late morning, early and late afternoon, early and late night).

These dictionaries can be exploited to extract indicators.

Definition 3 (Entropy). *The* entropy *function returns the normalized entropy in* $[0, 1]$ *of a dictionary* x*. It is defined as:*

$$entropy(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i) / \log_b n$$

The entropy tends to 0 when the user behavior with respect to the observed variable is systematic, tends to 1 when the behavior is not predictable. These

[2] The choice of describing a listening with these attributes is related to the case study. Additional attributes can be used when available from the data. We highlight that listening means that the song was played and not necessarily entirely listened.

indicators are similar to those related to shopping behavior described in [10, 12]. We define the entropy for songs, artists, albums, genres, days and time-slots as $e_{s_u} = entropy(g_u)$, $e_{a_u} = entropy(a_u)$, $e_{b_u} = entropy(b_u)$, $e_{g_u} = entropy(g_u)$, $e_{d_u} = entropy(d_u)$ and $e_{t_u} = entropy(t_u)$.

A pattern we consider is the top listened artist, genre, etc.

Definition 4 (Top). *The* top *function returns the most supported item in a dictionary. It is defined as:*

$$top(X) = \underset{(x,y)\in X}{argmax}(y)$$

We define the top for songs, artists, albums and genres as $\hat{s_u} = top(s_u)$, $\hat{a_u} = top(a_u)$, $\hat{b_u} = top(b_u)$ and $\hat{g_u} = top(g_u)$.

Moreover, we want to consider for each user the set of representatives, i.e. significantly most listened, subsets of artists, albums, and genres.

Definition 5 (Repr). *The* repr *function returns the most representative supported items in a dictionary. It is defined as:*

$$repr(X) = \underset{(x,y)\in X}{knee}(y)$$

The result of $repr(X)$ contains a set of preferences such that their support is higher than the support of most of the listening done with respect to other artists, albums, and genres. For example if user u has $g_u = \{(rock, 0.4), (pop, 0.3), (folk, 0.1), (classic, 0.1), (house, 0.1)\}$. Then the result of $repr(g_u)$ will bet $\{(rock, 0.4), (pop, 0.3)\}$. This result is achieved by employing a technique known as "knee method" [28] represented by the function $knee(\cdot)$. It sorts the vector according to the supports, and it returns as most representative the couples with support greater or equal than the support corresponding to the *knee* in the curve of the ordered frequencies. We define the most representative for songs, artists, albums and genres as $\tilde{s_u} = repr(s_u)$, $\tilde{a_u} = repr(a_u)$, $\tilde{b_u} = repr(b_u)$ and $\tilde{g_u} = repr(g_u)$. Obviously we have $\hat{g_u} \subseteq \tilde{g_u} \subseteq g_u$ that holds also for songs, albums and artists.

Finally, in order to understand how each user is related with her *friends* in terms of preferences we define the set of friends of a user u as $f_u = \{v_1, \ldots, v_n\}$ where $\forall v_i \in U,\ v_i \in f_u$. The ego-network of each user u is modeled by f_u.

By applying the definitions and the functions described above on the user listening L_u we can turn the raw listening data of a user into a complex personal data structure that we call *Personal Listening Data Model* (PLDM). The PMDL characterizes the listening behavior of a user by means of its *indicators*, *frequencies* and *patterns*.

Definition 6 (Personal Listening Data Model). *Given the listening L_u we define the* personal listening data model *as*

$$
\begin{aligned}
P_u = \langle & |L_u|, |S_u|, |A_u|, |B_u|, |G_u|, && \textit{indicators} \\
& e_{s_u}, e_{a_u}, e_{b_u}, e_{g_u}, e_{d_u}, e_{t_u}, && \textit{indicators} \\
& s_u, a_u, b_u, g_u, d_u, t_u, && \textit{frequencies} \\
& \hat{s_u}, \hat{a_u}, \hat{b_u}, \hat{g_u}, && \textit{patterns} \\
& \tilde{s_u}, \tilde{a_u}, \tilde{b_u}, \tilde{g_u}, && \textit{patterns} \\
& f_u \rangle && \textit{friends}
\end{aligned}
$$

It is worth to notice that, that according to the procedures followed in [18, 15], the PLDM can be extracted by following a parameter-free approach.

4 LastFM Case Study

In this section, we discuss the benefits derivable from using PLDM while analyzing the data extracted from a famous music-related online social network: *Last.Fm*. In Last.Fm people can share their own music tastes and discover new artists and genres on the bases of what they, or their friends, like. In such a service, each user produces is characterized by two elements: the social structure it is embedded in and her own listening. Through each listening, a user expresses a preference for a certain song, artist, album, genre, and take place in a certain time. Using Last.Fm APIs[3] we retrieved the last 200 listening, as well as the social graph $G = (U, E)$, of about $30,000$ users resident in the UK[4]. For each user $u \in U$, given the listening L_u we calculated her PLDM P_u. Using such individual model, we then performed a two-stage analysis aimed at: *(a)* describing how Last.Fm users can be characterized, 4.1, and *(b)* analyzing if, and how, the Last.Fm social structure reflects homophily behaviors, 4.2.

4.1 Who I am? PLDM Analysis.

The first analysis we report is related to the *indicators* of the PLDMs $\{P_u\}$ extracted. In Fig. 2 are reported the distributions of the number of users which have listened a certain number of songs $|S_u|$, artists $|A_u|$, albums $|B_u|$ and genres $|G_u|$. The first distribution is right-skewed with most of the users who have listened to about 140 songs (this implies that some tracks have been listened more than once). On the other hand, the other distributions are left-skewed: a typical user listened to about 60 artists, 70 albums and 10 genres.

Fig. 3 depicts the distributions of the entropy. It emerges that users are much more systematic with respect to the listening time (day of week and time of the

[3] `http://www.last.fm/api/`, retrieval date 2016-04-04

[4] The code, along with the ids of seed users used in this study, is available at `https://github.com/GiulioRossetti/LastfmProfiler`. The complete dataset is not released to comply with Last.fm TOS.

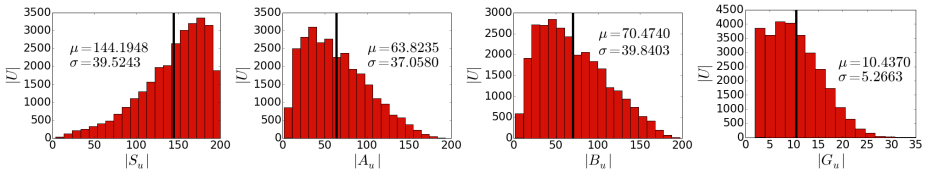

Fig. 2: Distributions of the number of songs $|S_u|$, artists $|A_u|$, albums $|B_u|$ and genres $|G_u|$ respectively. The black vertical lines highlight the means.

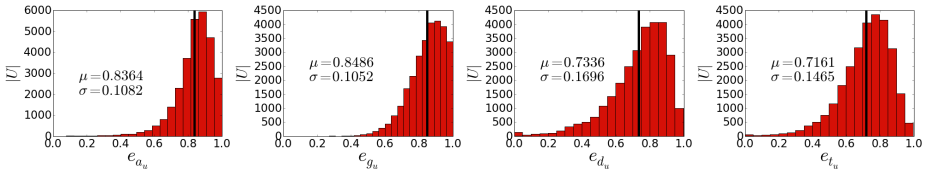

Fig. 3: Distributions of entropy for artists e_{a_u}, genre e_{g_u}, day of week e_{d_u} and time of day e_{t_u} respectively. The black vertical lines highlight the means.

day) than with respect to what they listen to. This behavior is in opposition to what happens in shopping [10]. Apparently, since the artist and genre entropy are right-skewed, it seems that most of the users are not very systematic with respect to the genre or to the artist. This can indicate that it is very unlikely that it exists a unique prevalence towards a certain artist or genre.

In Fig. 4 (left), we observe the heat-map of the correlations among the indicators. Some of them like $|A_u|$, $|B_u|$, $|G_u|$ are highly correlated[5] ($cor(|A_u|, |B_u|) = 0.8569, cor(|G_u|, |B_u| = 0.6358)$): the higher the number of artists or genres, the higher the number of albums listened. Other interesting correlations are $cor(|B_u|, e_{g_u}) = -0.3275$ and $cor(|B_u|, e_{a_u}) = 0.5483$. Their density scatter plots are reported in Fig. 4 *(center, right)*: the higher the number of albums listened, the lower the variability with respect to the genre and the higher the variability with respect to the artists. From this result, we understand that a user listening to many different albums narrows its musical preferences toward a restricted set of genres and that it explores these genres by listening to various artists of this genre and not having a clear preference among these artists.

A user can get benefit from a smart visualization of the PLDM *indicators* obtaining a novel level of *self-awareness* of her listening behavior. For instance, a user could discover that is listening to a great variety of artists but that they all belong to the same genre and that she always listens to them following the same pattern of songs. A possible reaction could be to start her new listening with an unknown artist belonging to a different genre to enlarge her musical knowledge and discover if she really dislikes certain genres or just had never the occasion to listen to them. Moreover, due to the continuously growing size of the personal raw listening dataset, the PLDM can be recalculated in different time windows so that the user can observe changes and/or stability in the listening profile.

[5] The *p-value* is zero (or smaller than 0.000001) for all the correlations.

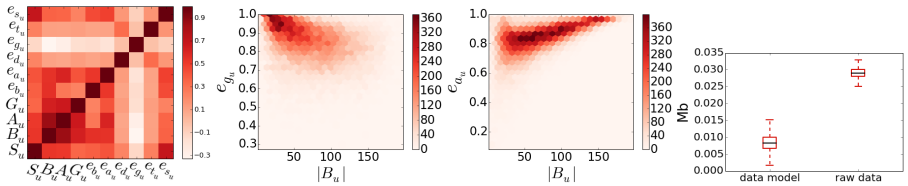

Fig. 4: (*left*) Correlation matrix; (*center-left*) Scatter density plots of number of albums $|B_u|$ versus genre entropy e_{g_u} and (*center-right*) versus artists entropy e_{a_u}; (*right*) Storage for the model.

PLDM efficiency, Storage Analysis. We report in Fig. 4(right) the boxplots of the storage occupancy of the data model PLDMs (left) and for the raw listening (right). The storage required by the data model is typically one third of the storage required by the raw data. Moreover, the storage space of the data model will not grow very much when storing more listening because the number of possible genres, artists, albums, songs is limited, while the number of listening grows continuously. Thus, an average storage of $0.01Mb$ together with a computational time of max 5 sec per user guarantees that the PLDM could be calculated and stored individually without the need for central service.

Frequency And Patterns Analysis. When dealing with music listening data, it is common to identify users by looking only to their most listened genre/artist. In order to prove that this assumption does not represent the users' preferences properly, we exploit the knowledge coming from the frequency vectors. We analyze the frequency vectors a_u, g_u, the top listened $\hat{a_u}$, $\hat{g_u}$, and the most representative $\tilde{a_u}$, $\tilde{g_u}$. In order to simplify the following discussion, we will refer to the sets $\tilde{a_u}$ and $\tilde{g_u}$ equivalently as $\tilde{x}$ and to the artists and genres contained in such sets as *preferences*. In Fig. 5 is depicted the result of this analysis for the genre (top row) and artist (bottom row)[6].

The first column shows the distribution of the number of users with respect to the number of representative genres $|\tilde{g_u}|$ and artists $|\tilde{a_u}|$. In both cases, the smallest value is larger than 1, indicating that each user has more than a preference. On the other hand, a large part of all the genres and artists listened are removed when passing from x to $\tilde{x}$. Indeed, the mean for the genres passes from 10 to 3, the mean for the artist passes from 60 to 10.

The second column in Fig. 5 illustrates the distribution of the number of users with respect to the maximum difference in frequencies between the listening preference obtained as $max(\tilde{x}) - min(\tilde{x})$. Both for genres and artists, the mode of this value is close to zero. This proofs that the highest preferences are similar in terms of listening for the majority of the users.

The third column shows the distributions of the users with respect to the most listened artist support *mas* and most listened genre support *mgs* given

[6] The analysis of b_u have similar results (not reported due to lack of space).

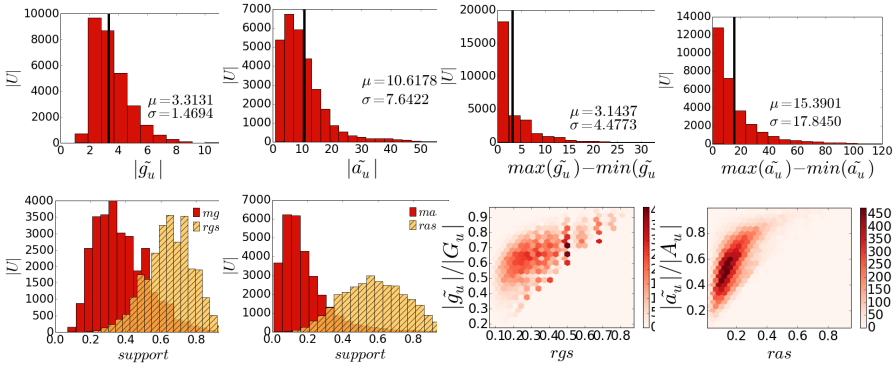

Fig. 5: Frequencies analysis for genre (left) and artist (right). *First row*: distribution of number of users w.r.t the number of representative preferences. *Second row*: distribution of number of users w.r.t the maximum difference in frequencies between the listening preference. *Third row*: distribution of number of users w.r.t the support given by the representative preferences. *Last row*: density scatter plot between the representative preferences support and the ratio of their number on the number of all the possible artists or genres.

by $mas = v|(a, v) = \hat{a_u}$ and $mgs = v|(g, v) = \hat{g_u}$ respectively, versus the representative artist support ras and representative genre support rgs given by $ras = sum(v|(a, v) \in \tilde{a_u})$ and $rgs = sum(v|(g, v) \in \tilde{g_u})$. From these distributions is evident the increase of the support when are considered, not only the preferred, but also all the representative preferences.

The last column reports a density scatter plot between the representative preferences support (rgs and the ras) and the ratio of their number on the number of all the artists or genres listened, i.e. $|\tilde{a_u}|/|A_u|$ and $|\tilde{g_u}|/|G_u|$ respectively. Since the higher concentration of the points is tends to be around 0.2 with respect to the x-axis and around 0.5 with respect to the y-axis we have that for most of the users it is sufficient a limited number of preferences (but more than one) to reach a very high level of support. This concludes that each user can be described by a few preferences that highly characterize her.

Finally, it is interesting to observe how the total support of the users and consequently the ranks of the top ten artists and genres change when the preferences in $|\tilde{g_u}|$ and $|\tilde{a_u}|$ are considered instead of those in $|\hat{g_u}|$ and $|\hat{a_u}|$. We report in Table 1 the top ten of the top listened genres and artists and the top ten of the most representative genres and artists with the users support, i.e., the percentage of users having that genre or artist as $\hat{g_u}$ or $\hat{a_u}$, and $\tilde{g_u}$ or $\tilde{a_u}$. We can notice how for the two most listened genres (rock and pop) there is a significant drop in the total support, vice-versa the other genres gain levels of support. The overall rank in the genre top ten is not modified very much. On the other hand, a completely new rank appears for the artists with a clear redistribution of the support out of the top ten. This last result is another proof that the user's preferences are systematic, but they are not towards a unique genre or artist, while they are towards groups of preferences.

	$\{\hat{g_u}\}$	sup	$\{\hat{a_u}\}$	sup	$\{\tilde{g_u}\}$	sup	$\{\tilde{a_u}\}$	**sup**
1	Rock	53.86	The Beatles	0.75	Rock	13.41	David Bowie	0.29
2	Pop	19.64	David Bowie	0.72	Pop	9.73	Arctic Monkeys	0.26
3	Hip Hop	5.05	Kanye West	0.56	Hip Hop	5.16	Radiohead	0.24
4	Electronic	2.21	Arctic Monkeys	0.54	Inide Rock	4.39	Rihanna	0.24
5	Folk	2.03	Rihanna	0.51	Folk	4.31	Coldplay	0.23
6	Punk	1.74	Lady Gaga	0.48	Electronic	4.26	The Beatles	0.22
7	Inide Rock	1.65	Taylor Swift	0.47	Punk	4.07	Kanye West	0.21
8	Dubstep	0.90	Radiohead	0.43	House	2.63	Muse	0.19
9	House	0.85	Muse	0.38	R&B	2.53	Florence	0.19
10	Metal	0.84	Daft Punk	0.37	Emo	2.11	Lady Gaga	0.19

Table 1: Top ten of the top listened ($\{\hat{g_u}\}$, $\{\hat{a_u}\}$) and most representative ($\{\tilde{g_u}\}$, $\{\tilde{a_u}\}$) genres and artists with corresponding support.

4.2 Who are my friends? PLDM, network and homophily.

So far, we focused on describing how individual users can be characterized by their listening patterns; however, sometimes self-awareness by itself is not sufficient to realize *who we are*. In order to understand *where we are* positioned with respect to the mass or with respect to our friends, we need to compare ourselves with them and to calculate the degree of the differences.

Given two users $u, v \in U$ it is possible to calculate the similarity between them by comparing their PLDMs P_u and P_v. By exploiting the previous result, we decided to compute two distinct families of similarities:

- *music-taste* similarity: computed on the most representative music preferences, e.g. $\tilde{g_u}$, instead of complete frequency dictionaries for artist, album and genre;
- *temporal* similarity computed on the day/timeslot frequency dictionaries.

We can analyze the similarity among two users by using the *cosine similarity function* among their frequency dictionaries: for example given $\tilde{g_u}$ and $\tilde{g_v}$ for u and v, we measure their similarity as $cos(g_u, g_v) = \frac{g_u * g_v}{||g_u|| ||g_v||}$.

To understand if, and how, friendship ties affect the listening behavior and users' homophily we calculated the similarities among all the pairs of users in U (we call this set A), and we compared these distributions with the ones obtained by filtering out the nodes that are not directly connected in the social graph. Fig. 6 reports the distribution between pairs of users for artist, album, genre, day, and time-slot. Quite surprisingly, we can observe nearly exactly the same distributions[7] when considering all the pairs in A or just the friends (F). This means that users' ties in Last.Fm social network are not driven by a special listening behavior: the friends in the users' ego-networks are a sample of all the users inscribed to the system. Another interesting result is that genre, day and time-slot distributions are "reverse tilde" $\backsim$-shaped. There is a peak of pairs which are not similar at all (similarity equals to zero), and a growing trend of pairs of users which are more and more similar up to another peak of quite similar use: just a few couples are identical. On the other hand, the distributions for

[7] The Pearson correlations ranges in $[0.96, 0.99]$, p-value $\ll 1.0e^{-60}$.

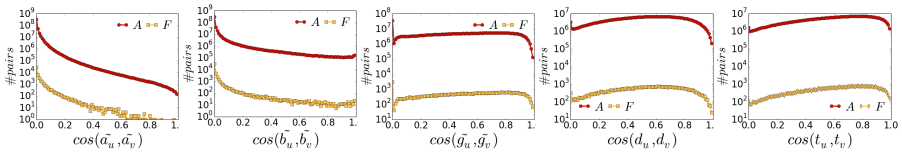

Fig. 6: Distributions between all the pairs of users A and between users which are friends F for artists, albums, genres, days and time-slots (from left to right).

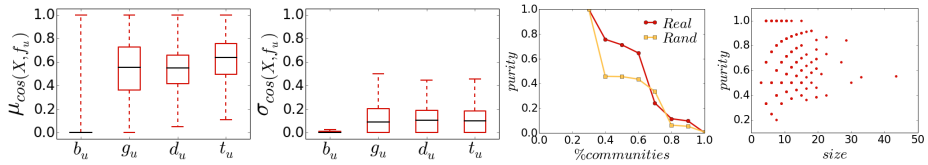

Fig. 7: (left) Boxplots of ego-network indicators μ and σ for album, genre, day and time-slots; (right) Community Discovery results.

artists are long-tailed, while those for album are U-shaped with a peak between most similar and a peak between most different.

Ego-Networks and homophily. According to [2, 9], we decided to characterize each user with respect to her listening behavior and the listening behavior of her friends. We described the ego-network and the homophily of each user (for each analyzed feature) through two additional *indicators* μ and σ. We indicate with μ the inter-quartile mean and with σ the standard deviation of the cosine similarity calculated on the Last.Fm friends f_u of a given user u. The higher is μ, the more homophilous is u with her friends w.r.t. a certain variable X (where X can be the genre, album, etc.). The higher is σ, the more various is the similarity between u and her friends f_u w.r.t. a certain variable X. Fig. 7 depicts the boxplots of μ (left) and σ (right) for album, genre, day and time-slot. We indicate with $cos(X, f_u)$ the cosine of a certain variable X calculated between user u and her friends f_u. Most of the users have a low μ indicator for the album, but many users have quite high μ indicators for the genre, day and time-slot. The variability σ is in line with the previous indicator: the higher the similarity, the higher the variability of the features.

Segmentation Analysis. By exploiting the previous indicators μ and σ we investigate the existence of different groups of listeners with respect to their listening taste compared with those of their friends. We applied the clustering algorithm K-Means [28] by varying the number of clusters $k \in [2, 50]$. By observing the sum of squared error [8] we decided to select 8 as the number of clusters. In Fig. 8 are described the normalized radar charts representing the centroids and the size of the clusters.

Cluster D is the cluster with the lowest indicators. It contains the users who are not very similar to their friends. If we observe the left part of the radars representing clusters B and G, we can notice that they are comparably

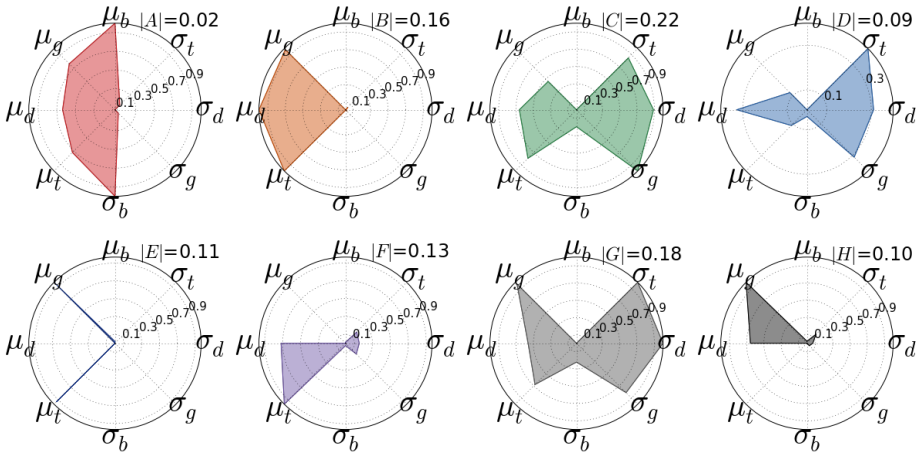

Fig. 8: Radar charts for the centroids of the clusters extracted on the indicators of friendship homophily based on PMDLs.

pronounced in terms of users having friends with similar listening behavior (time-slot, day and genre). However, cluster G has also a great variety with respect to these features. On the contrary, cluster C contains only users having a great variability but not significant similarity in preferences with their friends. Clusters E, F and H have variability very low and are complementary in terms of μs. The first one contains users similar w.r.t. time-slot and genre, the second one users similar w.r.t. day and time-slot, the last one users similar w.r.t genre and day. Finally, cluster A, the smallest one, contains users who are similar w.r.t album besides genre, day and time.

Once identified clusters of similar users, we analyzed if their characterization reflects on the network structure: are community formed by users belonging to the same cluster? To achieve this goal we design a three-step approach:

1. *partitioning* of the social graph $G = (U, E)$ in mesoscale topologies by applying a community discovery;
2. *labeling* of each node within a community with the identifier of the cluster it belongs to;
3. *evaluating* the level of *purity* of each community as the relative support of the most shared node label.

Among various community discovery algorithms, we decided to adopt a state-of-art bottom-up approach: Demon [7]. Demon works on the assumption that, in a social scenario, communities emerge from the choices of individuals: each Last.Fm user directly chooses her friends, and the community she belongs to is implicitly described by this bottom-up wiring pattern. Demon extract micro-communities starting from the ego-network graph of each user and then recombines them in order to identify stable and dense mesoscale structures without suffering the so-called "scale problem" that affects other approaches based on modularity (e.g., Louvain [5]). Moreover, the chosen algorithm has proven to be

one of the best solutions while the final task was to identify network substructures able to bound homophilic behaviors [27].

By applying Demon we obtained 2160 communities. Most of the communities are pure with respect to the clustering labels. Indeed 30% of the communities are perfectly pure and 60% of the communities have a purity higher than the 0.67 (see Fig. 7 *(center-right)*). We compared this result against a random model obtained through 100 random permutations of the clustering labeling in the communities. The line represents the average of these simulations. Even though the shape of the distribution is similar, high levels of purity are not reached for a considerable portion of the communities. Thus, in general, the users in a community tend to belong to the same cluster. This result lets us conclude that, even if each user has its own peculiar profile, it tends to be surrounded by peers that share similar behavior with respect to the listening tastes compared with those of her friends. Communities are not composed by users that necessarily listen to the same artist/album/genre or that use the service during the same day/time-slot. Conversely, they group together users having the same degree of erratic behaviors. What emerges is that service usage drives people to connect. For example, users that like to listen to various genres tend to surround themselves with people with high music preference entropy (maybe to maximize the exposure to novelty). Vice-versa, users that like to listen to few genres tend to surround themselves with friends with music taste narrowed towards specific genres (maybe to deeply explore various artists of those genres) As highlighted by Fig. 7 *(right)* this result is not affected by the size of the community even though it seems that for larger communities the purity tends to 0.5.

5 Conclusion

The endless growth of individual data is requiring efficient models able to store information and tools for automatically transforming this knowledge into a personal benefit. In this paper, we have presented the Personal Listening Data Model (PLDM). The PLDM is designed to deal with musical preferences and can be employed for many applications. By employing the PLDM on a set of 30k Last.Fm users we endorsed the potentiality of this data structure. We have shown how our modeling approaches can be used to increase the self-awareness of Last.Fm users enabling for a succinct description of music tastes as well as service usage habits. We have discussed how the indicators composing the PLDM can be exploited to produce a user segmentation able to discriminate between different groups of listeners. Finally, we studied the correlations among the segments identified and the modular structure of the Last.Fm social graph. From this last analysis clearly emerges that Last.Fm users tend to cluster, in the network sense, with peers having a similar degree of music entropy and/or similar temporal listening behaviors. In the future, we would like to implement a real web service where a user can provide his Last.Fm username and a personal dashboard exploiting all the features contained in the PLDM, as well as her similarity to her friends, is shown. The dashboard would allow self-awareness and self-comparison with

other users, with similar users or with the user's friends. In this way, a user could enlarge his musical experience, try novel tracks, and increase her musical education because knowledge comes from listening.

Acknowledgment

This work is partially supported by the European Community H2020 programme under the funding schemes: INFRAIA-1-2014-2015: Research Infrastructures G.A. 654024 *SoBigData* (`http://www.sobigdata.eu`), G.A. 78835 *Pro-Res* (`http://prores-project.eu/`), and G.A. 825619 *AI4EU* (`https://www.ai4eu.eu/`), and G.A. 780754 *Track & Know* (`https://trackandknowproject.eu/`).

References

1. Abiteboul, S., André, B., Kaplan, D.: Managing your digital life. Communications of the ACM **58**(5), 32–35 (2015)
2. Al Zamal, F., Liu, W., Ruths, D.: Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. ICWSM **270** (2012)
3. Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F.: Analysis of ego network structure in online social networks. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom). pp. 31–40. IEEE (2012)
4. Bischoff, K.: We love rock 'n' roll: analyzing and predicting friendship links in last.fm. In: Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012. pp. 47–56 (2012)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008)
6. Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., He, X.: Music recommendation by unified hypergraph: combining social media information and music content. In: International conference on Multimedia. pp. 391–400. ACM (2010)
7. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Uncovering hierarchical and overlapping communities with a local-first approach. ACM Transactions on Knowledge Discovery from Data (TKDD) **9**(1) (2014)
8. Draper, N.R., Smith, H., Pownell, E.: Applied regression analysis, vol. 3. Wiley New York (1966)
9. Guidotti, R., Berlingerio, M.: Where is my next friend? recommending enjoyable profiles in location based services. In: CN, pp. 65–78. Springer (2016)
10. Guidotti, R., Coscia, M., Pedreschi, D., Pennacchioli, D.: Behavioral entropy and profitability in retail. In: International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2015)
11. Guidotti, R., Monreale, A., Nanni, M., et al.: Clustering individual transactional data for masses of users. In: SIGKDD. pp. 195–204. ACM (2017)
12. Guidotti, R., Rossetti, G., Pappalardo, L., et al.: Market basket prediction using user-centric temporal annotated recurring sequences. In: 2017 International Conference on Data Mining (ICDM). pp. 895–900. IEEE (2017)

13. Guidotti, R., Rossetti, G., Pedreschi, D.: Audio ergo sum: A personal data model for musical preferences. In: Federation of International Conferences on Software Technologies: Applications and Foundations. pp. 51–66. Springer (2016)
14. Guidotti, R., Sassi, A., Berlingerio, M., Pascale, A., Ghaddar, B.: Social or green? a data-driven approach for more enjoyable carpooling. In: 2015 18th International Conference on Intelligent Transportation Systems. pp. 842–847. IEEE (2015)
15. Guidotti, R., Trasarti, R., Nanni, M.: Tosca: Two-steps clustering algorithm for personal locations detection. In: International Conference on Advances in Geographic Information Systems (SIGSPATIAL). ACM (2015)
16. Guidotti, R., Trasarti, R., Nanni, M.: Towards user-centric data management: individual mobility analytics for collective services. In: SIGSPATIAL. ACM (2015)
17. Guidotti, R., Trasarti, R., et al.: There's a path for everyone: A data-driven personal model reproducing mobility agendas. In: DSAA. pp. 303–312. IEEE (2017)
18. Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Towards parameter-free data mining. In: nternational conference on Knowledge discovery and data mining (SIGKDD). pp. 206–215. ACM (2004)
19. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology pp. 415–444 (2001)
20. de Montjoye, Y.A., Shmueli, E., Wang, S.S., Pentland, A.S.: openpds: Protecting the privacy of metadata through safeanswers. PloS one **9**(7), e98790 (2014)
21. Park, M., Weber, I., Naaman, M., Vieweg, S.: Understanding musical diversity via online social media. In: AAAI Conference on Web and Social Media (2015)
22. Pennacchioli, D., Rossetti, G., Pappalardo, L., et al.: The three dimensions of social prominence. In: Social Informatics, pp. 319–332. Springer (2013)
23. Putzke, J., Fischbach, K., Schoder, D., Gloor, P.A.: Cross-cultural gender differences in the adoption and usage of social media platforms - an exploratory study of last.fm. Computer Networks **75**, 519–530 (2014)
24. Rawlings, D., Ciancarelli, V.: Music preference and the five-factor model of the neo personality inventory. Psychology of Music **25**(2), 120–132 (1997)
25. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: the structure and personality correlates of music preferences. Journal of personality and social psychology **84**(6), 1236 (2003)
26. Rossetti, G., Guidotti, R., Miliou, I., Pedreschi, D., Giannotti, F.: A supervised approach for intra-/inter-community interaction prediction in dynamic social networks. Social Network Analysis and Mining **6**(1), 86 (2016)
27. Rossetti, G., Pappalardo, L., Kikas, R., Pedreschi, D., Giannotti, F., Dumas, M.: Community-centric analysis of user engagement in skype social network. In: ASONAM. pp. 547–552. IEEE (2015)
28. Tan, P.N., Steinbach, M., Kumar, V., et al.: Introduction to data mining, vol. 1. Pearson Addison Wesley Boston (2006)
29. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: Myway: Location prediction via mobility profiling. Information Systems (2015)
30. Vescovi, M., Moiso, C., Pasolli, M., Cordin, L., Antonelli, F.: Building an ecosystem of trusted services via user control and transparency on personal data. In: Trust Management IX, pp. 240–250. Springer (2015)
31. Vescovi, M., Perentis, C., Leonardi, C., Lepri, B., Moiso, C.: My data store: toward user awareness and control on personal data. In: International Joint Conference on Pervasive and Ubiquitous Computing. pp. 179–182. ACM (2014)
32. Zheleva, E., Guiver, J., Mendes Rodrigues, E., Milić-Frayling, N.: Statistical models of music-listening sessions in social media. In: Proceedings of the 19th international conference on World wide web. pp. 1019–1028. ACM (2010)