Audio Ergo Sum
A Personal Data Model for Musical Preferences

Riccardo Guidotti[1,2(✉)], Giulio Rossetti[1,2], and Dino Pedreschi[1]

[1] KDDLab, University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy
{guidotti.riccardo,rossetti.giulio,pedreschi.dino}@di.unipi.it
[2] KDDLab, ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy
{guidotti.riccardo,rossetti.giulio}@isti.cnr.it

Abstract. Nobody can state "Rock is my favorite genre" or "David Bowie is my favorite artist". We defined a Personal Listening Data Model able to capture musical preferences through indicators and patterns, and we discovered that we are all characterized by a limited set of musical preferences, but not by a unique predilection. The empowered capacity of mobile devices and their growing adoption in our everyday life is generating an enormous increment in the production of personal data such as calls, positioning, online purchases and even music listening. Musical listening is a type of data that has started receiving more attention from the scientific community as consequence of the increasing availability of rich and punctual online data sources. Starting from the listening of 30k Last.Fm users, we show how the employment of the Personal Listening Data Models can provide higher levels of self-awareness. In addition, the proposed model will enable the development of a wide range of analysis and musical services both at personal and at collective level.

1 Introduction

The unstoppable rise of smartphones joint with their increasing ability of collecting individual information is creating a huge increment in the production of personal data. Personal information like visited locations, web-searches, purchases, phone calls and even music listening are collected and stored without any clear benefit for the user. Consequently, it is being defined the need for a personal model to manage and exploit these large amounts of data.

In the last years in the scientific community is taking place the idea of the *personal data store*. A personal data store is a personal, digital identity management service controlled by an individual where each user can choose at which level she wants to share her own data [3]. In our context, we would like that a personal data store could allow an individual not only the data storage and management, but also the automatic extraction of systematic behaviors and the providing of proactive suggestions on the basis of the user's profile [7].

Since music is a pervasive dimension of our life, and due to the abundance of online data sources like Spotify, iTunes and Last.Fm, we propose a *Personal*

P. Milazzo et al. (Eds.): STAF 2016, LNCS 9946, pp. 51–66, 2016.
DOI: 10.1007/978-3-319-50230-4_5

Listening Data Model (*PLDM*) able to capture the characteristics and the systematic patterns which are present in our musical listening behavior. The PLDM is built on a set of personal listening represented by an abstract data type taken as input. A listening is formed by the song listened, the artist of the song, the album, the genre and the listening time-stamp.

A crucial component of the PLDM are the *indicators* extracted from the listening features. They summarize the listener and explain her level of repetitiveness in the listening. Moreover, in the PLDM we define some listening *patterns* coming from the listening *frequencies*. These patterns are the top listened genre, artist, album etc. and the most representative preferences. In addition, the PLDM contains the frequent listening *sequences*. Those are the typical repetitions followed by the user during a listening session. In short, the proposed data model is an instance of the personal data store specialized for listening data and equipped to provide an improved level of self-awareness.

We employed the PLDM to study Last.Fm users. Last.Fm is an online platform, where people can listen music, share their own musical tastes and discover new artists and genres on the bases of what they, or their friends, like. We retrieved the last 200 listening of about 30k users resident in the UK. We calculated the PLDM for each user given their listening. The obtained PLDMs allowed us to estimate how the Last.Fm audience is segmented in terms of repetitiveness in their listening. There are some well defined classes: listeners systematic with respect to the listening day or time hour, listeners which are predictable with respect to the artists or with respect to the genre, and also "random" listeners. Another finding is that the musical profile of each user is best outlined using a limited set of distinct musical preferences, but not by a unique liking. Furthermore, we explain how the PLDM can enable the development of a broad range of musical analysis and services both at personal and at collective level.

The paper is organized as follows. Section 2 surveys the works related to personal data model and Last.Fm. Section 3 describes our model for analyzing musical listening. In Sect. 4 are presented the analysis of the PLDM applied to Last.Fm users, while Sect. 5 provides an outline of different possible applications. Finally, Sect. 6 summarizes conclusion and future works.

2 Related Work

The need to handle individual data is leading to the development of personal models able to deal with and summarize human behavior. These data models can be generic or specific with respect to the type of data. In [3] 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. openPDS is oriented to the protection of the metadata shared and on the privacy of the data contained in the system. Similarly, in [9] the authors analyzed a new personal data ecosystem centered around the role of *Bank of Individuals Data*, i.e. a provider of personal data management services enabling people to exploit their personal data. In [16] the authors presented *My Data Store*, a tool

allowing people to control and share their personal data. A test with a small set of real users showed improvement over the users' awareness of their personal data and the perceived usefulness of the tool. My Data Store has been integrated in [15] into a framework that permits the development of trusted and transparent services and apps whose behavior can be controlled by the user, allowing the growth of an eco-system of personal data-based services. Finally, the proposal described in [1] is that each user can select which applications have to be run on which data, facilitating in this way diversified services on a personal server. In such a way, the personal server would contain all the user's favorite applications and all the user's data that are currently distributed, fragmented, and isolated.

The majority of the works in the literature [1,3,9] focus their attention on the architecture of the personal data store and on how to treat data sharing and privacy issues. Hence, the main difference between the personal data model proposed and those present in the literature is that our focus is to obtain an added value from the personal data through the application of data mining techniques. Indeed, we aim to apply the methodological framework proposed in [7] for mobility data to analyze personal musical preferences. The authors proposed a framework for personal mobility data able to automatically perform individual data mining and to provide proactive suggestions for supporting decisions. An application of this approach in mobility data can be found in the *MyWay* system [14]. MyWay is a predictive system based on individual mobility profiles which exploits systematic behaviors models to predict human movements.

To the best of our knowledge this work is the first attempt to define a data model able to capture human listening behavior. We believe that the treatment of musical listening is becoming valuable because in the last decade the music world has started receiving more attention from the scientific community. Last.Fm offers a privileged playground to study different phenomena related to the online music consumption. Hence, by following the example of some recent works, we decided to test our personal data model on this dataset. In [11] the authors measured different dimensions of social prominence on a social graph built upon 70k Last.Fm users whose listening were observed for 2 years. By analyzing the *width*, the *depth*, and the *strength* of local diffusion trees, the authors were able to identify patterns related to individual music genres. In [10] the authors formally defined the effect of social influence providing new models and evaluation measures for real-time recommendations with very strong temporal aspects. The authors of [12] analyzed the cross-cultural gender differences in the adoption and usage of Last.Fm: *(i)* men listen to more pieces of music than women, *(ii)* women focus on fewer musical genres and fewer tracks than men. Finally, in [2] 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 to the same concerts.

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 term of listening.

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

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

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 occurred, the *song* listened, the *artist* which sings the song, the *album* the song belongs to, and the *genre* of the artist.

From the set of listening L_u, for each user we can extract the set of songs S_u, artists A_u, albums B_u and genres G_u. Their sizes ($|\cdot|$) are valuable *indicators*.

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

The user behavior can be summarized through frequency dictionaries indicating the support (i.e. relative number of occurrences) of the listening features.

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

$$sup(X, L) = \{(x, y) | y = |Y|/|L| \wedge x \in X \wedge Y \subseteq L s.t. \forall l \in Y, x \in l\} \quad (1)$$

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 $T = \{(2{-}8], (8{-}12], (12{-}15], (15{-}18], (18{-}22], (22{-}2]\}$ contains the time slots of the day.

These dictionaries can be exploited to extract indicators and patterns.

Definition 3 (Entropy). *Given dictionary $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the* entropy *function returns the normalized entropy defined as*

$$entropy(X) = \frac{-\sum_{i=1}^{n} \mathrm{P}(y_i) \log_2 \mathrm{P}(y_i)}{\log_2 n} \quad \in [0, 1] \quad (2)$$

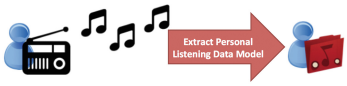

Fig. 2. The raw listening of a user L_u can be turn into a Personal Listening Data Store P_u extracting the songs S_u, artists A_u, albums B_u and genres G_u and by applying to them the functions *sup*, *top*, *repr*, *entropy*, *getseq* and $freqseq$.

The entropy tends to 0 when the user behavior is systematic, tends to 1 when the behavior is not predictable. These indicators are similar to those related with shopping behavior described in [5]. We define the entropy for songs, artists, albums, genres, days and time-slots as $e_{s_u} = entropy(s_u)$, $e_{a_u} = entropy(a_u)$, $e_{b_u} = entropy(b_u)$, $e_{g_u} = entropy(g_u)$, $e_{d_u} = entropy(d_u)$ and $e_{t_u} = entropy(t_u)$.

The simplest pattern we consider is the most listened song, artist, genre, etc.

Definition 4 (Top). *Given dictionary* $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, *the* top *function returns the most supported item. It is defined as:*

$$top(X) = \underset{(x,y)\in X}{argmax}(y) \tag{3}$$

We define the most listened 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 $\hat{g_u} = top(g_u)$, respectively.

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

Definition 5 (Repr). *Given dictionary* $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, *the* repr *function returns the most representative supported items. It is defined as:*

$$repr(X) = \underset{(x,y)\in X}{knee}(y) = \underset{(x,y)\in X^*, y'\in X'}{argmax}(|y - y'|) \tag{4}$$

where X^* *is* X *sorted with respect to the supports* y, $X' = \{y' | y' = mx' + n\}$ *with* $m = (max(sup(X)) - min(sup(X)))/|X|$ *and* $n = min(sup(X))$.

The method $repr(X)$ returns a set of preferences with a support higher than the support of most of the other listening. For example if $g_u = \{(rock, 0.4), (pop, 0.3), (folk, 0.1), (classic, 0.1), (house, 0.1)\}$, $repr(g_u)$ returns $\{(rock, 0.4), (pop, 0.3)\}$.

This result is achieved by employing the *knee* method [13]. Given a dictionary X, the *knee* method sorts the pairs (x_i, y_i) according to the supports generating X^*. Then, it selects the point x_k^* on the support curve X^* which has the maximum distance $|y_k^* - y_k'|$ with the correspondent point x_k' in X', where X' is the straight line passing through the minimum and the maximum point of the curve described by X^*. In this way the *knee* x_k^* is different for each user because it is driven by personal data. Finally, the method returns the pairs with a support greater or equal than the support y_k of the knee x_k. We define the most representative 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 $\tilde{g_u} = repr(g_u)$, respectively. Obviously we have $\hat{g_u} \subseteq \tilde{g_u} \subseteq g_u$ that holds also for songs, albums and artists.

Finally, we want to define the frequent sequences of listening to capture the typical sequences of the listeners. Given the set of listening L_u we can extract for each day a sequence with respect to a certain feature.

Definition 6 (Listening Sequence). *We define a* listening sequence $seq = [i_1, \ldots, i_n]$ *as a list built by concatenating the items of the listening L in a given time window* τ*, ordered by time-stamp and describing a feature of the listening.*

The function $getseq(X, L) = Seq_u = \{seq_1, \ldots, seq_m\}$ orders the listening by time-stamp, divide them in sequences and returns a set of ordered items describing a certain feature, i.e. songs, albums, genres, artists. We name them $Seq_u^S = getseq(S_u, L)$, $Seq_u^A = getseq(A_u, L)$, $Seq_u^B = getseq(B_u, L)$, $Seq_u^G = getseq(\ G_u, L)$ respectively for songs, artists, albums and genres.

In order to extract the frequent pattern sequences we define the function.

Definition 7 (FreqSeq). *The* freqseq *function returns the closed [13] most frequent sequences with at least minsup occurrences. It is defined as*

$$F = freqseq(Seq_u, minsup) \tag{5}$$

where $F = \{(seq_1, sup_1), \ldots, (seq_n, sup_n)\}$ *is a set containing the frequent sub-sequences and their support,* seq_i *is a sub-sequence properly contained or equals to one of the sequences in* Seq_u*,* $sup_i \geq minsup$ *is its support and minsup is the minimum support, for* $1 \leq i \leq n$*.*

By employing the previous functions on L_u, we can obtain for each user the set of frequent sequences $F_{S_u} = freqseq(Seq_u^S, minsup)$, $F_{A_u} = freqseq(Seq_u^A, minsup)$, $F_{B_u} = freqseq(Seq_u^B, minsup)$ and $F_{G_u} = freqseq(Seq_u^G, minsup)$.

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 (see Fig. 2) that we call *Personal Listening Data Model* (PLDM). The PLDM characterizes the listening behavior of a user by means of its *indicators*, *frequencies* and *patterns* (see Fig. 3).

Definition 8 (Personal Listening Data Model). *Given the listening* L_u *of a user u we define the user* personal listening data model *as*

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

Fig. 3. The PLDM is formed by *indicators* ($|L_u|$, $|S_u|$, $|A_u|$, $|B_u|$, $|G_u|$, and entropy values), by *frequencies* (the support dictionaries) and by *patterns* (most listened preference, most representative preferences and frequent sequences).

It is worth to notice that according to the procedures in [6,8], the PLDM can be extracted through a parameter free approach. The only parameter is *minsup*, but we set $minsup = 3$ to capture all the meaningful frequent subsequence: $minsup = 1$ is useless, $minsup = 2$ is too low because there may by a repetition just by chance.

4 LastFM Case Study

In this section we show the benefits derivable from the application of the PLDMs on the data extracted from a famous music website called Last.Fm. In particular, we will show that the information which is generally reported on the main page of many social network or web services (like the most listened song, artist or genre in Last.Fm) are not good enough to represent the user's preferences. Conversely, a structured data model describing the user behavior like the PLDM can achieve this goal, also providing to the user personal access to her data.

Last.Fm is an online social network, where people can share their own music tastes and discover new artists and genres on the bases on what they, or their friends, like. Each user produces data about her own listening. Through each listening a user expresses a preference for a song, artist, album, genre and take place in a certain time. Using Last.Fm APIs[1] we retrieved the last 200 listening of about 30,000 users U resident in the UK. Given the listening L_u, we calculated the PLDM P_u for each user $u \in U$.

4.1 Data Models Analysis

The first analysis we report is related to the *indicators* of the PLDMs $\{P_u\}$ extracted. In Fig. 4 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, i.e. most of the users have listened 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 listens about 60 artists, 70 albums and 10 genres.

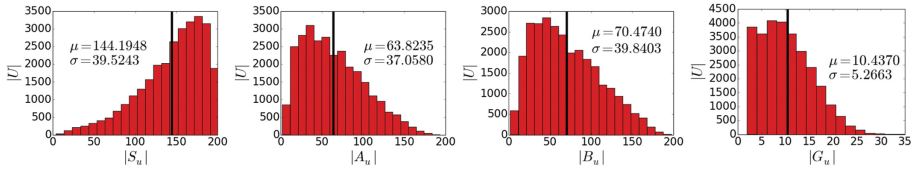

Fig. 4. Distributions of the number of songs $|S_u|$, artists $|A_u|$, albums $|B_u|$ and genres $|G_u|$ respectively. The black vertical lines highlight the means.

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

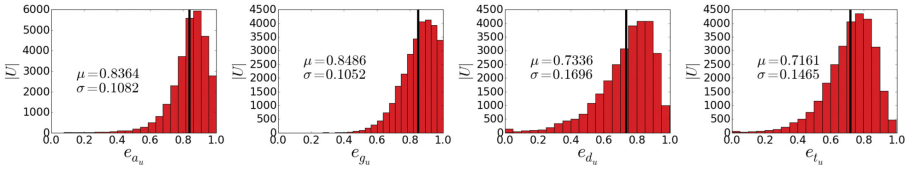

Fig. 5. Distributions of entropy for artists e_{a_u}, genre e_{g_u}, day of week e_{d_u} and time of day e_{t_u} respectively. The black vertical lines highlight the means.

Figure 5 depicts the distributions of the entropy[2]. It emerges that users are much more systematic with respect to the listening time (day of week and time of the day) than with respect to what they listen. This behavior is in opposition to what happens in shopping [5]. Since the artist and genre entropy are right-skewed, it seems that most of the users are not very predictable with respect to the genre or to the artist. This is a first clue that is very unlikely that exists a unique prevalence towards a unique artist or genre.

Figure 6 (left) shows the heat-map of the correlations among the indicators. Some of them like $|A_u|$, $|B_u|$ and $|G_u|$ are highly correlated[3] ($cor(|A_u|, |B_u|) = 0.86$, $cor(|G_u|, |B_u| = 0.64)$): the higher the number of artists or genres, the higher the number of albums listened. Other interesting correlations are $cor(|B_u|, e_{g_u}) = -0.33$ and $cor(|B_u|, e_{a_u}) = 0.55$. Their density scatter plots are reported in Fig. 6 *(center, right)*. They tell us that 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 various artists of this genre and not having a clear preference among these artists.

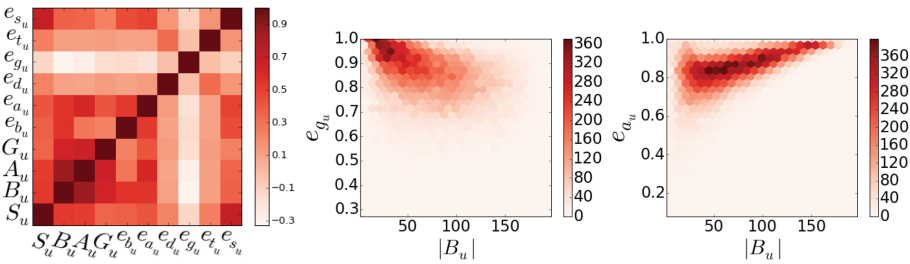

Fig. 6. Correlation matrix (*left*)): the darker the more positively correlated, the lighter the more negatively correlated. Scatter density plots of number of albums $|B_u|$ and genre entropy e_{g_u} (*center*) and number of albums $|B_u|$ and artists entropy e_{a_u} (*right*).

[2] Not all of them are reported due to lack of space.
[3] The *p-value* is zero (or smaller than 0.000001) for all the correlations reported.

4.2 Segmentation Analysis

The second analysis we propose investigate the existence of different groups of listeners with respect to their *indicators* in the PLDMs $\{P_u\}$. We applied the clustering algorithm K-Means [13] by varying the number of clusters $k \in [2, 30]$. By observing the trend of the sum of squared error [4] we decided to select 5 as the number of clusters. In Fig. 7 are described the radar charts representing the centroids while in Table 1 are reported the value of the centroids and the size of the clusters.

Table 1. Centroids for the entropy and size of the clusters extracted.

	e_{t_u}	e_{d_u}	e_{s_u}	e_{a_u}	e_{b_u}	e_{g_u}	size
A	0.8067	0.8442	0.9744	0.8591	0.8794	0.8461	0.44
B	0.7092	0.7234	0.9305	0.7001	0.6732	0.8862	0.13
C	0.4672	0.3366	0.9254	0.7438	0.7717	0.8751	0.06
D	0.5568	0.7687	0.9748	0.8666	0.8855	0.8383	0.19
E	0.7484	0.5624	0.9775	0.8739	0.8918	0.8306	0.19

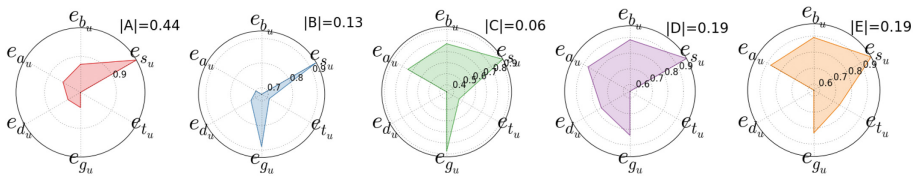

Fig. 7. Radar charts for the centroids of the clusters extracted on the PMDLs.

The most populated cluster is A. It contains the majority of the listeners. It seems that these listeners use the web service without a specific listening schema and that with a high probability they reproduce the tracks using the random function. However, a peculiarity of these users, is that they are more repetitive than users in the other clusters with respect to the genres.

In opposition with A, users in clusters B and C do not have a set of genres which is clearly preferred on top of the others, but they are the most systematic users in terms of albums and artists listened. This means that they like a concise set of artists regardless of their genre and they keep listening only them. The main difference between these two clusters is that users of cluster B are the most systematic in terms of albums and artists, while those of clusters C are the most regular with respect to the use of Last.Fm in specific days and time slots.

Finally, users in clusters D and E are similar to those in cluster A with respect to the level of repetitiveness of listening of genres, artists and albums. On the other hand, how is highlighted by the last two radars in Fig. 7, they are complementary with respect to the day of the week and to time of listening.

Users in cluster D do not have a specific day of the week but use the service constantly at the same time (e.g. during gym session or during specific working areas). Conversely, users in cluster E do not have a specific time slot but use the service periodically in specific days of the week (e.g. during the weekend).

We can conclude that exists a clear distinction among different groups of listeners. From the clustering information originated from the PLDM, a user could learn that is focusing too much on a certain genre or on certain artists and that is not exploring what is outside her "musical confidence zone".

4.3 Sequences Analysis

In this section we make use of the frequent sequences to give a first proof that the most listened genre is not a good candidate to be representative for the user preferences. We remark that a frequent sequence is, for example, a concatenation of genres listened many times in a specific order.

We report in Table 2 the ten most listened genres and artists with the users support, i.e. the percentage of users having that genre or artist as $\hat{g_u}$ or $\hat{a_u}$. To analyze the frequent sequences, for each PLDMs $\{P_u\}$ we considered the most listened genres $\{\hat{g_u}\}$ and the most supported patterns in the genre frequent sequences $\{F_{G_u}\}$ (i.e. the pattern with the highest support). Then, for each genre $g \in G$ we built two sets $F_{G_u}^{\hat{g_u}}$ and $\neg F_{G_u}^{\hat{g_u}}$. $F_{G_u}^{\hat{g_u}}$ contains the most supported patterns of each user having $g = \hat{g_u}$ and containing g into the pattern sequence, while $\neg F_{G_u}^{\hat{g_u}}$ contains the most supported patterns of each user having $g = \hat{g_u}$ and not containing g into the pattern sequence. Figure 8 shows the distribution of the number of genres with respect to the ratio of this two sets $|F_{G_u}^{\hat{g_u}}|/|\neg F_{G_u}^{\hat{g_u}}|$. A ratio smaller than one indicates that the most listened genre is not present in the most supported patterns, vice-versa a ratio greater than one means that the most listened genre is present in the most supported patterns. The higher the ratio the more present is $\hat{g_u}$ in the most supported pattern in F_{G_u}.

Table 2. Ten of most listened genres and artists considering $\{\hat{g_u}\}$ and $\{\hat{a_u}\}$.

	Genre	sup %	Artist	sup %
1	Rock	53.86	The Beatles	0.75
2	Pop	19.64	David Bowie	0.72
3	Hip Hop	5.05	Kanye West	0.56
4	Electronic	2.21	Arctic Monkeys	0.54
5	Folk	2.03	Rihanna	0.51
6	Punk	1.74	Lady Gaga	0.48
7	Indie Rock	1.65	Taylor Swift	0.47
8	Dubstep	0.90	Radiohead	0.43
9	House	0.85	Muse	0.38
10	Metal	0.84	Daft Punk	0.37

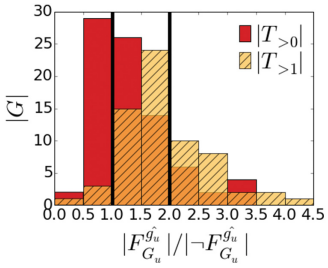

Fig. 8. Distribution of $|G|$ with respect to the ratio of $|F_{G_u}^{\hat{g_u}}|/|\neg F_{G_u}^{\hat{g_u}}|$.

What emerges is that when we consider patterns which have at least two different genres in a sequence (e.g. rock, pop) (labeled with $|T_{>1}|$ in Fig. 8), then for most of the genres the ratio is greater than 1.5. On the other hand, if we consider patterns without any constraint in the number of different genres in a sequence (e.g. rock, rock, rock) (labeled with $|T_{>0}|$ in Fig. 8), than we have that the mode of the distribution is lower than 1.

This result implies that if we consider any kind of sequence, than the most listened genre is among the genres in these patterns but it becomes a significant genre only when patterns with more than a genre are considered. This means that the most listened genre is frequently listened together with other genres.

4.4 Frequency Analysis

In this section we exploit the knowledge of the frequency vectors to demonstrate that the most listened genre, album and artist considered alone do not represent properly the preferences of the users. To this aim we look at the frequency vectors a_u, g_u, the top listened $\hat{a_u}$, $\hat{g_u}$, and the most representative $\tilde{a_u}$, $\tilde{g_u}$. 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. 9 is depicted the result of this analysis for genre (top row) and artist (bottom row)[4]. 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

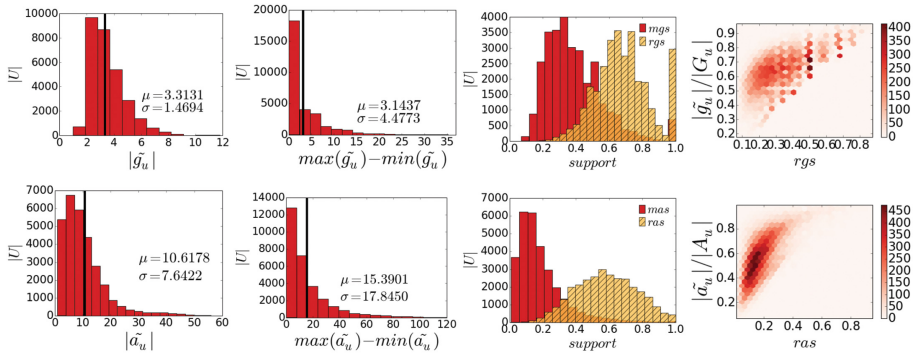

Fig. 9. Frequencies analysis for genre (top row) and artist (bottom row). *First column*: distribution of number of users w.r.t the number of representative preferences. *Second column*: distribution of number of users w.r.t the maximum difference in frequencies between the listening preference. *Third column*: distribution of number of users w.r.t the support given by the representative preferences. *Last column*: density scatter plot between the representative preferences support and the ratio of their number on the number of all the possible artists or genres.

[4] Similar results are obtained for album but they are not reported due to lack of space.

are removed when passing from x to $\tilde{x}$. Indeed, the mean for the genres decreases from 10 to 3, the mean for the artist diminishes from 60 to 10.

The second column in Fig. 9 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, defined as:

$$mas = v \text{ s.t. } (a, v) = \hat{a_u}, \; mgs = v \text{ s.t. } (g, v) = \hat{g_u}$$

and the representative artist support, ras, and representative genre support, rgs, defined as:

$$ras = sum(v|(a, v) \in \tilde{a_u}), \; rgs = sum(v|(g, v) \in \tilde{g_u})$$

From these distributions is evident the increase of the support when not only the top but also all the representative preferences are considered.

The last column reports a density scatter plot of the representative preferences support (rgs and ras) and the ratio of their size on the size of A_u and G_u, i.e. $|\tilde{a_u}|/|A_u|$ and $|\tilde{g_u}|/|G_u|$ respectively. Since the higher concentration of points tends to be $\sim$0.2 with respect to the x-axis and $\sim$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 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 $|\hat{a_u}|$ (see Table 3). 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 complete 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 user's preferences are systematic but they are not towards a unique genre or artist, while they are towards groups of preferences.

4.5 Storage Analysis

To enhance the portability of the PLDM, we report in Fig. 10 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 0.01 Mb together with a computational time of max 5 sec per user, guarantees that the PLDM could be calculated and stored individually without the need of a central service.

Table 3. Top ten of the most listened genres and artists considering $\{\tilde{g_u}\}$ and $\{\tilde{a_u}\}$.

	Genre	sup %	Artist	sup %
1	Rock	13.41	David Bowie	0.29
2	Pop	9.73	Arctic Monkeys	0.26
3	Hip Hop	5.16	Radiohead	0.24
4	Indie Rock	4.39	Rihanna	0.24
5	Folk	4.31	Coldplay	0.23
6	Electronic	4.26	The Beatles	0.22
7	Punk	4.07	Kanye West	0.21
8	House	2.63	Muse	0.19
9	R&B	2.53	Florence	0.19
10	Emo	2.11	Lady Gaga	0.19

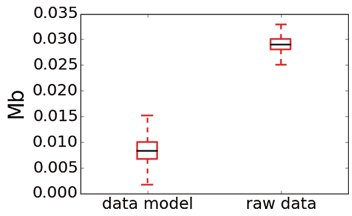

Fig. 10. Data storage for the data model *(left)* and for the raw data *(right)*.

5 Applications

The PLDM described can be easily applied for many purposes and for a wide range of tasks. In this section we will try to structure some application proposals. A first diversification can be made with respect to the main purpose: *analysis* and *services*. Another categorization can be made with respect to the type of data required: *individual* and *collective*. Before going forward it is worth to notice that the computation needed to calculate the PLDM is very small and each user could potentially have it calculated on her own personal device without requiring an external service. Consequently, privacy issues in real applications can be treated by adopting the frameworks in [3,15]: the PLDM only belongs to the user that can decide if she wants or not to disclose it (or part of it) to other users.

The simplest example of *individual analysis* is the user *self-awareness*. Through a smart visualization of the features of the PLDM the user can obtain an unexpected new level of consciousness of her listening behavior. For example a user could discover that is listening 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 starting a new listening with an unknown artist belonging to a different genre to enlarge her musical knowledge, possibly discovering new musical preferences. 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.

Nevertheless, sometimes only the self-awareness is not sufficient to realize *who we are* if we do not compare ourselves with the others (*collective analysis*). Thus, if a portion of users agrees to share some features of the PLDM it becomes possible to understand how much we differ from the mass and *where we are* positioned with respect to the others. For example we could discover that our

most representative genres are the same of the mass but that we are much more systematic than others.

In addition, there are very diversified categories of listeners and comparing ourselves with all the others can be not meaningful. Users segmentation at a collective level can reveal these categories. Then the knowledge of the membership to a category and the comparison with the users belonging to the same category can reveal more interesting results. As shown in Sect. 4.2, user segmentation can be obtained by applying clustering techniques on the indicators, patterns and frequencies. According to this, a third party *collective service* provider could exploit shared PLDM to offer recommendations services for artists, song, genre etc. Furthermore, different types of recommendations could be provided according to the type of user in the diffusion process [11] and considering if a user is good in discovery novel successful songs.

Finally, each user can make use of the PLDM for *individual services*. Some examples are the creation of personal play-lists coming from the prediction of the desire of the user for a certain genre or artist, and the automatic reproduction in certain days and time of the day. According to the personal data store framework these individual services can be integrated and extended with collective knowledge bringing to the user an upgraded level of services.

6 Conclusion

The endless growing 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. It is formed by *indicators* of the musical behavior, listening *patterns* and vectors containing the listening *frequencies*. By employing the PLDM on a set of 30k Last.Fm users we proved the potentialities of this model. We have shown how the indicators of PLDM can be exploited to produce a users segmentation able to discriminate between different groups of listeners. Moreover, the patterns and frequency vectors of the PLDM have been used to prove that information like the most listened genre or artist are not enough to represent the musical preferences of a user. Finally, we have proposed a wide set of applications of the PLDMs at individual and collective level both for analytic purposes and for the development of novel services.

In the future, it would be interesting to consider in the Last.Fm PLDM also the friendship dimension in order to estimate and evaluate the level of homophily of each user with respect to different listening and musical aspects. In addition, we would like to implement a real web service where a user can provide her Last.Fm username and a personal dashboard exploiting all the features contained in the PLDM 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 her musical experience, try novel tracks and increase her musical education because knowledge comes from listening.

Acknowledgements. This work was partially supported by the European Communitys H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement 654024 *"SoBigData: Social Mining & Big Data Ecosystem"*, http://www.sobigdata.eu, and under the founding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement 641191 *"CIMPLEX Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories"*, https://www.cimplex-project.eu.

References

1. Abiteboul, S., André, B., Kaplan, D.: Managing your digital life. Commun. ACM **58**(5), 32–35 (2015)
2. Bischoff, K.: We love rock 'n' roll: analyzing and predicting friendship links in last.fm. In: Web Science 2012, WebSci 2012, Evanston, IL, USA. 22–24 June 2012, pp. 47–56 (2012)
3. 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)
4. Draper, N.R., Smith, H., Pownell, E.: Applied Regression Analysis, vol. 3. Wiley, New York (1966)
5. Guidotti, R., Coscia, M., Pedreschi, D., Pennacchioli, D.: Behavioral entropy and profitability in retail. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. IEEE (2015). 36678 2015
6. Guidotti, R., Trasarti, R., Nanni, M., Tosca: two-steps clustering algorithm for personal locations detection. In: 23rd International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2015). ACM (2015)
7. Guidotti, R., Trasarti, R., Nanni, M.: Towards user-centric data management: individual mobility analytics for collective services. In: MobiGIS Workshop Co-located with ACM SIGSPATIAL 2015. ACM (2015)
8. Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Towards parameter-free data mining. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 206–215. ACM (2004)
9. Moiso, C., Minerva, R.: Towards a user-centric personal data ecosystem the role of the bank of individuals' data. In: 2012 16th International Conference on Intelligence in Next Generation Networks (ICIN), pp. 202–209. IEEE (2012)
10. Pálovics, R., Benczúr, A.A.: Temporal influence over the last.fm social network. Social Netw. Anal. Mining **5**(1), 4:1–4:12 (2015)
11. Pennacchioli, D., Rossetti, G., Pappalardo, L., Pedreschi, D., Giannotti, F., Coscia, M.: The three dimensions of social prominence. In: Jatowt, A., et al. (eds.) SocInfo 2013. LNCS, vol. 8238, pp. 319–332. Springer, Heidelberg (2013)
12. 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. Comput. Netw. **75**, 519–530 (2014)
13. Tan, P.-N., Steinbach, M., Kumar, V., et al.: Introduction to Data Mining, vol. 1. Pearson Addison Wesley, Boston (2006)
14. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: Myway: Location prediction via mobility profiling. Inf. Syst. (2015)

15. Vescovi, M., Moiso, C., Pasolli, M., Cordin, L., Antonelli, F.: Trust management IX. In: Damsgaard Jensen, C., Marsh, S., Dimitrakos, T., Murayama, Y. (eds.) IFIPTM 2015. IAICT, vol. 454. Springer, Heidelberg (2015)
16. Vescovi, M., Perentis, C., Leonardi, C., Lepri, B., Moiso, C.: My data store: toward user awareness and control on personal data. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 179–182. ACM (2014)