USER INSIGHTS ON DIVERSITY IN MUSIC RECOMMENDATION LISTS

Kyle Robinson[1] **Dan Brown**[1] **Markus Schedl**[2]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Canada
[2] Institute of Computational Perception, Johannes Kepler University Linz, Austria
kyle.robinson@uwaterloo.ca, dan.brown@uwaterloo.ca, markus.schedl@jku.at

ABSTRACT

While many researchers have proposed various ways of quantifying recommendation list diversity, these approaches have had little input from users on their own perceptions and preferences in seeking diversity. Through an exploratory user study, we provide a better understanding of how users view the concept of diversity in music recommendations, and how they might optimise levels of intra-list diversity themselves. In our study, 17 participants interacted with and rated the suggestions from two different recommendation systems. One provided static top-7 collaborative filtering recommendations, and the other provided an interactive slider to re-rank these recommendations based on a continuous diversity scale. We also asked participants a series of free-form questions on music discovery and diversity in semi-structured interviews. User-preferred levels of diversity varied widely both within and between subjects. Although most users agreed that diversity is beneficial in music discovery, they also noted a risk of dissatisfaction from too much diversity. A key finding is that preference for diversification was often linked to user mood. Participants also expressed a clear distinction between diversity within existing preferences, and outside of existing preferences. These ideas of inner and outer diversity are not well defined within the bounds of current diversity metrics, and we discuss their implications.

1. INTRODUCTION

As music consumption has moved from physical media to digital collections to streaming, people have changed the way they discover new music. As with other forms of consumption which have made the shift to digital media and marketplaces such as movies, television, and consumer products, data on music listening habits is more prevalent than ever. Accordingly, systems which use this data to market or recommend new content to users have become ubiquitous. These *music recommender systems* aim to provide satisfying music recommendations to users across a wide variety of contexts [23].

One common way of recommending music is to create a ranked list where the items are formed by the top-*n* recommendations, as produced by the used recommendation algorithm, sorted by recommendation relevance. To judge the quality of recommendations, various forms of accuracy metrics have been proposed. Typically borrowed from the field of *information retrieval*, these accuracy measurements aim to quantify how well a recommendation (or set of recommendations) aligns with a user's known preferences, or in some cases how satisfied a user will be with those recommendations [10].

In addition to accuracy, various other metrics have been proposed [10, 12]. These aptly named *beyond-accuracy* metrics include novelty, coverage, serendipity, and diversity [10]. Novelty relates to items which are unknown to the user, coverage relates to the proportion of items that can be recommended (item coverage) or to the proportion of users for which at least one recommendation can be made (user coverage), serendipity relates to the unexpectedness of a recommendation, and diversity relates to the dissimilarity of recommended items [10]. We focus our attention to diversity as it is well researched, and easily understood for music [10, 12, 25].

Diversity in *music recommender systems* is well researched, but we are unaware of any research which specifically explores user provided perceptions of diversity. Additionally, most implementations of diversity treat the metric as a static variable between or within users. We argue that desire for diversity in recommendations may instead be situationally dependent so we present users with an interactive system which allows them to select diversity as a continuous trade off against accuracy across numerous personalized recommendation lists. Here, alongside a live user study using this prototype system, we present the results of semi-structured interview questions in order to address the following research questions:

- RQ1: How do users feel about diversity in personalized music recommendation lists?
- RQ2: How might users optimise their own level of diversity in personalized recommendation lists?

We found that users presented a range of definitions for diversity, linked ideal diversity levels to their mood, and distinguished between what we call inner and outer diversity. When asked to optimise their own level of diversity using our system selections differed greatly within and between subjects.

Attribution: Kyle Robinson, Dan Brown, Markus Schedl. "User Insights on Diversity in Music Recommendation Lists", 21st International Society for Music Information Retrieval Conference, Montréal, Canada, 2020.

2. BACKGROUND & RELATED WORK

2.1 Diversity in Recommender Systems

Alongside novelty, coverage, and serendipity, diversity has long been identified as an important metric in providing satisfying automated recommendations to users across varying domains [12]. Diversity in this context traces back to information retrieval tasks, where it was used to resolve ambiguity in search queries [3]. Within recommender systems, diversity prevents over-personalization of recommendations to users, thereby increasing user satisfaction with recommendations [12, 13]. Research on diversity in recommender systems is extensive, and numerous different definitions have been proposed [13]. More generally, recommender system diversity has been described as the opposite of *similarity* [1, 10]. Among the most commonly researched and implemented definitions of diversity in *music recommender systems* is intra-list diversity (ILD) which measures the average pairwise dissimilarity of items using some chosen similarity metric; typically calculated using content features [1, 32].

2.2 Optimising for Diversity

Research on selecting optimal levels of diversity for recommender systems is extensive. In their original paper defining diversity as the opposite of similarity, Bradley and Smyth show that traditional recommender system outputs are not diverse, and diversity, in one metric, can be increased with minimal negative impact on accuracy [1]. Ziegler et al. further showed that user satisfaction with recommendation lists relies on more than accuracy by computing precision, recall, and satisfaction curves in a large user study [32]. Studies following this theme of incorporating existing diversity metrics with minimal negative impact on accuracy and/or satisfaction are plentiful [21, 31]. Whereas these works applied a global level of diversity to recommendations, recent work has focused on selecting levels of diversity on a per-user basis through user modeling [5, 8, 17, 18]. Interactive systems which allow users to explore recommendations through diversity have been explored outside of the music domain, but these systems aim to abstract diversity into a user interface rather than allow for user selection of existing diversity metrics [22, 27, 30].

Differences in user perceived diversity levels have been identified across varying recommendation algorithms [6], and varying levels of intra-list diversification [29]. Finally, user listening habits on diversity have been extracted from social networks [7] and playlists [20].

We are not aware of any research which explores user provided perceptions of diversity in personalized music recommendations, or allows them to directly modify existing diversity metrics on the fly. We begin to fill in this gap by providing knowledge on how well formalizations of diversity align with user perceptions of diversity.

3. METHODOLOGY

To control all aspects of recommendation and diversity inclusion, and to minimise restricting participants' consumption method, we implemented a collaborative filter recommender. We used Last.fm as a source of raw listening data, and presented song previews in the form of standardised 30 second track previews from Spotify.

3.1 Interactive Recommendation Lists

3.1.1 Data

We collected a total of 341,764,569 unique listening events (LEs) from 51,669 unique users whose region was set to North America using the Last.fm API. Users were found by crawling the Last.fm social graph using the *user.getFriends* endpoint. We had a limit of 10,000 LEs accepted per user, and only accepted LEs between January 12, 2019 and when we collected them in February 2020. The median number of LEs per user is 7744, 25th percentile is 3502, and 75th$^{\text{th}}$ percentile is 9842.

We used a simple key consisting of artist and track name tuples in order to identify individual tracks. The final user-track-interaction matrix, used to generate recommendations (see Section 3.1.2), contains 141,205,668 non-zero entries (play counts) across 12,300,857 unique artist-track tuples, resulting in a 51,669x12,300,857-sparse matrix. This system does not account for potentially inaccurate metadata obtained from Last.fm, but does account for the same track across different releases. Entries in this matrix are integers which correspond to the number of unique times a user (row) played the track (column). An anonymized version of this data is available upon request.

3.1.2 Collaborative Filtering & Diversity

For generating recommendations we used an Alternating Least Squares (ALS) matrix factorization algorithm which is designed specifically for implicit feedback data sets [8, 11]. This algorithm results in one vector for each user consisting of a non-negative real number (recommendation relevance) for each track in the database; higher numbers are considered more relevant recommendations. The ALS collaborative filter recommender was implemented using the *Implicit* python library [9], and was trained using the dataset described in Section 3.1.1. Hyperparameters were optimised using 5-fold cross-validation and Mean Average Precision for top-10 recommendations (MAP@10) over 60 iterations of randomized search resulting in 160 factors, 28 iterations, a scaling factor of $\alpha = 774$, and regularization term of $\lambda = 1$.

The trained collaborative filter recommender was used to generate top-400 track recommendation lists for a single Last.fm username (see Section 3.2). To facilitate multiple recommendation lists per-user we split this list evenly into four smaller lists of 100 tracks each. Each track within each of the four lists was assigned a rank from 1-100 with one being the most relevant. In order to measure diversity we used the latent vectors generated for each track during matrix factorization as descriptors. Similar to previous work [8, 29], we calculated a form of ILD (d_i) by summing the Euclidean distance of one track's descriptors (v_i) from all other descriptors (v_j) in each top-100 list.

$$d_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} ||v_i - v_j|| \qquad (1)$$

This calculation differs from previous work in that diversity is only calculated once and not as part of a greedy diversification algorithm. Higher values of d_i correspond to more diverse tracks in relation to others in the list. Tracks are assigned additional ranks from 1-100 where rank one is the most diverse. We are left with four unique top-100 recommendation lists for a given user where each track is assigned a rank for relevance (R_i), and diversity(D_i).

The final ranking (F_i) is calculated as a trade off between relevance and diversity controlled by a convex combination of both ranks, with a diversity parameter β.

$$F_i = (1 - \beta) * R_i + \beta * D_i \tag{2}$$

The user interface, shown in Figure 1, displays the top-7 tracks of each top-100 recommendation list based on F_i in the form of 30 second previews using Spotify Play Button widgets. [1] We chose to use top-7 recommendation lists to ensure user study session times under 70 minutes (see Section 3.2). An interactive slider that controls the value of β is situated above the song previews. The left of this slider corresponds with $\beta = 0$ and the right side corresponds with $\beta = 1$ with a step size of 0.001. A *Well Known* button appears to the left of each song preview allowing users to remove songs which are not new to them.

Due to differences in the music collection available on Spotify and our own music database, as well as to avoid false-positives in retrieving song previews, we omitted all songs which did not match exact artist and song string queries to Spotify. This typically resulted in final recommendation lists of 95-100 tracks each.

3.2 User Study

Participants were recruited on the University of Waterloo campus through internal email lists and posters. After completing a digital information consent form participants were asked to complete a brief survey. As part of this survey they were asked to provide their Last.fm usernames, or alternatively were provided instructions on how to set up a Last.fm account and record their listening events to it. We required that participants had a minimum of 5 hours of LEs recorded before continuing to the interactive portion.

The interactive portion of the study involved a pre-interaction interview, two conditions of 4 trials using four unique recommendation lists, and a post-interaction interview. Interviews were semi-structured. Pre-interaction interview questions focused on the importance of music discovery to the participant, how the participant finds new music, and what a diverse list of personalized recommendations means to them. Post-interaction interview questions focused on the perceived effect of the slider on recommendations, the static or variable nature of their selections across trials, and positives and negatives of diversity in music recommendations.

Trials 1-4 consisted of static top-7/100 recommendation lists each corresponding with one evenly split quarter of their top-400 recommendations as ranked only by recommender output (relevance) (see Section 3.1.2). The user-interface was similar to Figure 1 but without the slider. Participants were asked to listen to each preview, remove well known tracks, mark if they were familiar with the artist, and rate the recommendation on a four-point Likert scale of ’Strongly Dislike’, ’Dislike’, ’Like’, or ’Strongly Like’. Only once every track was rated could the participant move to the next trial.

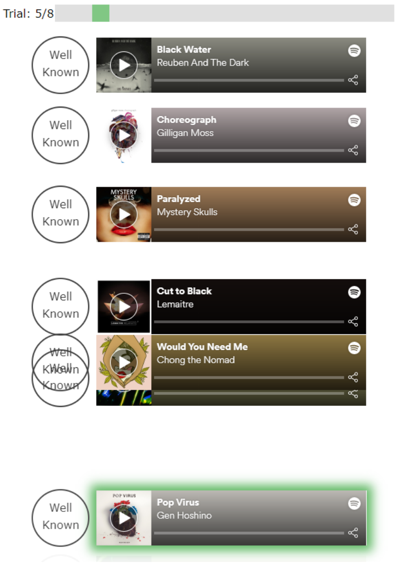

Figure 1. The mid-motion user interface state directly after moving the diversity slider seen at the top. The top 7/100 songs as ranked by Equation (2) are displayed as 30 second song previews. As the slider is moved the songs shift from the old order to the new order over a period of 2 seconds. Songs which leave the top 7 move off the bottom and new songs appear from the bottom highlighted in green for 5 seconds. The circular *Well Known* buttons on the left remove songs from the list entirely.

Trials 5-8 consisted of the same ranked lists as trials 1-4 (minus tracks marked as well-known) with the addition of the interactive slider to re-rank the larger hidden list based on the participants’ selected level of diversity (see Section 3.1.2). The user-interface can be seen in Figure 1. Participants were not told what the slider did and were instructed to find the position on the slider that resulted in the most satisfying recommendation list as a whole while removing tracks that were well known to them. Once the participant locked in this position they were again asked to mark if they were familiar with each song’s artist, and rate each individual recommendation on the same four-point Likert scale before moving to the next trial.

Between each trial participants completed a survey with questions on their satisfaction with the final recommendation list, the level of diversity in the recommendation list, and how well the recommendation list portrayed the definition of diversity they provided in their pre-interview survey. Participants were paid $10 CAD upon completing the interactive portion of the study.

Pre- and post-interaction interviews were transcribed, and comments were then sorted into three categories: in-

[1] https://developer.spotify.com/documentation/widgets/generate/play-button/

teraction, music discovery, and diversity. Similar to other qualitative music consumption studies we extracted individual ideas as statements from transcriptions and proceeded to build connections and groupings through affinity diagramming [2, 19]. Main ideas were highlighted and categorized into groupings of similar themes, and finally counts of each theme were collected. We specifically focused on responses regarding diversity.

4. RESULTS

We recruited 18 participants, and removed one participant for marking all recommendations as *Well Known*, leaving 17 total participants. The median participant age was 23; the oldest was 29 and the youngest 19. Each user session took 50-70 minutes inclusive of interviews. Some sessions were completed face to face, and others involved the users connecting remotely to the interactive system.

4.1 Music Discovery

When asked how they discovered new music, 9 said they used Spotify, 9 used YouTube, 5 used movies and/or television, 4 relied on friends, 3 used radio, and 4 used some other online service such as Amazon or Soundcloud. The importance and frequency of finding new music varied significantly from user to user, and no clear patterns were observed. Some users noted that the primary reason they use music services such as Spotify is to enable easier music discovery. When asked how important finding new music is to them, one user reported previously spending 5 hours per week looking for new music, but added:

> “While it’s still very important to me, I basically don’t do it very often on my own anymore; I rely on Spotify to do almost all of it for me.”

4.2 Recommendations

None of the participants had an existing Last.fm account, and the length of time during which users recorded their listening histories to Last.fm varied from one to three weeks. The median percentage of user LEs which existed in our CF database was 95%, with a max of 100% and a min of 65%. Median LE counts per-user used for recommendation generation were 256, with a max of 1156 and min of 86. All users marked and removed fewer than 100 tracks as well known across all trials, with the exception of one user who marked and removed 208.

When asked to rate individual recommendations on a 4-point Likert scale (Strongly Dislike, Dislike, Like, Strongly Like) 72.69% of songs were rated as ’Like’ or ’Strongly Like’ after locking in the diversity slider, and 74.79% in static lists. In addition to rating individual songs, participants were asked if they were satisfied with the list of recommended music for every trial. On a 5-point Likert scale (Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree) 75% of diversified recommendation lists resulted in a positive response, with 50% for static recommendation lists.

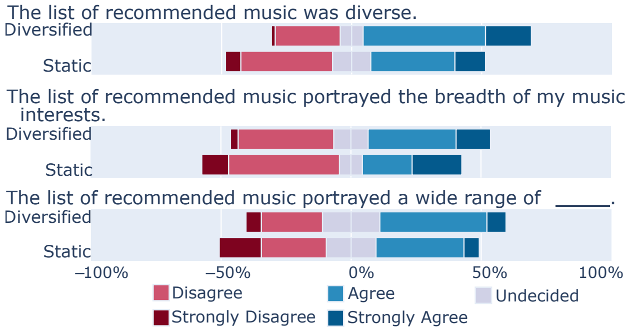

Figure 2. Responses to Likert questions completed after every recommendation list, split between static lists and lists which were selected using the diversity slider. The final question was customized for each individual using their own definition of diversity obtained during the pre-interaction interview (see Section 4.4).

4.3 Interactive Diversity

In addition to the task of selecting an optimal position for the diversity slider, participants were asked a series of questions on how diverse they felt each recommendation list was. Responses to these questions can be seen in Figure 2. In order to visualise how participant responses on diversity align with their diversity selections, Figure 3 shows all 17 user’s diversity selections coded with their Likert response on diversity. User selections varied greatly between their own recommendation lists and between other users’. Likert responses for perceived diversity did not fall in line with levels of β.

As a part of the post-interaction interview participants were asked to identify what they thought the slider was changing within their recommendation lists. Of the 17 participants, 5 identified it to increase diversity directly, 3 identified some change in genres, and 4 had no explanation. The remaining participants identified the slider to change the perceived gender of vocalists, increase ’newness’, increase distaste, increase quality, and decrease quality. In one case where a participant identified the slider to effect genre they stated:

> “I noticed initially that the first side of the slider was giving me a bunch of songs from different genres. The more I was sliding it the more it was giving me the songs. . . from the genre which I like.”

In another case where a participant was unable to identify the effect of the slider and was asked what they would like the slider to do they answered:

> “The way I imagined it was. . . less diverse on the one side and more and the other side. That’s something I could definitely use.”

When asked about their experience using the system some users expressed difficulty in remembering which locations of the slider they preferred most, and frustration over which songs remained on the list and which were moved off. In total, 10 users preferred interacting with the static list, and 7 preferred using the interactive slider.

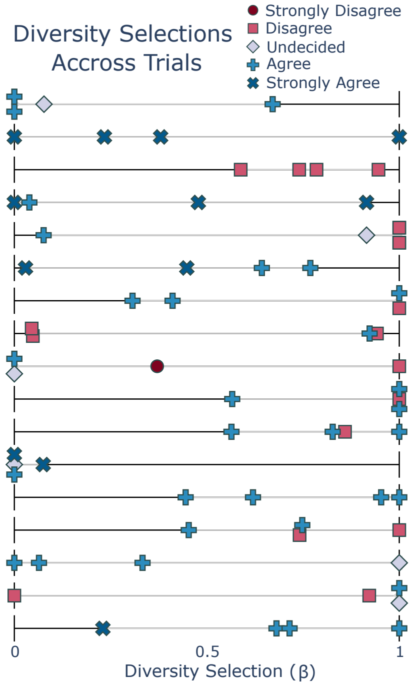

Figure 3. All user selections for diversity using the slider found in Figure 1. The legend corresponds to Likert responses to: "The list of recommended music was diverse."

4.4 User Perceptions of Diversity

During the pre-interaction interview participants were asked what they would mean if they were looking for diverse recommendations. In addition to their open ended responses, they were challenged to come up with a single word or idea that could be used in place of diversity. Of the responses to this question, 13 answered a difference in **genres**, 2 answered **cultural differences**, and the remaining participants responded with **originality**, **variety**, and differences in **artists**.[2] These definitions were used to complete the third question in Figure 2.

Coding of participants' comments on diversity in their own personalized music recommendations resulted in two primary themes which we labeled **diversity meaning**, and **listener mood**. Comments which we classified under **diversity meaning** are deeply intertwined with personal definitions of diversity, and can be more specifically categorized into what we identify as **inner and outer diversity**; that is music within the bounds of existing preferences, and music outside of these bounds. In answering the interview question on the meaning of diverse recommendations, 8 participants made reference to a preference for this idea of inner or outer diversity. Participant comments expressing a preference for inner diversity include:

> "Diverse in the–within the boundaries of the things that I like."

> "I like a playlist which recommends me songs on the genre I like. . . the important thing is to get diversified music in my genre only. . . to stay in the same genre but diversity in artists."

> "A diverse music recommendation I think should still be within the category of music that I usually listen to, but it should be different artists or different albums that I haven't listened to so far."

Comments expressing a preference for outer diversity include:

> "[Diverse recommendations are] something new, something exciting. Something that I'm not used to, like I've never heard before."

> "[Diverse recommendations] would be music from other genres that maybe I haven't listened to very much, but still somewhat akin to the ones that I have listened to."

Secondary in frequency of occurrence to diversity meaning, **mood** was explicitly mentioned by 7 participants. Only 2 participants mentioned context. Participants referenced mood as a primary factor in how much diversity they want in their music recommendations at any given time. Notable comments on mood include:

> ". . . depending on my mood–whether I'm looking for more of the same things that I already like–I could set that slider to show me less diverse music–if I'm in the mood."

> "[I] like a piece of music right now because of the mood that I am in, but I might not like it while I'm listening to a very different kind of music. So diversity is good but I think in a weird way the recommender system should know when to recommend it."

> "Sometimes you're in the mood of listening to one specific–like you don't want [a] diverse playlist. You just want to listen to sad songs. You just want a playlist that has a sad song. You don't want diversity"

> "If you're in a melancholic mood and then you don't have a very diverse playlist of melancholic music then you'd be happy about your music because that's your mood."

Participants also provided their thoughts on the positives and negatives of diversity in personalized recommendations, and a summary of these thoughts can be found in Table 1. Participants generally felt that while diversity could enable music discovery, it also increased the risk of disliking some recommendations.

5. DISCUSSION

In this study we provided a primary analysis of user perceptions on diversity in personalized music recommendations. We also provided users an opportunity to directly optimise a diversity metric which until now had been algorithmically optimised for them. Although our results do not hold statistical power due to the small sample size, our semi-structured interviews facilitated valuable insights and answers to our posed research questions. These insights add to the growing number of other qualitative works in Music Information Retrieval research [2, 14–16, 19].

[2] One participant was unable to choose between genre and culture.

	More Diversity	Less Diversity
Pos.	Music Discovery (N=11) Preference Discovery (N=4) Interesting (N=4)	Likely to Like (N=2)
Neg.	Likely to Dislike (N=8) Dissatisfaction/Annoyance (N=2)	Restrictive (N=4) Repetitive (N=2) High Risk/Reward (N=2) Unremarkable (N=2)

Table 1. Positives and negatives of more and less diversity in recommendation lists expressed by participants.

5.1 RQ1: How do users feel about diversity in personalized music recommendation lists?

Despite a large variance in user's feelings towards diversity in music recommendations, their ideas on its positives and negatives (Table 1) mostly align with the metric's purpose of reducing over-personalization. Beyond this, however, users attached more complex ideas such as personal preference discovery and interestingness to more diversity. Ideas such as this may in part explain the higher levels of satisfaction reported by users given more diverse recommendations.

The prevalence of mood in participants descriptions of diversity is especially notable when compared to the lack of references to their context. As more focus is directed towards context-aware recommender systems [24], careful attention should be paid to not assume that ideal diversity levels can be determined by context alone. Diversity optimisation may also serve as an ideal jumping off point for mood-based recommendation [4,26]. In designing systems which incorporate diversity, it is also important to note that preferred diversity levels may not remain static on an individual user basis.

Although most participants described diversity as a difference in genres, genre was not the exclusive answer. To some participants, a recommendation list which spans genre may not be considered diverse unless those genres span a range of cultures, and to other users a recommendation list which spans artists in just one genre may be considered diverse.

The occurrence of inner and outer diversity–that is diversity within the bounds of existing preference, and outside of those bounds–was an unexpectedly binary result, and neither of these ideas are well defined by existing beyond-accuracy metrics. Inner diversity is not well described as novelty, nor is outer diversity well described by serendipity. The idea of inner diversity does however align with idea of user genre coverage [28]. More research on the universality of inner and outer diversity preference is clearly required.

In their foundational paper on diversity in *information retrieval*, Clarke et al. use a query for 'jaguar' as an example to show the usefulness of diversity; a diverse response might include the cars, the cats, and the classic Fender guitar [3]. In the case of music recommendations, all diverse responses may be simultaneously correct to one user, and incorrect to another.

5.2 RQ2: How might users optimise their own level of diversity in personalized recommendation lists?

The interactive system we implemented (Figure 1) represents a first attempt in allowing users to optimise diversity metrics in line with how they are optimised in existing studies. As such, all variables other than the level of diversity (Equation (2)) were fixed. We note that in allowing users to remove well-known songs the system represents a specific use for diversity in discovering novel music.

Diversity selections accross the interactive trials varied widely within and between users. Ideally in Figure 1, users' Likert ratings would be distributed with positive responses on the right ($0.5 \leq \beta \leq 1$), and negative responses on the left ($0 \leq \beta \leq 0.5$). While results do not follow this distribution, the responses in Figure 2 show that users generally found the slider system to enable more diversity. We hypothesise a combination of three reasons for these results. First, the Likert survey provided no frame of reference for diversity and participants used their own idiosyncratic definitions. Second, the users' responses were heavily impacted by music previewed before locking in a diversity value. Third, the diversity metric did not match users' models of diversity. All three of these hypotheses should be considered for future implementations.

We also note that while our selection of CF recommender and diversity metric have a basis in previous work, there are countless combinations of them which may be used to comprise of a system such as ours. Also, more recent music recommendation algorithms based on deep neural networks could be investigated [24].

6. CONCLUSION & FUTURE WORK

The work we present here provides a much needed connection between quantitative diversity metrics and user perceptions of diversity in music recommendation lists. Through analysis of semi-structured interviews with 17 participants we identified two primary themes on user selections for diversity: listener mood, and diversity meaning. More specifically many users expressed a clear distinction between diversity within the bounds of their existing preferences, and diversity outside of these preferences. This inner and outer diversity was often expressed as a binary preference. Additionally, we found that when given the ability to select their own level of diversity in recommendation lists, user selections varied widely within and between subjects.

Much future work is required in order to generalize our qualitative findings to a larger population, and further inform *music recommender systems* from the user's perspective. Additionally, we plan to explore the connection between listener mood, and their preference for inner and outer diversity. Diversity in music recommendations should have at least as solid a foundation in user perception as in *information retrieval.*

7. REFERENCES

[1] K. Bradley and B. Smyth. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, pages 85–94. Citeseer, 2001.

[2] L. Chen, W. Wu, and L. He. How Personality Influences Users' Needs for Recommendation Diversity? In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2013-April, pages 829–834. Association for Computing Machinery, apr 2013.

[3] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*, pages 659–666, 2008.

[4] S. Deng, D. Wang, X. Li, and G. Xu. Exploring User Emotion in Microblogs for Music Recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.

[5] T. Di Noia, J. Rosati, P. Tomeo, and E. D. Sciascio. Adaptive multi-attribute diversity for recommender systems. *Information Sciences*, 382-383:234–253, mar 2017.

[6] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, pages 161–168, New York, New York, USA, oct 2014. Association for Computing Machinery, Inc.

[7] K. Farrahi, M. Schedl, A. Vall, D. Hauger, and M. Tkalčič. Impact of listening behavior on music recommendation. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014*, pages 483–488, 2014.

[8] B. Ferwerda, M. Graus, A. Vall, M. Tkalčič, and M. Schedl. The Influence of Users' Personality Traits on Satisfaction and Attractiveness of Diversified Recommendation Lists. In *Proceedings of the 4th workshop on emotions and personality in personalized services (EMPIRE 2016)*, pages 43–47, Boston, USA, 2016.

[9] B. Frederickson. Fast Python Collaborative Filtering for Implicit Datasets. `https://github.com/benfred/implicit`, 2019.

[10] A. Gunawardana and G. Shani. Evaluating recommender systems. In *Recommender Systems Handbook, Second Edition*, pages 265–308. 2015.

[11] Y. Hu, C. Volinsky, and Y. Koren. Collaborative filtering for implicit feedback datasets. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 263–272, 2008.

[12] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42, 2016.

[13] M. Kunaver and T. Požrl. Diversity in recommender systems – A survey. *Knowledge-Based Systems*, 123:154–162, 2017.

[14] J. H. Lee. How similar is too similar?: Exploring users perceptions of similarity in playlist evaluation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, pages 109–114, 2011.

[15] J. H. Lee and R. Price. Understanding users of commercial music services through personas: Design implications. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pages 476–482, 2015.

[16] J. H. Lee, L. Pritchard, and C. Hubbles. Can we listen to it together?: Factors influencing reception of music recommendations and post-recommendation behavior. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 663–669, nov 2019.

[17] A. L'Huillier, S. Castagnos, and A. Boyer. Understanding usages by modeling diversity over time. In *CEUR Workshop Proceedings*, volume 1181, pages 81–86, 2014.

[18] F. Lu and N. Tintarev. A diversity adjusting strategy with personality for music recommendation. In *CEUR Workshop Proceedings*, volume 2225, pages 7–14, 2018.

[19] S. Y. Park, A. Laplante, J. H. Lee, and B. Kaneshiro. Tunes together: Perception and experience of collaborative playlists. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 723–730, 2019.

[20] L. Porcaro and E. Gomez. 20 Years of Playlists : a Statistical Analysis on Popularity and Diversity. *Proceedings of the 20th International Symposium on Music Information Retrieval, ISMIR 2019*, (July):4–11, 2019.

[21] M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *RecSys'12 - Proceedings of the 6th ACM Conference on Recommender Systems*, pages 19–26, 2012.

[22] J. B. Schafer, J. A. Konstan, and J. Riedl. Meta-recommendation systems: User-controlled integration of diverse recommendations. In *International Conference on Information and Knowledge Management, Proceedings*, pages 43–51, 2002.

[23] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas. Music recommender systems. In *Recommender systems handbook*, pages 453–492. Springer, 2015.

[24] M. Schedl, H. Zamani, C. W. Chen, Y. Deldjoo, and M. Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, jun 2018.

[25] M. Slaney and W. White. Measuring playlist diversity for recommendation systems. In *Proceedings of the ACM International Multimedia Conference and Exhibition*, pages 77–82, New York, New York, USA, 2006. ACM Press.

[26] M. Tkalčič, A. Košir, and J. Tasič. Affective recommender systems: The role of emotions in recommender systems. In *CEUR Workshop Proceedings*, volume 811, pages 9–13, 2011.

[27] C. H. Tsai and P. Brusilovsky. Leveraging interfaces to improve recommendation diversity. *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 65–70, 2017.

[28] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, pages 209–216, 2014.

[29] M. C. Willemsen, B. P. Knijnenburg, M. P. Graus, L. C. Velter-Bremmers, and K. F. Eindhoven. Using latent features diversification to reduce choice difficulty in recommendation lists. In *CEUR Workshop Proceedings*, volume 811, pages 14–20, 2011.

[30] D. Wong, S. Faridani, E. Bitton, B. Hartmann, and K. Goldberg. The Diversity Donut: Enabling participant control over the diversity of recommended responses. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1471–1476, 2011.

[31] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor. Auralist: Introducing serendipity into music recommendation. In *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 13–22, 2012.

[32] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*, page 22, New York, New York, USA, 2005. Association for Computing Machinery (ACM).