RESEARCH

Support the Underground: Characteristics of Beyond-Mainstream Music Listeners

Dominik Kowald1*, Peter Muellner1, Eva Zangerle2, Christine Bauer3, Markus Schedl4,5 and Elisabeth Lex6

Abstract

Music recommender systems have become an integral part of music streaming services such as Spotify and Last.fm to assist users navigating the extensive music collections offered by them. However, while music listeners interested in mainstream music are traditionally served well by music recommender systems, users interested in music beyond the mainstream (i.e., non-popular music) rarely receive relevant recommendations. In this paper, we study the characteristics of beyond-mainstream music and music listeners and analyze to what extent these characteristics impact the quality of music recommendations provided. Therefore, we create a novel dataset consisting of Last.fm listening histories of several thousand beyond-mainstream music listeners, which we enrich with additional metadata describing music tracks and music listeners. Our analysis of this dataset shows four subgroups within the group of beyond-mainstream music listeners that differ not only with respect to their preferred music but also with their demographic characteristics. Furthermore, we evaluate the quality of music recommendations that these subgroups are provided with four different recommendation algorithms where we find significant differences between the groups. Specifically, our results show a positive correlation between a subgroup’s openness towards music listened to by members of other subgroups and recommendation accuracy. We believe that our findings provide valuable insights for developing improved user models and recommendation approaches to better serve beyond-mainstream music listeners.

Keywords: Music Recommender Systems; Acoustic Features; Last.fm; Clustering; User Modeling; Fairness; Popularity Bias; Beyond-Mainstream Users

1 Introduction

In the digital era, users have access to continually increasing amounts of music via music streaming services such as Spotify and Last.fm. Music recommender systems have become an essential means to help users deal with content and choice overload as they assist users in searching, sorting, and filtering these extensive music collections. Simultaneously, both music listeners and artists benefit from the employed segmentation and personalization approaches that are typically leveraged in music recommendation approaches [1]. As a result, users with different preferences and needs can be targeted in various ways with the goal that all users are presented the information and content that they need or prefer. This also means that current recommendation techniques should serve all users equally well, independent of their inclination to popular content.
Present work. In the paper at hand, we focus on music consumers who listen to music beyond the mainstream (i.e., users who listen to non-popular music) in the music streaming platform Last.fm\footnote{https://www.last.fm}. As highlighted in Figure 1, current recommender systems do not work well for consumers of beyond-mainstream music (see Section 3.5 for details on this analysis). In contrast, music consumers who listen to popular music seem to get better recommendations. This finding is not essentially new. In fact, it is a widely-known problem that recommender systems (and those based on collaborative filtering, in particular) are prone to popularity bias, which leads to the behavior that long-tail items (i.e., items with few user interactions) have little chance being recommended. This phenomenon is also present across different application domains such as movies \cite{4} or music \cite{5}.

Our previous work \cite{6} has shown that users interested in beyond-mainstream music tend to have larger user profile sizes (i.e., individual users show a high(er) number of distinct artists they have listened to) compared to users interested in mainstream music. The observation that beyond-mainstream music listeners produce a substantial amount of digital footprints motivates the need to improve the recommendation quality for this group. However, although related research has already studied the long-tail recommendation problem (e.g., \cite{7, 8, 9, 10}; see Section 2 for a more detailed discussion of related work), it is still a fundamental challenge to understand and identify the characteristics of beyond-mainstream music and beyond-mainstream music listeners. Additionally, related work \cite{11} has shown that the group-specific concepts of openess and diversity influence recommendation quality, where openess is defined as across-group diversity (i.e., do users of one group listen to the music of other groups?) and diversity is defined as within-group variability (i.e., how dissimilar is the music listened to by users within groups?). Thus, we are also interested in the correlation between the characteristics of beyond-mainstream music and music listeners with openess and diversity patterns as well as with recommendation quality. Concretely, our work is guided by the following research question:
**RQ:** What are the characteristics of beyond-mainstream music tracks and music listeners, and how do these characteristics correlate with openness and diversity patterns as well as with recommendation quality?

To address this research question, we create, provide, and analyze a novel dataset called *LFM-BeyMS*, which contains complete listening histories of more than 2,000 beyond-mainstream music listeners mined from the Last.fm music streaming platform. Besides, our dataset is enriched with acoustic features and genres of music tracks. Using this enriched dataset, we identify different types of beyond-mainstream music via unsupervised clustering applied to the acoustic features of music tracks. We then characterize the resulting music clusters using music genres. Then, we assign beyond-mainstream users to the clusters to further divide the beyond-mainstream users into subgroups. We study how the characteristics of these beyond-mainstream subgroups correlate with openness and diversity patterns as well as with recommendation quality measured through prediction accuracy.

**Findings and contributions.** We identify four clusters of beyond-mainstream music in our dataset: (i) *C_folk*, music with high acousticness such as “folk”, (ii) *C_hard*, high energy music such as “hardrock”, (iii) *C_ambi*, music with high acousticness and high instrumentalness such as “ambient”, and (iv) *C_elec*, music with high energy and high instrumentalness such as “electronica”. By assigning users to these clusters, we get four distinct subgroups of beyond-mainstream music listeners: (i) *U_folk*, (ii) *U_hard*, (iii) *U_ambi*, and (iv) *U_elec*. We also find that these groups differ considerably with respect to the accuracy of recommendations they receive, where group *U_ambi* gets significantly better recommendations than *U_hard*. When relating our results to openness and diversity patterns of the subgroups, we find that *U_ambi* is the most open but least diverse group, while *U_hard* is the least open but most diverse group. This is in line with related research [11], which has shown that openness is stronger correlated with accurate recommendations than diversity. This means that users are more likely to accept recommendations from different groups (i.e., openness) rather than varied within a group (i.e., diversity).

Summed up, our contributions are:

- We identify more than 2,000 beyond-mainstream music listeners on the Last.fm platform and enrich their listening profiles with acoustic features and genres of music tracks listened to (Sections 3.1–3.4).
- We validate related research by showing that beyond-mainstream music listeners receive a significantly lower recommendation accuracy than mainstream music listeners (Section 3.5).
- We identify four clusters of beyond-mainstream music using unsupervised clustering and characterize them with respect to acoustic features and music genres (Section 4.1).
- We define four subgroups of beyond-mainstream music listeners by assigning users to the music clusters and discuss the relationship between openness, diversity, and recommendation quality of these groups (Section 4.2).
- To foster reproducibility of our research, we make available our novel *LFM-BeyMS* dataset via Zenodo and the entire Python-based implementation of our analyses via Github.
We believe that our findings provide useful insights for creating user models and recommendation algorithms that better serve beyond-mainstream music listeners.

2 Related Work

We identify three strands of research that are relevant to our work: (i) modeling of music preferences, (ii) long-tail recommendations, and (iii) popularity bias in music recommender systems.

**Modeling of music preferences.** A multitude of factors [12] influences musical tastes and musical preferences of users. Characteristics of music listeners and music preferences have been studied in various research domains [13], ranging from music sociology [14] and psychology [15] to music information retrieval and music recommender systems [1]. Studies on music listening behavior showed that personal traits and long-term music preferences are correlated as people tend to prefer music styles that align with their personalities [16, 17]. Furthermore, related work found a relationship between music and motivation [18], music and emotion [19, 20, 21, 22] or both personality and emotion [23]. Openness, a personality trait from the Five Factor Model [24], has also been shown to positively influence a user’s preference for music recommendations [11]. Specifically, the authors of [11] found that people tend to prefer recommendations from different kinds of music (i.e., openness) rather than varied within a specific kind of music (i.e., diversity). Others showed that familiarity has a positive influence on music preferences [25, 26] and that music preferences may change over time [27]. Another strand of research on modeling users’ music preferences leverages content features, e.g., acoustic features. It has been shown that the distribution of acoustic features of a user’s preferred genre substantially influences the user’s choice of music within other genres [28]. Also, acoustic features have been utilized to model users’ preferences under different contextual conditions, in order to refine recommendation quality [29]. Based on tracks’ acoustic features, the authors of [30] identified several types of music, and subsequently modeled each user by linearly combining the acoustic features of the music types. In contrast to these works, we focus on using acoustic features of music tracks for modeling and clustering beyond-mainstream music. Additionally, we link these beyond-mainstream music clusters to music genres and users in our Last.fm data sample.

**Long-tail recommendations.** Related research [8, 9] has found that individual music consumption is biased towards popular music and that usage data for less popular music is scarce. Due to the scarcity problem, items with no or few ratings (i.e., long-tail items) have little chance of being recommended [7]. As a consequence, users that particularly favor items with few ratings or interactions are less likely to be recommended those items that they like [5]. That is problematic because many users, from time to time, prefer niche music [10]. Therefore, such users are not well served as a result of their preference for less popular items. That has been attributed to popularity bias, which corresponds to over-representation of popular items in the recommendation lists [31, 32, 33]. Abdollahpouri et al. [4] studied popularity bias in a dataset of movies (i.e., the MovieLens 1M dataset [34]) from the user
perspective. Their study showed that commonly used recommendation techniques tend to deliver worse recommendations to users who prefer less popular movies. In our work [6], we found evidence for popularity bias in a Last.fm dataset and showed that traditional personalized recommendation algorithms such as collaborative filtering deliver worse recommendations for consumers of niche music. In the present work, we aim to gain a deeper understanding of the behavior and preferences of this beyond-mainstreaminess user group. Thus, in contrast to existing works in long-tail recommendations, we focus on the user rather than the item perspective.

**Popularity bias in music recommender systems.** Music recommender systems [1] are crucial tools in online streaming services such as Last.fm, Pandora, or Spotify. They help users find music that is tailored to their preferences. The basis of music recommender systems are user models derived from users’ listening behavior, user properties such as personality (e.g., [35]), content features of music, or hybrid combinations of both, e.g., [36, 37, 38, 39]. As discussed earlier, due to insufficient amounts of usage data for less popular items, many music recommendation algorithms do not provide useful recommendations for consumers of less popular and niche items. As a remedy, in [40], an approach is suggested that divides music consumers into experts and novices according to their long tail distribution in their playlists. These experts are then converted to nodes with bidirectional links connecting all the experts. These links are created to perform link analysis on the graph and to assign fine-grained weights to songs. The presented approach helps add music from the long-tail into the recommendation list. In our previous research [41], we use a framework [42] that employs insights from human memory theory to design a music recommendation algorithm that provides more accurate recommendations than collaborative filtering-based approaches for three groups of users, i.e., low-mainstream, medium-mainstream and high-mainstream users. While the awareness of popularity bias in music recommender systems increases (e.g., [43]), the characteristics of music consumers whose preferences lie beyond popular, mainstream music are still not well understood. In the present work, we shed light on the characteristics of such beyond-mainstream music consumers and relate them to openness and diversity patterns as well as recommendation quality. With this, we aim to provide useful insights for creating novel music recommendation models that mitigate popularity bias.

### 3 Preliminaries

We investigate the characteristics of beyond-mainstream music listeners in a dataset mined from Last.fm, a popular music streaming platform. We characterize the tracks in our dataset with acoustic features. Besides, we compare the recommendation accuracy of beyond-mainstream music listeners with the one of mainstream music listeners to motivate our subsequent analysis of the characteristics of beyond-mainstream music listeners.

#### 3.1 Acoustic Music Features

For our analyses, we characterize music tracks using acoustic features that describe the content of a given track. Following the lines of, e.g., [44, 45, 46, 30], we rely
on acoustic features provided by the Spotify API as a compact characterization of tracks. The following eight features are extracted from the audio signal of a track:

**Danceability** captures how suitable a track is for dancing and is computed based “on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity”.

**Energy** describes the perceived intensity and activity of a track and is based on the dynamic range, perceived loudness, timbre, onset rate, and general entropy of a track.

**Speechiness** captures the presence of spoken words in a track. High speechiness values indicate a high degree of spoken words (e.g., an audiobook), whereas medium values indicate tracks with both music and speech (e.g., rap music). Low values represent typical music tracks.

**Acousticness** measures the probability that the given track only contains acoustic instruments.

**Instrumentalness** quantifies the probability that a track contains no vocals, i.e., the track is instrumental.

**Tempo** measures the rate of the track’s beat in beats per minute.

**Valence** describes the “emotional positiveness” conveyed by a track (i.e., cheerful and euphoric tracks reach high valence values).

**Liveness** measures the probability that a track was performed live, i.e., whether an audience is present in the recording.

### 3.2 Enriched Dataset of Music Listening Events

To study characteristics of beyond-mainstream users and their listening preferences, we create a novel dataset called **LFM-BeyMS** that contains the required information for such analyses. We base our work on a dataset gathered from the Last.fm music platform, which we considerably enrich with the music tracks’ acoustic features (see Section 3.1) [47]. Additionally, we combine this data with mainstreaminess information of Last.fm users (see Section 3.3) as well as music genre information to identify beyond-mainstream listeners and music (see Section 3.4).

An overview of our new **LFM-BeyMS** dataset and its data sources is depicted in Figure 2. As shown, the starting point for our new dataset is the publicly available **LFM-1b** dataset of music listening information shared by users of the online music platform Last.fm [48]. **LFM-1b** contains listening histories of 120,322 users; their listening records (or “listening events”) have been created between January 2005 and August 2014. They sum up to over 1.1 billion listening events (LEs), where each LE is described by an (anonymous) user identifier, the artist name, the album name, the track name, and the timestamp of the listening event. Also, the **LFM-1b** dataset includes demographics of some users (i.e., country, age, and gender).

To enrich the **LFM-1b** dataset to suit our task, we utilize our previously created **CultMRS** music recommendation dataset [49]. This dataset contains 55,191 users, who have listened to a total of 26,022,625 distinct tracks, captured by a total of 807,890,921 listening events [47].

To further enrich the dataset with music acoustic features, we gather the acoustic features described in Section 3.1 for the tracks remaining in the dataset after the filtering described above. To this end, we rely on the Spotify API to gather...
content-based acoustic features for each track. Particularly, we search tracks using the <track, artist, album> triples extracted from the LFM-1b dataset using the Spotify search API to gather the Spotify track URI of each track by using all three parts of the triple in a conjunctive query. In total, this allowed gathering 4,326,809 Spotify URIs. For the remainder of the tracks, we were not able to retrieve a URI. We attribute this to two factors: either the searched track is not provided by Spotify or the track, artist, and album information cannot be matched to a Spotify track unambiguously. Subsequently, we use the obtained track URI to query the acoustic features API, which returns the acoustic features of a given track (cf. Section 3.1). In a subsequent cleaning step, we remove all tracks for which the Spotify API did not provide the full set of acoustic features.

That procedure provides us with a set of 3,478,399 unique tracks and their acoustic features. Within the LFM-1b dataset, this amounts to 13.36% of the distinct tracks. Overall, these account for as much as 48.89% of all listening events (i.e., the tracks listened to by users) of the LFM-1b dataset. The resulting dataset, now enriched by acoustic music descriptors, comprises a total of approximately 394 million listening events of 55,149 users. In Table 1 (column “CultMRS”), we provide further descriptive statistics of the CultMRS dataset. We refine this dataset to create our new LFM-BeyMS dataset (column “BeyMS in Table 1), which consists of BeyMS, i.e., data to study the characteristics of beyond-mainstream music lis-
Table 1 Descriptive statistics of the CultMRS dataset and our novel LFM-BeyMS dataset. CultMRS comprises acoustic features of tracks. LFM-BeyMS is based on CultMRS and consists of BeyMS and Recommendation. Our analyses of beyond-mainstream music listeners utilize BeyMS and our recommendation experiments utilize Recommendation, which includes listening events of both users with beyond-mainstream and mainstream music taste.

| Item                     | CultMRS [49] | LFM-BeyMS (our novel dataset) |
|--------------------------|--------------|-------------------------------|
| Users                    | 55,149       | 2,074                         |
| Tracks                   | 3,478,399    | 157,444                       |
| Artists                  | 337,840      | 14,922                        |
| Listening Events (LEs)   | 394,944,868  | 4,916,174                     |
| Min. LEs per user        | 1            | 3                             |
| Q1 LEs per user          | 1,442        | 1,254                         |
| Median LEs per user      | 5,667        | 2,048                         |
| Q3 LEs per user          | 9,738        | 3,239                         |
| Max. LEs per user        | 399,210      | 10,536                        |
| Avg. LEs per user        | 7,161.41 (± 10,326.91) | 2,371.526 (± 1,520.629) | 4,022.990 (± 1,898.060) |

teners, and Recommendation, i.e., data to conduct recommendation experiments of beyond-mainstream and mainstream music listeners.

3.3 Identifying Beyond-Mainstream Music Listeners

To identify beyond-mainstream music listeners, for each user, we compute a mainstreaminess score, which is generally defined as the overlap between a user’s individual listening history and the aggregated listening history of all Last.fm users in the dataset. In this vein, the mainstreaminess score reflects a user’s inclination to music listened to by the Last.fm mainstream listeners (i.e., the “average” Last.fm listener in the dataset). In [50], several measures of user mainstreamness are defined. Out of these, we choose the M-global-R-APC definition since it yielded good results in context-based music recommendation experiments for the LFM-1b dataset, as evidenced in [50]. The M-global-R-APC measure approximates a user’s mainstreaminess score by computing Kendall’s $\tau$ [51] rank correlation between the user’s vector of artist play counts and the global vector of artist play counts (aggregated over all users in the dataset). This definition also explains the name of the measure, where “M” refers to mainstreaminess, “global” indicates the global perspective, “R” stands for rank correlation, and “APC” refers to artist play counts.

Next, we describe how we identify our beyond-mainstream users via filtering the users by the number of listening events (see Figure 3 and Section 3.3.1) and by mainstreaminess scores (see Figure 4 and Section 3.3.2).

3.3.1 Filtering Users by the Number of Listening Events

For our study, we select the users so that listeners of different levels of “listening activity” are equally represented. We conduct a Gaussian kernel density estimation (KDE) [52] on the distribution of listening events over users to estimate the continuous probability density function (PDF) [53]. However, KDE estimates the PDF via discrete bins and hence, it is necessary to approximate the gradient via the principle of finite differences. The gradient of the PDF helps us identifying regions of increasing or decreasing probability.

Figure 3 shows that two large subsets of users exist that exhibit either very few or an abundance of listening events. For our analysis, we consider only users who are not in one of the subsets as mentioned earlier. On the one hand, we exclude
users with too little data available for studying their listening behavior; and on the other hand, we exclude so-called power listeners who might bias our analyses. Furthermore, such users with a very high number of listening events are often radio stations, which do not contribute reliable data to our investigations.

Hence, we define lower and upper bounds regarding the number of users’ listening events to include in our study, such that the rate of change in terms of the number of listening events is minimal and stable within these boundaries. That requires the gradient of the region within the lower and upper bound to be near zero (i.e., $\pm 10^{-6}$). By computing the second-order accurate central differences [54], we obtain an approximation of the gradient and find the longest cohesive region fulfilling the requirements between a lower bound of 4,688 and an upper bound of 14,787 listening events per user, which leads to 12,814 users.

3.3.2 Filtering Users by Mainstreaminess Scores
Figure 4 illustrates the mainstreaminess distribution of the 12,814 users that we have extracted based on the number of listening events. Here, mainstreaminess is defined according to the $M$-global-$R$-$APC$ definition taken from [50] (explained in Section 3.3). By setting an appropriate upper bound, we aim to exclude mainstream music listeners. In other words, we aim to set the upper bound to the beginning of the distribution’s bulk, which is motivated as follows: Firstly, the first inflection point (i.e., maximal gradient) of a Gaussian distribution is found at $E[X] - std(X)$, where $E[X]$ is the expectation, and $std(X)$ is the standard deviation of the Gaussian random variable $X$. Secondly, the first inflection point of a Gaussian distribution is equivalent to the 15.9-percentile. By setting the mainstreaminess threshold to
this point, we intend to omit the majority of users and hence, only consider the 15.9% of users with the lowest mainstreaminess scores. Utilizing this upper bound on the mainstreaminess score, we obtain a set of 2,074 beyond-mainstream users. Furthermore, the Gaussian assumption can be strengthened by the observation that the 2,074 beyond-mainstream users represent 16.19% of users. In the remainder of this paper, we refer to this set of beyond-mainstream music listeners as BeyMS.

3.4 Identifying Beyond-Mainstream Music
We aim to study beyond-mainstream listeners in terms of their music taste. We characterize music via its acoustic features, as described in Section 3.1, and also investigate genres as an alternative way to describe a music track via conventional categories. As the LFM-1b dataset does not contain genre annotations of tracks and the Spotify API only provides genres on artist level\(^6\), we leverage the tags assigned to each track by Last.fm users to identify genre annotations. To obtain these tags, we use the respective Last.fm API endpoint\(^7\). After having fetched the tags for each track, we de-capitalitize them and remove all non-alpha-numeric characters. Since not all tags used by Last.fm users correspond to actual music genres (e.g., the “seenlive” tag is used to indicate that a user has seen an artist performing this track live), we use a fine-grained music genre taxonomy consisting of 3,034 genres that are also utilized by Spotify, which we gather from the EveryNoise service (2019-07-24)\(^8\). Specifically, for each track listened to by any of our BeyMS users, we remove all tags that are not part of the EveryNoise genre taxonomy, using a case-insensitive matching approach.
We note that Last.fm users tend to assign very general genre tags to a large number of tracks, such as “pop” or “rock”. To remove these coarse-grained genres and to identify fine-grained beyond-mainstream music genres, we calculate the inverse document frequency (IDF) [55] metric of our genre-track distribution by treating genres as terms and tracks as documents, i.e., $IDF(g) = \log_{10} \frac{|T|}{|\{t \in T \mid g \in G_t\}|}$. More precisely, the IDF-score of genre $g$ is determined by relating the number of all tracks $|T|$ to the number of tracks annotated with genre $g$ where $|G_t|$ is the set of genres assigned to track $t$. This way, a coarse-grained genre receives a small IDF-score, while a fine-grained genre receives a high IDF-score. Figure 5 shows the IDF-score distribution of the top-100 genres in ascending order (i.e., from coarse-grained to fine-grained). Here, we identify two groups of genres, where the first group consists of 6 genres with small IDF-scores, and the second group consists of 94 genres with high IDF-scores. The visual inspection of Figure 5 indicates that the lower bound of 0.90 serves as a discriminant between these two groups of coarse-grained and fine-grained genres. Consequently, we remove the 6 coarse-grained genres (i.e., “rock”, “pop”, “electronic”, “metal”, “alternativerock”, “indierock”) from the genre assignments of our tracks, which leads to 157,444 out of 799,659 tracks listened to by BeyMS users with at least one remaining genre. In total, these tracks are annotated with 1,418 unique genre identifiers.

We are aware of the fact to our track filtering procedure leads to incomplete listening profiles of users. Since we rely on genres to describe beyond-mainstream music, these filtering steps are necessary for our study. To ensure that the BeyMS users’ reduced listening profiles are still representative of their music preferences, we further investigate the consequences of the filtering procedure. Here, we find that a user’s listening history (i.e., the entirety of a user’s listening events) is reduced to 61% on average. However, we also find that there are only 62 of the 2,074 BeyMS users, for whom the listening history is reduced to less than 20%. For these users most affected by the filtering, we compare the acoustic feature distributions of their listened tracks before and after the filtering steps, and find that filtering only marginally affects the acoustic feature distributions (i.e., average change in mean $= 0.0098 \pm 0.0148$). This means that the acoustic feature distribution contained in the user’s profile is highly robust against the filtering. The statistics of BeyMS are summarized in column “BeyMS” in Table 1.

### 3.5 Recommendations for Beyond-Mainstream Music Listeners

In order to compare the recommendation accuracy of recommendations received by the users of our BeyMS group and by mainstream users, we construct a dataset consisting of BeyMS’s listening events and the listening events of an equally-sized group of mainstream users. Therefore, we define the MS user group as 2,074 (i.e., the size of our BeyMS group) randomly-chosen users with a mainstreaminess score that is higher than the upper bound for low mainstreaminess, identified in Figure 4. Furthermore, the MS users are also in between the lower and upper bounds for listening events identified in Figure 3. As shown in Table 1 (column “Recommendation”), the dataset used for the evaluation of recommendations contains data of 4,148 distinct BeyMS and MS users, 1,084,922 distinct tracks, and 16,687,363 listening events.
We use the Python-based open-source recommendation library Surprise\textsuperscript{j} to compute and evaluate recommendations. One advantage of using Surprise is that it provides built-in recommendation algorithms as well as a standardized evaluation pipeline, which enhances the reproducibility of our research. Since Surprise is focused on rating prediction, we formulate our music recommendation scenario also as a rating prediction problem, in which we predict the preference of a target user $u$ for a target track $t$. As done in [56], we model the preference of $t$ for $u$ by scaling the play count (i.e., number of listening events) of $t$ by $u$ to a range of $[1; 1,000]$ using min-max normalization. We perform this normalization on the individual user level to ensure that all users share the same preference value ranges. Thus, with this method, we ensure that each user’s most listened track has a preference value of 1,000, while their least listened track has a preference value of 1. To ensure that this min-max normalization procedure does not disrupt the play count distribution of our users, we compare the original play count distribution with the normalized distribution and find that both distributions are strongly right-skewed. Specifically, we find very similar distributions for large amounts of our play count data.

We utilize a selection of Surprise’s built-in recommendation methods consisting of one baseline approach (i.e., UserItemAvg), two neighborhood-based approaches (i.e., UserKNN and UserKNNAvg), and one matrix factorization-based approach (i.e., NMF). Specifically, UserItemAvg predicts the average play count in the dataset by also accounting for deviations of $u$ and $t$, for example, if a user $u$ tends to have more listening events than the average Last.fm user [57]. UserKNN [3] is a user-based collaborative filtering approach and is calculated using $k = 40$ nearest neighbors and the cosine similarity metric, which are the default settings of Surprise. UserKNNAvg is an extension of UserKNN [3] that also takes the average rating of

\begin{figure}[ht]
\centering
\includegraphics[width=0.7\textwidth]{figure5.png}
\caption{IDF-score distribution of the top-100 genres in ascending order (i.e., from coarse-grained to fine-grained). The 6 coarse-grained genres below the lower bound of 0.90 are removed from the genre assignments, i.e., “rock”, “pop”, “electronic”, “metal”, “alternativerock”, “indierock”.}
\end{figure}
Table 2 Mean absolute error (MAE) results for the two user groups MS and BeyMS of different mainstreaminess and a selection of standard recommendation algorithms. A one-tailed Mann-Whitney-U test ($\alpha = 0.0001$) provides significant evidence, indicated by ***, that all algorithms perform worse on BeyMS than on MS in terms of MAE. Furthermore, NMF (as shown in bold) outperforms the other three approaches UserItemAvg, UserKNN and UserKNNAvg.

| User group | UserItemAvg | UserKNN | UserKNNAvg | NMF      |
|------------|-------------|---------|------------|----------|
| BeyMS      | 63.4608***  | 71.6694*** | 67.5770*** | 57.7703*** |
| MS         | 61.2562     | 68.4894 | 63.3985    | 54.8182  |
| Overall    | 62.2315     | 69.8962 | 65.2469    | 56.2492  |

We favor MAE over the commonly used root mean squared error (RMSE) due to several pitfalls, especially regarding the comparison of groups with different numbers of observations [58]. Here, we perform 5-fold cross-validation leading to 5 different 80/20 train-test splits and average the MAE over the 5 folds. NMF clearly outperforms UserItemAvg as well as the two neighborhood-based methods (i.e., UserKNN and UserKNNAvg) both for the two user groups (see rows “BeyMS” and “MS”) separately and overall without distinguishing between the user groups (see row “Overall”). Additionally, we conduct a one-tailed Mann-Whitney-U test ($\alpha = 0.0001$), where we define the null-hypothesis as the MAE for MS being larger than or equal to the MAE for BeyMS. Results marked with *** indicate that the null-hypothesis was rejected for every fold. Thus, all algorithms (including NMF) provide a significantly larger error for BeyMS than for MS. In other words, recommendation quality is significantly better for users with mainstream taste than for users who prefer beyond-mainstream music across all recommendation approaches.

These initial results underpin the need to study the characteristics of the BeyMS user group that receives worse recommendations. The corresponding experiments are presented in the next section.

### 4 Characteristics of Beyond-Mainstream Music and Listeners

We identify the types of beyond-mainstream music using unsupervised clustering and characterize these types with respect to acoustic features and music genres. Besides, we detect subgroups of beyond-mainstream music listeners by assigning users to these clusters and evaluate the recommendation quality obtained for these subgroups. Finally, we discuss the recommendation quality with respect to openness and diversity. For this, we relate to the definitions given by [11]:
Openness is the across-groups diversity (or categorical diversity) and describes if users of one group also listen to the music of other groups. Diversity is the within-groups diversity (or thematic diversity) and describes the dissimilarity of music listened to by users within groups. Based on the findings of [11], we would expect that subgroups with high openness should receive more accurate recommendations than subgroups with high diversity.

4.1 Clustering and Characterizing Beyond-Mainstream Music
To study the different types of music listened to by the users in our BeyMS group, we conduct a cluster analysis. Specifically, we cluster the 157,444 tracks listened to by BeyMS users, where each track is described by the eight acoustic features danceability, energy, speechiness, acousticness, instrumentalness, tempo, valence, and liveness (see Section 3.1). We scale the value ranges of these features to [0, 1] using min-max normalization. The use of latent representations of musical elements such as tracks was shown to be efficient in the area of music information retrieval [59, 60, 30]. Furthermore, for visually analyzing the obtained music clusters and decreasing computation time, we favor a reduction of dimensionality to two dimensions.

We conduct experiments with a broad body of dimensionality reduction methods, i.e., linear and nonlinear principal component analysis (PCA) [61], locally linear embedding [62], multidimensional scaling [63], Isomap [64], spectral embedding [65], t-distributed stochastic neighbor embedding (t-SNE) [66] and uniform manifold approximation and projection (UMAP) [67]. We visually inspected the 2-dimensional feature spaces created by these methods with regards to the clustering quality, and we obtained the visually most homogeneous results with UMAP. Moreover, UMAP has already been successfully used in the music domain [30] and thus, we use it for the remainder of our experiments. Specifically, we utilize the open-source implementation of UMAP [68], which requires four parameters: (i) the distance metric $M$ in the input space, (ii) the number of latent dimensions $D$, (iii) the minimum distance of points in the latent space $d_{min}$, and (iv) the number of neighbors of a point $N$. Based on experimentation and related literature (e.g., [68]), we set the distance metric $M$ to the Euclidean distance, the number of latent dimensions $D$ to 2, the distance $d_{min}$ to 0.1 and the number of neighbors $N$ to 15.

In a next step, we perform clustering on the dimensionality-reduced acoustic features of tracks. Again, we conduct experiments with various clustering methods, i.e., DBSCAN [69], K-Means [70], Gaussian mixture models [71], affinity propagation [72], spectral clustering [73], hierarchical agglomerative clustering [74], OPTICS [75] and HDBSCAN* [76]. Here, we obtain the best results with respect to cluster cohesion and separation using HDBSCAN*. Furthermore, HDBSCAN* was also already used by related work to cluster music items [77]. We employ the open-source implementation of HDBSCAN* [78] that requires four parameters: (i) the minimum cluster size $s_{min}$ that defines the minimum size of a group of points to consider a cluster, (ii) the minimum number of samples in the neighborhood of a core point $N_{min}$, which quantifies how conservative the clustering is, (iii) $\varepsilon$, which enables the recovery of DBSCAN clusters if the $s_{min}$ value is not reached, and (iv) the scaling of the distance $\alpha$, which is another measure of the clustering’s conservativeness. In detail, $\alpha$ scales the distance between two points, which determines
Figure 6 Music clustering results obtained with HDBSCAN* and UMAP for the 2-dimensional mapping. The outputs are four clusters with the following cluster sizes: 12,148 (blue, hatch: /), 92,798 (green, hatch: +), 7,629 (orange, hatch: o) and 30,379 (pink, hatch: x) tracks. 14,490 of our 157,444 BeyMS tracks have not been assigned to a cluster.

whether these points are merged into a cluster. This scaling is used in the construction of HDBSCAN*’s hierarchy of clusterings. Again, we find the best-suited parameters based on experimentation and related literature (e.g., [76]). Specifically, we require each cluster to comprise a sufficiently large number of tracks to increase the level of significance of our subsequent experiments. We expect the existence of very small music clusters and thus, search for the optimal value of the minimal cluster size $s_{\text{min}}$ in the search space of $\{1,000; 1,025; \ldots; 1,475; 1,500\}$, where we obtain the best results with respect to the within-cluster variance for $s_{\text{min}} = 1,375$. Furthermore, tightly packed clusters without any contribution of noise should be favored. In other words, all points within a cluster should be within the neighborhood of at least one core point. Thus, we set the minimal number of samples in the neighborhood $N_{\text{min}} = s_{\text{min}} = 1,375$. The remaining two parameters are set to their default values, i.e., $\varepsilon = 0$ and $\alpha = 1$.

Figure 6 shows the results of the clustering process using HDBSCAN* and UMAP for the 2-dimensional mapping. This process leads to four music clusters. Here, the green cluster (hatch: +) is the largest one with 92,798 tracks, followed by the pink cluster (hatch: x) with 30,379 tracks and the blue cluster (hatch: /) with 12,148 tracks. The smallest cluster is the orange one (hatch: o) as it contains 7,629 tracks. The remaining 14,490 of our 157,444 BeyMS tracks have not been assigned to a cluster and thus, will not be included in further analyses and interpretations. Next, we describe how we name these clusters based on their music genre distributions.

4.1.1 Genre Distributions

In Figure 7, we illustrate the top-10 genres of the four music clusters. For this, we refer to the genre IDF-scores presented in Section 3.4 and weight each genre
assigned to a track in a cluster with its corresponding IDF-score. For example, if a genre with an IDF-score of 1.4 is assigned to 1,000 tracks in a cluster, it is visualized as an aggregated genre IDF-score of 1,400 in the corresponding plot of Figure 7.

Based on the genre distributions, we label each cluster according to its top genre. With respect to the blue cluster (hatch: /) in Plot (a), we find top genres such as “folk” and “singersongwriter”, which typically reflect music with high acousticness. In the remainder of this paper, therefore, we refer to this cluster as $C_{\text{folk}}$. The top genres of the green cluster (hatch: +) in Plot (b) are typical high energy music genres such as “hardrock”, “punk”, “poprock”, and “hiphop”. Based on this, we name this cluster $C_{\text{hard}}$.

For the orange cluster (hatch: o) in Plot (c), we find genres that reflect music with high acousticness and high instrumentalness such as “ambient”, “experimental”, “newage”, and “postrock”. As “ambient” clearly dominates the genre distribution for this cluster, we name this cluster $C_{\text{ambi}}$. Similarly to $C_{\text{folk}}$, this cluster contains music with high acousticness; yet, while $C_{\text{folk}}$ is characterized by low instrumentalness music, $C_{\text{ambi}}$ is characterized by a high level of instrumentalness. Finally, Plot (d) shows the genre distribution of the pink cluster (hatch: x) with “electronica” as the top genre, which leads to the name $C_{\text{elec}}$ for this cluster.

Thus, both $C_{\text{elec}}$ and $C_{\text{hard}}$, consist of high energy music but in contrast to $C_{\text{hard}}$, $C_{\text{elec}}$ also comprise high instrumentalness values. This also makes sense when look-
ing at other top genres of $C_{elec}$ such as “deathmetal” and “blackmetal” where guttural vocal techniques are often mistakenly classified as another type of instrument [79].

To compare the genre distributions among the four music clusters, we illustrate the relative genre frequency distribution of the clusters in Figure 8. The relative frequency of a genre $g$ depicts the fraction of listening events of tracks within a cluster $c$ that are annotated with $g$. Here, we only show genres with a minimum relative genre frequency of 0.1. We see that there are clearly dominating genres in $C_{folk}$ and $C_{ambi}$, whereas the genre distributions in $C_{hard}$ and $C_{elec}$ are more evenly distributed. When relating this finding to the findings of Figure 7, we clearly see that the results correspond to each other: $C_{hard}$ and $C_{elec}$ contain a more diverse genre spectrum (e.g., “hardrock” and “hiphop” are both part of $C_{hard}$’s top genres) than $C_{folk}$ and $C_{ambi}$ (e.g., in $C_{ambi}$’s top genres, we find “ambient” and “darkambient”).

4.1.2 Acoustic Feature Distributions

To understand the musical content of these four music clusters, we analyze the acoustic feature distributions of the four music clusters using boxplots in Figure 9. This visualization does not show any obvious differences with respect to danceability and tempo among the four clusters. For the acoustic features energy, speechiness, acousticness, valence, and liveness, there are similar values for the cluster pairs $C_{folk}$ and $C_{ambi}$, and $C_{hard}$ and $C_{elec}$. We observe differences between these two cluster pairs with respect to energy and acousticness. While $C_{hard}$ and $C_{elec}$ provide high energy values and small acousticness values, $C_{folk}$ and $C_{ambi}$ feature small energy values and high acousticness values.

In contrast, for instrumentalness, we see similar values for the cluster pairs $C_{folk}$ and $C_{hard}$ as well as for $C_{ambi}$ and $C_{elec}$. We observe very high values for $C_{ambi}$
Figure 9 Distribution of the eight acoustic features for the four music clusters. While the clusters do not show obvious differences with respect to danceability and tempo, we find large differences with respect to energy, acousticness and instrumentalness.

and $C_{elec}$, and very small values for $C_{folk}$ and $C_{hard}$. This difference is also visible in Figure 6 in the form of the gap between $C_{folk}$ and $C_{hard}$ on the left, and $C_{ambi}$ and $C_{elec}$ on the right.

Summing up, in $C_{folk}$, we find music with low energy, high acousticness, and low instrumentalness; $C_{hard}$ contains music with high energy, low acousticness, and low instrumentalness; in $C_{ambi}$, we observe music with low energy, high acousticness, and high instrumentalness; and in $C_{elec}$, we find high energy, low acousticness, and high instrumentalness. Thus, these findings are in line with the genre distributions presented in Figure 7.
Table 3  Descriptive statistics of the four subgroups. Here, $|U|$ is the number of users, $|A|$ is the number of artists, $|T|$ is the number of tracks, $|LE|$ is the number of listening events, $|G|$ is the number of genres, $|LE_u|$ is the average number of listening events per user, $\bar{T}_u$ is the average number of tracks per user and $\bar{\text{Age}}$ is the average age (along with the standard deviation) of users in the group.

| Subgroup | $|U|$ | $|A|$ | $|T|$ | $|LE|$ | $|G|$ | $|LE_u|$ | $\bar{T}_u$ | $\bar{\text{Age}}$ (std.) |
|----------|------|------|------|------|------|------|-------|------------------|
| $U_{\text{folk}}$ | 369  | 9,559 | 72,663 | 702,635 | 811 | 1,904.160 | 549.650 | 24.639 ($\pm$ 10.369) |
| $U_{\text{hard}}$ | 919  | 11,966 | 107,952 | 2,150,246 | 1,274 | 2,339.767 | 557.470 | 23.867 ($\pm$ 8.912) |
| $U_{\text{ambi}}$ | 143  | 6,869  | 39,649  | 224,327 | 918  | 1,568.720 | 473.308 | 29.571 ($\pm$ 14.138) |
| $U_{\text{elec}}$ | 642  | 11,814 | 105,907 | 1,416,354 | 1,005 | 2,073.028 | 549.650 | 24.639 ($\pm$ 7.886) |

4.2 Assigning and Studying Beyond-Mainstream Music Listeners

In the next step, we assign the 2,074 BeyMS users to the four music clusters to categorize them into four distinct beyond-mainstream subgroups for further analyses.

For each user $u$, we count the number of listening events $LE_{u,c}$ that $u$ has contributed to the tracks in each cluster $c$, where $c \in C = \{C_{\text{folk}}, C_{\text{hard}}, C_{\text{ambi}}, C_{\text{elec}}\}$. Then, we assign $u$ to the cluster $c$ for which the number of contributed listening events $LE_{u,c}$ is the highest. However, because we have varying cluster sizes, the probability of $u$ listening to a track $t$ of the two larger clusters $C_{\text{hard}}$ and $C_{\text{elec}}$ is much higher than for the two smaller clusters $C_{\text{folk}}$ and $C_{\text{ambi}}$, although $C_{\text{folk}}$ and $C_{\text{ambi}}$ could be more representative choices for $u$. Thus, similar to the IDF distribution of genres (see Figure 5), we take advantage of the IDF scoring to reduce the influence of the larger clusters and to assign higher weights to the smaller clusters.

Specifically, these cluster IDF-scores are given by $IDF(c) = \log_{10} \frac{|T|}{|\{t \mid c_t \in c\}|}$, i.e., by relating the number of all tracks $|T|$ to the number of tracks in cluster $c$ where $c_t$ is the music cluster assigned to track $t$. That lets us define the user–cluster weight $w_{u,c}$ for user $u$ and cluster $c$ as $w_{u,c} = IDF(c) \cdot LE_{u,c}$.

Consequently, users are assigned to the highest weighted music cluster and thus, a subgroup $U_c$ for cluster $c$ is given by $U_c = \{u \in U : \arg \max_{c \in C}(w_{u,c})\}$. Out of the 2,074 BeyMS users, we can assign 2,073 users to these subgroups. Thus, only 1 user listened to tracks not contained in any cluster in Figure 6. Similar to the naming scheme of music clusters, we label the subgroups according to the name of their assigned music cluster. Hence, we obtain four subgroups $U_{\text{folk}}, U_{\text{hard}}, U_{\text{ambi}},$ and $U_{\text{elec}}$.

Table 3 provides basic descriptive statistics of these four resulting subgroups. Here, $U_{\text{hard}}$ is the largest subgroup with $|U| = 919$ users, followed by $U_{\text{elec}}$ with $|U| = 642$ users, $U_{\text{folk}}$ with $|U| = 369$ users, and $U_{\text{ambi}}$ with $|U| = 143$ users. The differences with respect to the number of users also correspond to the differences regarding the number of artists $|A|$, the number of tracks $|T|$, and the number of listening events $|LE|$ contained in the clusters. In the case of the number of genres $|G|$, this differs slightly because the users in the smaller $U_{\text{ambi}}$ cluster listen to more genres (i.e., 918) than the bigger $U_{\text{folk}}$ cluster (i.e., 811). This indicates that the users in $U_{\text{ambi}}$ listen to a broader set of music than the users in $U_{\text{folk}}$.

Considering the average number of listening events per user (i.e., $|LE_u|$) and the average number of tracks per user (i.e., $\bar{T}_u$), we see that, while there is little difference between $U_{\text{hard}}$ and $U_{\text{elec}}$ with respect to $|LE_u|$ and $\bar{T}_u$ is much higher for $U_{\text{elec}}$ (i.e., 670.402) than for $U_{\text{hard}}$ (i.e., 557.470). This indicates that, although the number of listening events is nearly the same, users of $U_{\text{elec}}$ tend to listen to a wider
Figure 10 Radar plot illustrating the contribution of each music cluster to a subgroup. While the weight distribution of \( U_{hard} \) and \( U_{elec} \) is rather narrow, it is more broad in case of \( U_{folk} \) and \( U_{ambi} \) suggesting that these groups are more open to music outside the own music cluster.

set of tracks than users of \( U_{hard} \). With respect to the average age of the users \( \overline{Age} \), we see that the users of \( U_{folk} \) and \( U_{ambi} \) are the oldest ones, and users of \( U_{hard} \) and \( U_{elec} \) are the youngest ones. However, it is worth noting that the group with the highest average age (i.e., \( U_{ambi} \)) also shows by far the highest standard deviation of age (i.e., 14.138 years).

In Figure 10, we show the contribution of each music cluster to each subgroup in the form of a radar plot. For this, we use the user-cluster weights \( w_{u,c} \) introduced before and calculate the average weight over all users in cluster \( c \). One consequence of the IDF scoring applied to \( w_{u,c} \) is that the weight contributions of a user group to the four clusters does not sum up to 1, which eventually influences the interpretation of the values shown in Figure 10. However, in return, these values account for the varying cluster sizes and can also be interpreted as preference weights for a user group towards a specific music cluster.

We observe that the weight distribution of the two larger subgroups \( U_{hard} \) and \( U_{elec} \) is rather narrow, which indicates that these users do not listen to many tracks of other clusters. Contrary to that, the weights of the two smaller subgroups \( U_{folk} \) and \( U_{ambi} \) are more broadly distributed over the four music clusters. This suggests that users of \( U_{folk} \) and \( U_{ambi} \) are more open to music outside of their own music cluster than users of \( U_{hard} \) and \( U_{elec} \).

4.2.1 Correlation of Music Clusters and Beyond-Mainstream Subgroups
To better understand the correlations and connections between the music clusters and subgroups, we plot the Pearson correlation matrix of the four music clusters as a heatmap in Figure 11. Here, we represent each music cluster \( c \) by a 2,073-dimensional
vector (i.e., one entry for each user) consisting of the user–cluster weights $w_{u,c}$, introduced before. Each element in the matrix is then calculated using the Pearson correlation measure based on these cluster vectors. For example, if there is a positive correlation between two clusters, we assume that a user who enjoys music from the one cluster likely also enjoys music from the other cluster. This can give us also an indication of the openness of a subgroup for music mainly listened to by other subgroups. Specifically, for $C_{folk}$, we see a positive correlation between $C_{folk}$ and $C_{ambi}$, and a negative correlation between $C_{folk}$ and both, $C_{hard}$ as well as $C_{elec}$. Users listening to the music of $C_{hard}$ seem to represent the most closed subgroup as $C_{hard}$ because it solely has negative correlations with all other clusters, especially with $C_{ambi}$ and $C_{elec}$. In contrast, users listening to the music of $C_{ambi}$ seem to represent the most open subgroup as $C_{ambi}$ has positive correlations with two other clusters, i.e., $C_{folk}$ and $C_{elec}$. The fourth cluster, $C_{elec}$, is negatively correlated with $C_{folk}$ and especially with $C_{hard}$, and positively correlated with $C_{ambi}$. These results are also in line with the ones shown in Figure 10, in which we identify the users of $U_{ambi}$ as more open music listeners than the ones of $U_{hard}$.

In order to relate the openness of the subgroups to the diversity of the users within the subgroups, we calculate the average pairwise user similarity using the cosine similarity metric computed on the users’ genre distributions, i.e., number of listening events per genre. Figure 12 shows the resulting boxplots for the four identified subgroups (i.e., $C_{folk}$, $C_{hard}$, $C_{ambi}$, and $C_{elec}$). Figure 12 shows that users in $U_{hard}$ and $U_{elec}$ have a rather small average pairwise user similarity and, thus, exhibit a more diverse listening behavior, whereas users in $U_{folk}$ and $U_{ambi}$ tend to listen to more similar music genres and, thus, have a narrow listening behavior within the group. Summed up, we find pronounced differences with respect to openness and
variety across the subgroups. Although $U_{ambi}$ is the most open subgroup (i.e.,
also listens to music of other subgroups), it is also the least diverse subgroup (i.e.,
the users within the group listen to very similar music). That observation is in
line with what is shown in Figures 7, and Figure 8. Here, we see that $C_{ambi}$, i.e.,
the most tightly connected music cluster to $U_{ambi}$, contains the dominating genre
“ambient” as well as genres that are strongly associated with this dominating genre
(e.g., “darkambient”). For $U_{hard}$, we observe the opposite. While it is the least
open subgroup, it is also the most diverse one (e.g., it contains “hardrock” as well
as “hiphop” listeners).

4.2.2 Recommendations for Beyond-Mainstream User Subgroups
In Section 3.5, we have shown that the recommendation accuracy of four person-
alized recommendation algorithms is significantly worse for BeyMS users than for
MS users. Now, we extend this analysis and evaluate the recommendation accuracy
of these algorithms for the four subgroups (i.e., $U_{folk}$, $U_{hard}$, $U_{ambi}$, and $U_{elec}$).

Table 4 shows our results with respect to the mean absolute error (MAE). Addition-
ally, we analyze these results with respect to statistically significant differences
in Table 5 by performing ANOVA ($\alpha = .01$) and a subsequent Tukey-HSD test
($\alpha = .05$). Here, we report pairwise differences as significant (marked with **), if
both ANOVA and Tukey-HSD were significant across all five folds (see Section 3.5
for details on the experimental setup).

We see that among all algorithms, the significantly worst accuracy results (i.e.,
the highest MAE scores) are achieved for the $U_{hard}$ subgroup. Next, $U_{folk}$, $U_{ambi}$
and $U_{elec}$ reach significantly better (i.e., lower MAE scores) than $U_{hard}$ for all algo-
rithms. However, there is no statistically significant difference between the recom-
Table 4 Mean absolute error (MAE) measurements for the four subgroups and four personalized recommendation algorithms. NMF (in bold) outperforms all other algorithms for all subgroups. Among the subgroups, the best accuracy results (i.e., lowest MAE scores) are reached by $U_{ambi}$, while the worst accuracy results (i.e., highest MAE scores) are reached by $U_{hard}$. To facilitate comparison, we also show the MAE measurements for the BeyMS and MS user groups.

| Subgroup | UserItemAvg | UserKNN | UserKNNAvg | NMF    |
|----------|-------------|---------|-------------|--------|
| $U_{folk}$ | 63.2143     | 70.3049 | 67.4406     | 57.2278 |
| $U_{hard}$ | 65.1464     | 73.1949 | 69.2855     | 59.6887 |
| $U_{ambi}$ | 60.5558     | 69.8315 | 65.5708     | 54.2073 |
| $U_{elec}$ | 62.2894     | 71.0387 | 66.1499     | 56.6209 |
| BeyMS    | 63.4608     | 71.6694 | 67.5856     | 57.7703 |
| MS       | 61.2562     | 68.4894 | 63.3985     | 54.8192 |

Table 5 Statistically significant differences between pairs of subgroups, as determined by ANOVA ($\alpha = .01$) and a subsequent Tukey-HSD test ($\alpha = .05$).

| Subgroup | UserItemAvg | UserKNN | UserKNNAvg | NMF  |
|----------|-------------|---------|-------------|------|
| $U_{folk}$ | **         | **      | **          | **   |
| $U_{hard}$ | **         | **      | **          | **   |
| $U_{ambi}$ | **         | **      | **          | **   |
| $U_{elec}$ | **         | **      | **          | **   |

recommendation accuracy of $U_{folk}$ and $U_{elec}$. The overall best accuracy results (i.e., lowest MAE scores) are reached for the $U_{ambi}$ subgroup. These results are also statistically significant when compared with the other subgroups for the NMF algorithm. NMF also gives the overall best accuracy results for all subgroups, which is in line with our results presented in Section 3.5 and in our previous work [6].

Furthermore, we find a relationship between openness, diversity, and recommendation quality. Here, $U_{hard}$ is the least open but most diverse subgroup and gets the worst recommendations, while $U_{ambi}$ is the most open but least diverse subgroup and gets the best recommendations. This is in line with the findings of [11], who have shown that users are more likely to accept recommendations from different groups (i.e., openness) rather than varied within a group (i.e., diversity). Thus, we find a relationship between the quality of recommendations provided to beyond-mainstream music listeners and openness as well as diversity patterns of these users.

Finally, in Figure 13, we visually compare the MAE scores reached by the best performing approach NMF for the four subgroups. Additionally, we depict the MAE score for BeyMS as a black dashed line and the MAE score for MS as a grey dashed line. We see that $U_{hard}$ reaches worse results than BeyMS while $U_{folk}$ and $U_{elec}$ reach slightly better results than BeyMS. Interestingly, $U_{ambi}$ not only reaches better results than BeyMS but also better results than MS. Although this improvement over MS is not statistically significant (according to a one-tailed Mann-Whitney-U test with $\alpha = .0001$), it shows that there is a large variety among BeyMS users, where specific subgroups (i.e., $U_{hard}$) are disadvantaged in terms of recommendation accuracy by recommendation algorithms while others (i.e., $U_{ambi}$) are not.

5 Conclusions and Future Work
In this paper, we shed light on the characteristics of beyond-mainstream music and music listeners. As our first contribution, we identified 2,074 beyond-mainstream music listeners (i.e., BeyMS) in the Last.fm platform, and subsequently created a novel dataset called LFM-BeyMS based on the listening histories of these users. We
Figure 13 Comparison of the mean absolute error (MAE) scores reached by NMF for the four subgroups with the ones reached by NMF for BeyMS (black dashed line) and MS (grey dashed line). While specific subgroups (i.e., $U_{hard}$) are treated in an unfair way by recommendation algorithms, others (i.e., $U_{ambi}$) are not.

Further enriched this dataset with (i) acoustic features of music tracks gathered from Spotify, and (ii) genre information of tracks derived from Last.fm tags and matched with the Spotify microgenre taxonomy. Additionally, for reasons of comparability, LFM-BeyMS contains data of 2,074 Last.fm users listening to mainstream music. Using this dataset, as our second contribution, we validated related research by showing that beyond-mainstream music listeners receive a significantly lower recommendation accuracy than mainstream music listeners by four standard recommendation algorithms (i.e., UserItemAvg, UserKNN, UserKNNAvg and NMF).

As our third contribution, we applied the clustering algorithm HDBSCAN* on the acoustic features of tracks listened by BeyMS and identified four clusters of beyond-mainstream music: (i) $C_{folk}$, music with high acousticness such as “folk”, (ii) $C_{hard}$, high energy music such as “hardrock”, (iii) $C_{ambi}$, music with high acousticness and instrumentalness such as “ambient”, and (iv) $C_{elec}$, music with high energy and instrumentalness such as “electronica”.

As our fourth contribution, we mapped these clusters to our BeyMS users, which led to four beyond-mainstream subgroups: (i) $U_{folk}$, (ii) $U_{hard}$, (iii) $U_{ambi}$, and (iv) $U_{elec}$. We analyzed these subgroups with respect to their openness (i.e., across-groups diversity – do users of one group listen to music of other groups?) and diversity (i.e., within-groups diversity – how dissimilar is the music listened to by users within groups?). Here, we found large differences between $U_{hard}$ and $U_{ambi}$. Although $U_{hard}$ is the most closed subgroup (i.e., users do not listen to music of other subgroups), it is also the most diverse subgroup (i.e., users listen to a diverse set of genres such as “hardrock” and “hiphop”). For $U_{ambi}$, we get opposite results: while it is the most open subgroup (i.e., users listen to music of other subgroups as well), it is also the least diverse one (i.e., the users within the group listen to very similar music such as “ambient” and “darkambient”). We related these characteristics of the subgroups to the recommendation quality of the four recommendation algorithms UserItemAvg, UserKNN, UserKNNAvg and NMF. Here, we found that $U_{hard}$ got music recommendations with lowest accuracy, while $U_{ambi}$ got music recommendations with highest accuracy. This is in line with related research [11], which has shown that openness is stronger correlated with accurate recommendations than diversity. $U_{ambi}$ even received better recommendations than the group of
mainstream music listeners. This result highlights that there are large differences between the subgroups of beyond-music listeners. Finally, to foster reproducibility of our research, we provide our novel LFM-BeyMS dataset via Zenodo as well as our source code via Github.

We believe that our findings provide useful insights for creating user models and recommendation algorithms that better serve beyond-mainstream music listeners. As it was shown in [6], beyond-mainstream music listeners tend to have larger user profile sizes than users interested in mainstream music, which means that they provide a substantial amount of listening interaction data for services such as Last.fm and Spotify. We assume that improving the recommendation quality for this active user group also leads to another effect, namely a more prominent exposure of (long-tail) music artists due to a better-connected recommendation network [80]. We leave such investigations to future work.

Limitations and future work. Despite the merits of this work, we are aware of its limitations. The first limitation we recognize is that our analyses are based on a sample of the Last.fm community. The extent to which their listening behavior is representative of the Last.fm community at large, or similar music streaming communities such as Spotify, needs further investigation.

Next, since we conducted a comparative study of the accuracy of recommender systems algorithms—and were therefore not interested to beat state-of-the-art algorithms—we focused on traditional algorithms (e.g., KNN-based collaborative filtering) instead of investigating the most current deep learning architectures, which would also require a much higher computational effort. Furthermore, an award-winning-paper by Dacrema et al. [81] has recently shown that traditional algorithms are able to outperform almost all deep learning architectures.

While our work serves as a first milestone towards better characterizing beyond-mainstream music and listeners of such music, future work should focus on user modeling techniques to individually target the different subgroups, for example by integrating knowledge about openness and diversity. With respect to analyzing openness and diversity of users and user groups, we would also like to work on a more formal definition of these dimensions, which would not only allow us to measure them more precisely but also to integrate them into the recommendation calculation process.

Additionally, since previous research has shown that the listener’s cultural background impacts the quality of music recommendations [47], we plan to compare the cultural and socioeconomic aspects of beyond-mainstream and mainstream music listeners. We plan to employ these aspects by means of Hofstede’s cultural dimensions [82] and the World Happiness Report [83].

Finally, another avenue for future work is the research in the area of fair music recommender systems. Here, we plan to build user models that are capable of accounting for the complex characteristics of beyond-mainstream music listeners presented in this paper. While we believe that more specialized user models could help to provide better recommendations for users who currently receive worse recommendations (e.g., the $U_{hard}$ subgroup identified in this paper), we also aim to highlight that such user models still need to be generalizable to avoid any unfair treatment
of other users. Hence, future research should work on achieving a specialization-
generalization trade-off in music recommender systems. We hope that our open
LFM-BeyMS dataset as well as our source code will be of use to the scientific
community for subsequent analyses.

Availability of Data and Materials
The CultMRS dataset can be found on Zenodo https://doi.org/10.5281/zenodo.3477842. Additionally, we
provide our novel LFM-BeyMS dataset via Zenodo: https://doi.org/10.5281/zenodo.3784764. Our
Python-based implementations are available via Github https://github.com/pmuellner/supporttheunderground.

Competing interests
The authors declare that they have no competing interests.

Funding
This work is funded by the TU Graz Open Access Publishing Fund and the Austrian Science Fund (FWF): V579.

Author’s contributions
All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgements
This work is supported by the Know-Center GmbH within the Austria FFG COMET program.

Endnotes
a. https://www.last.fm/
b. https://doi.org/10.5281/zenodo.3784764
c. https://github.com/pmuellner/supporttheunderground
d. https://developer.spotify.com/web-api/get-several-audio-features/
e. http://www.cp.jku.at/datasets/LFM-1b/
f. https://developer.spotify.com/web-api/search-item/
g. https://developer.spotify.com/documentation/web-api/reference-beta/#endpoint-get-an-artist
h. https://www.last.fm/api/show/track.getTopTags
i. http://everynoise.com/
j. http://surpriselib.com/

Author details
1 Know-Center GmbH, Graz, Austria. 2 University of Innsbruck, Innsbruck, Austria. 3 Utrecht University, Utrecht,
The Netherlands. 4 Johannes Kepler University Linz, Linz, Austria. 5 Linz Institute of Technology AI Lab, Linz, Austria. 6 Graz University of Technology, Graz, Austria.

References
1. Schedl, M., Knees, P., McFee, B., Bogdanov, D., Kaminskas, M.: Music recommender systems. In: Recommender Systems Handbook, pp. 453–492 (2015)
2. Luo, X., Zhou, M., Xia, Y., Zhu, Q.: An efficient non-negative matrix-factorization-based approach to
  collaborative filtering for recommender systems. IEEE Transactions on Industrial Informatics 10(2), 1273–1284
  (2014)
3. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender
  systems. ACM Transactions on Information Systems (TOIS) 22(1), 5–53 (2004)
4. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The unfairness of popularity bias in
  recommendation. In: RMSE Workshop at ACM Recsys (2019)
5. Celma Herrada, O., et al.: Music recommendation and discovery in the long tail. PhD thesis, Universitat
  Pompeu Fabra (2009)
6. Kowald, D., Schedl, M., Lex, E.: The unfairness of popularity bias in music recommendation: A reproducibility
  study. In: European Conference on Information Retrieval, pp. 35–42 (2020). Springer
7. Celma, O., Cano, P.: From hits to niches?: or how popular artists can bias music recommendation and
  discovery. In: Proceedings of KDD’2018 (Netflix Price Workshop) (2008)
8. Celma, O.: Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital
  Music Space, (2010)
9. Oord, A.v.d., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Proceedings of
  NIPS’2013, pp. 2643–2651. Curran Associates Inc., USA (2013)
10. Goel, S., Broder, A., Gabrilovich, E., Pang, B.: Anatomy of the long tail: ordinary people with extraordinary
tastes. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 201–210 (2010)
33. Jannach, D., Lerche, L., Kamehkhosh, I., Jugovac, M.: What recommenders recommend: an analysis of
32. Brynjolfsson, E., Hu, Y.J., Smith, M.D.: From niches to riches: Anatomy of the long tail. Sloan Management
31. Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S.: All the cool
30. Zangerle, E., Pichl, M.: Content-based user models: Modeling the many faces of musical preference. In:
28. Barone, M.D., Bansal, J., Woolhouse, M.H.: Acoustic features influence musical choices across multiple genres. 
27. Moore, J.L., Chen, S., Turnbull, D., Joachims, T.: Taste over time: The temporal dynamics of user preferences.
26. Pereira, C.S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S.L., Brattico, E.: Music and emotions in the 
25. Schubert, E.: The influence of emotion, locus of emotion and familiarity upon preference in music. Psychology of Music 35(3), 499–515 (2007)
24. Goldberg, L.R.: The structure of phenotypic personality traits. American psychologist 48(1), 26 (1993)
23. Juslin, P.N., Sloboda, J.A.: Music and Emotion: Theory and Research., (2001)
22. Yang, Y.-H., Chen, H.H.: Music Emotion Recognition, (2011)
21. Juslin, P.N., Laukka, P.: Expression, perception, and induction of musical emotions: A review and a 
20. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: characterization, 
19. Juslin, P.N., Sloboda, J.A.: Music and Emotion: Theory and Research., (2001)
18. Kim, Y., Aiello, L.M., Quercia, D.: Pepmusic: motivational qualities of songs for daily activities. EPJ Data 
17. Rentfrow, P.J., Gosling, S.D.: The content and validity of music-genre stereotypes among college students. 
16. Laplante, A.: Improving music recommender systems: What can we learn from research on music tastes? In: 
15. Deutsch, D.: Psychology of Music, (2013)
14. Adorno, T.W.: Introduction to the Sociology of Music, (1988)
13. Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., Elahi, M.: Current challenges and visions in music 
12. Schubert, E.: The influence of emotion, locus of emotion and familiarity upon preference in music. Psychology of Music 35(3), 499–515 (2007)
11. Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: A study 
10. Zangerle, E., Pichl, M.: Content-based user models: Modeling the many faces of musical preference. In: 19th 
9. Haas, R., Brandes, V.: Music that Works: Contributions of Biology, Neurophysiology, Psychology, Sociology, 
8. Barone, M.D., Bansal, J., Woolhouse, M.H.: Acoustic features influence musical choices across multiple genres. 
7. Moore, J.L., Chen, S., Turnbull, D., Joachims, T.: Taste over time: The temporal dynamics of user preferences. 
6. Rentfrow, P.J., Gosling, S.D.: The content and validity of music-genre stereotypes among college students. 
5. Juslin, P.N., Sloboda, J.A.: Music and Emotion: Theory and Research., (2001)
4. Goldberg, L.R.: The structure of phenotypic personality traits. American psychologist 48(1), 26 (1993)
3. Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., Elahi, M.: Current challenges and visions in music 
2. Brynjolfsson, E., Hu, Y.J., Smith, M.D.: From niches to riches: Anatomy of the long tail. Sloan Management 
1. Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: A study 

Page 27 of 28
47. Zangerle, E., Pichl, M., Schiedl, M.: User models for culture-aware music recommendation: Fusing acoustic and cultural cues. Transactions of the International Society for Music Information Retrieval (TISMIR) 3(1), 1–16 (2020). doi:10.5334/tismir.37
48. Schiedl, M.: The fm-1b dataset for music retrieval and recommendation. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 103–110 (2016). ACM
49. Zangerle, E.: Culture-Aware Music Recommendation Dataset. https://doi.org/10.5281/zenodo.3477842
50. Bauer, C., Schiedl, M.: Global and country-specific mainstreamness measures: Definitions, analysis, and usage for improving personalized music recommendation systems. PLOS ONE 14(6), 1–36 (2019). doi:10.1371/journal.pone.0217389
51. Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1/2), 81–93 (1938)
52. Sheather, S.J.: Density estimation. Statistical science, 588–597 (2004)
53. Davis, R.A., Li, K.-S., Politis, D.N.: Remarks on some nonparametric estimates of a density function. In: Selected Works of Murray Rosenblatt, pp. 95–100. Springer, New York (2011)
54. Quarteroni, A., Sacco, R., Saleri, F.: Numerical Mathematics. Springer, New York (2007)
55. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation (1972)
56. Schiedl, M., Bauer, C.: Distance-and rank-based music mainstreamness measurement. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 364–367 (2017). ACM
57. Koren, Y.: Factor in the neighbors: Scalable and accurate collaborative filtering. ACM Transactions on Knowledge Discovery from Data (TKDD) 4(1), 1 (2010)
58. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research 30(1), 79–82 (2005)
59. Moore, J.J., Chen, S., Joachims, T., Turnbull, D.: Learning to embed songs and tags for playlist prediction. In: ISMIR, vol. 12, pp. 349–354 (2012)
60. Levy, M., Sandler, M.: Learning latent semantic models for music from social tags. Journal of New Music Research 37(2), 137–150 (2008)
61. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural computation 11(2), 443–482 (1999)
62. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. science 290(5500), 2323–2326 (2000)
63. Kruskal, J.B.: Nonmetric multidimensional scaling: a numerical method. Psychometrika 29(2), 115–129 (1964)
64. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. science 290(5500), 2319–2323 (2000)
65. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849-856 (2002)
66. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(Nov), 2579–2605 (2008)
67. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software 3(29), 861 (2018)
68. McInnes, L., Healy, J., Saul, N., Grosanberger, L.: Umap: Uniform manifold approximation and projection. The Journal of Open Source Software 3(29), 861 (2018)
69. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231 (1996)
70. Bishop, C.M.: Pattern Recognition and Machine Learning, pp. 424–429. Springer, New York (2006)
71. Reynolds, D.: Gaussian mixture models. Encyclopedia of biometrics, 827–832 (2015)
72. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. science 315(5814), 972–976 (2007)
73. Shi, J., Malik, J.: Normalized cuts and image segmentation. Departmental Papers (CIS), 107 (2000)
74. Murtaza, F., Legendre, P.: Ward’s hierarchical agglomerative clustering method: which algorithms implement ward’s criterion? Journal of classification 31(3), 274–295 (2014)
75. Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J.: Optics: ordering points to identify the clustering structure. In: ACM Sigmod Record, vol. 28, pp. 49–60 (1999). ACM
76. McInnes, L., Healy, J.: Accelerated hierarchical density based clustering. In: Data Mining Workshops (ICDMW), 2017 IEEE International Conference On, pp. 33–42 (2017). IEEE
77. Yoo, S., Lee, K.: A data-driven approach to identifying music listener groups based on users’ playrate distributions of listening events. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 77–81 (2017)
78. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2(11) (2017)
79. York, W.: Voices from hell—the dark, not-so-dulcet cookie monster vocals of extreme metal. The San Francisco Bay Guardian, 14–20 (2004)
80. Lamprecht, D., Strohmaier, M., Helic, D.: A method for evaluating discoverability and navigability of recommendation algorithms. Computational social networks 4(1), 9 (2017)
81. Dacrema, M.F., Cremonesi, P., Jannach, D.: Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019, pp. 101–109 (2019)
82. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind vol. 3rd, revised. McGraw-Hill, New York, NY, USA (2010)
83. Helliswell, J.F., Layard, R., Sachs, J.: World Happiness Report 2016 Update. Sustainable Development Solutions Network, New York (2016)