Proceedings of the 1st Workshop on Designing Human-Centric Music Information Research Systems

2nd November, 2019
Delft, The Netherlands

ACM FAT* Network

Organization

Organizing Committee

Marius Miron	Joint Research Centre, European Commission
Vicky Charisi	Joint Research Centre, European Commission
Emilia Gómez	Joint Research Centre, European Commission

Program Committee

Perfecto Herrera	Universitat Pompeu Fabra, Barcelona
Bob Sturm	KTH Royal Institute of Technology, Stockholm
Arthur Flexer	The Austrian Research Institute for Artificial Intelligence, Vienna
Dmitry Bogdanov	Universitat Pompeu Fabra, Barcelona
Peter Knees	TU Wien
Andres Ferraro	Universitat Pompeu Fabra, Barcelona
Lorenzo Porcaro	Universitat Pompeu Fabra, Barcelona
Antonio Ramires	Universitat Pompeu Fabra, Barcelona
Augoustinos Tsiros	Music Tribe, UK

Support

The organization of this workshop has been possible thanks to the support from the European Commission's Joint Research Centre, Centre for Advanced Studies, HUMAINT project.

Proceedings of the 2nd International Workshop on Reading Music Systems, Delft, 2019

Edited by Marius Miron

Preface

Dear colleagues,

Technology and music have a centuries old history of coexistence: from luthiers to music information research. The emergence of machine learning for artificial intelligence in music technology has the potential to change the way music is experienced, learned, played and listened. This raises concerns related to its fair and transparent use, avoiding discrimination, designing sustainable experimental frameworks, and being aware of the biases the algorithms and datasets have.

The first edition of the Workshop Designing Human-Centric Music Information Research systems aims at bringing together people interested in discussing the ethical implications of our technologies and proposing robust ways to assess our system for discrimination, sustainability, and transparency.

We strongly believe that research on fairness, accountability, transparency advances through multidisciplinary research. Thus, this first edition hosts two keynotes talks which bring a refreshing perspective from two different fields, economics and human-computer interaction. First, Luis Aguiar, University of Zurich, presents "Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists". Second, Nava Tintarev, Delft University of Technology, presents "Supporting User Control for Music Recommendations".

We would like to thank our keynote speakers and the participants for their insightful presentations and for contributing to the discussion. Finally, we would like to thank Jaehun Kim and Ginny Ruiter who assisted us in organizing the venue.

Marius Miron, Vicky Charisi, Emilia Gómez

Contents

A PROPOSAL FOR A NEUTRAL MUSIC RECOMMENDER SYSTEM

Peter Knees
Faculty of Informatics
TU Wien
Vienna, Austria
peter.knees@tuwien.ac.at

ABSTRACT

In this paper, we propose an initiative for a neutral music recommender system for music consumption and research purposes, that acts as intermediary between the users and music streaming providers. Neutral is this context refers to the utility of the system in making recommendations and should result in the primary objective of satisfying users irrespective of potential side constraints. In contrast to existing individual systems that operate in multistakeholder environments and follow objectives beyond user satisfaction, our proposal entails accessing individual services via a joint proxy that manages an overall user profile and resorts to multiple services for provision of content with the primary goal of optimizing for user satisfaction. Such a system would not only provide a consolidated interface and consistent user experience over the union of accessible catalogs, but also give users increased levels of control over their data and the recommendation process.

1. MOTIVATION

Music recommender systems [10] are a central element in today's music consumption. They allow users to listen to the music they prefer and discover new music they might like from the vast catalogs of music streaming providers with minimal effort. However, while typical music streaming catalogs contain tens of millions of tracks, none covers all possible tracks and renditions. Furthermore, usage of a personalized system also bears the risks for users of disclosing personal and sensitive information and being targeted for advertising and marketing.

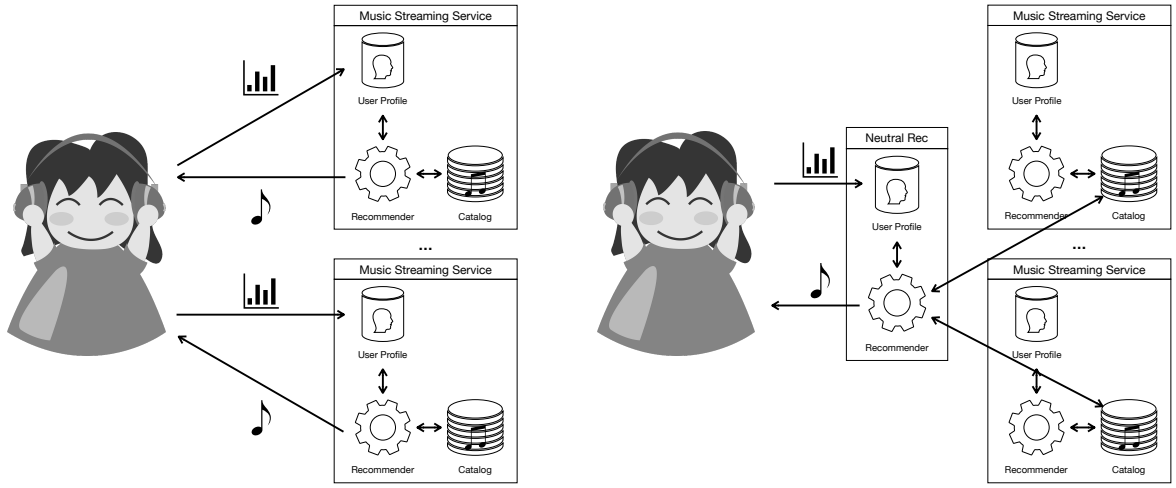
With an increasing number of players in the market and millions of users worldwide, a vast volume of interaction data and other traces of music listening is produced constantly. For academic research, typically, only small snapshots of this behavioral data are made available by commercial services for research purposes, limiting academic research to offline evaluation scenarios, cf. [11], or simulations of listening behavior under artificial online or lab conditions, e.g. [5]. Additionally, the circumstances under which this data came about are not always clear and—depending on the source, time span, and filtering—might exhibit a variety of biases (e.g., popularity, community, presentation, stakeholder, cf. [6]) possibly representative for a specific service or purpose but not for music listening per se.

In this position paper, we propose an architecture for an interface for music consumption and research purposes that acts as intermediary between the users and music streaming providers. The goal is to design this system with neutral utility (see Sec. 2), such that the system has the main objective of satisfying users irrespective of further constraints. Comparable to MovieLens for the movie domain [3], we aim at building a recommender system that collects preferences of users and matches them with accessible content across accessible service catalogs. Besides enabling access to a wider range of music in a consolidated interface and with a consistent user experience, such a system would give users increased levels of control over their data and the recommendation process. For academic research, such a system would allow deeper insights into music consumption behavior and provide an unprecedented in vivo experimentation setup.

2. MUSIC RECOMMENDATION AS MULTISTAKEHOLDER SCENARIO

Current music streaming platforms and their recommender systems operate in multistakeholder environments. As a consequence, they might follow objectives beyond pure user satisfaction, e.g., to optimize for revenue or preferences of content providers, cf. [2, 4, 8].

To model considerations of key stakeholders in recommender systems, Abdollahpouri et al. [1] have defined the multistakeholder framework. In the framework, utility and type of interaction of the key stakeholders C (consumer), P (provider of content), and S (system, i.e. the platform) can be expressed. In current music recommendation systems, the consumer has a personalized objective p with either passive or active interaction *; the provider has—presumably—a neutral objective n, as the provider has no preference in which users get which content, and a passive interaction $^-$, based on implicit feedback; and the system an aggregate or targeted utility a, t, as it gains from recommendations and has further objectives [2]; expressed as $\langle C_p^*, P_n^-, S_{a,t} \rangle$.



(a) Current music recommendation setting

(b) Aggregated and neutral music recommendation setting

Figure 1: Comparison of the current setting of music streaming with $1...n$ services and per-service recommendation (left) and the proposed setting with a neutral music recommender system acting as proxy and/or aggregator for $1...n$ different music services (right).

The goal of our proposal is to move from this situation to a situation expressed as $\langle C_p^*, P_n^-, S_n \rangle$, where the system has neutral utility n, resulting in the main objective of satisfying users regardless of side constraints. We set out to accomplish this transition by the idea and system architecture outlined next.

3. PROPOSED SYSTEM

In other domains, such as hotel bookings (Booking.com, trivago, etc.) or news recommendation (Google News, Blendle, etc.), meta-search portals facilitate access to multiple content providers via a consistent interface. Similarly, we propose to build a music streaming service aggregation platform, combining the catalogs of multiple platforms, while providing a unified experience based on the state of the art in research in both recommendation algorithms and user interface design.

Fig. 1 compares the current stakeholder-driven situation (left) with the envisioned neutral system (right). While in the current situation, a user subscribes to one or more services, each using an individual interface and storing a user profile for recommending music, in the envisioned scenario, the user registers $[1...n]$ accounts to services that allow access to their collection and unlimited streaming through an API, typically premium accounts.

In the following, we describe related efforts and discuss the implications of such a system from the perspectives of consumers, researchers, and services.

3.1 Related Work

Our proposed system resembles the MuSe system [9], a music recommendation management system, set out to provide an experimentation testbed for different music recommender algorithms. In contrast to MuSe, our approach is more holistic, as it presents itself to the user not as an evaluation testbed, but an integrated music streaming platform, aggregating the subscriptions and accounts of the user to develop a central user profile. As such, we expect the data collected to represent a less restricted selection and a more natural sample of real-world listening. This also entails the responsibility of user data management and providing a recommendation service that is competitive with or superior to existing platforms.

The Tomahawk project [1] is an open-source initiative to develop a multi-source and cross-platform music player. The goal is to combine a user's local library with free and commercial streaming services and enable playback from one of the accessible sources through an integrated interface. Connectivity with other users and simultaneous listening is another goal of the project.

ListenBrainz by the MetaBrainz Foundation [2] is an open-source project recreating an open data version of the AudioScrobbler functionality of Last.fm [3]. ListenBrainz allows users to import existing Last.fm listening profiles and stores public listening profiles with the goal of making it openly accessible for anyone interested, e.g. research initiatives [12].

3.2 Advantages for Consumers

Beside improved user experience, a neutral system offers several advantages for users. With the neutral system being in charge of giving recommendations, features often called for in academic research but not consistently implemented in commercial systems can be provided. Foremost this concerns matters of *transparency* and control, both regarding the functionality of the recommendation algorithm, as

[1] `https://github.com/tomahawk-player/tomahawk`
[2] `https://listenbrainz.org`
[3] `https://last.fm`

well as the user model and the data stored (cf. GDPR regulations within the European Union).

The latter aspect is an important feature and underlying motivation for this type of service, as usage of a personalized system also entails disclosing personal and potentially sensitive information through interaction, cf. [7]. This also bears the risk of being targeted for advertising and marketing. Recent work already addresses these risks and investigates the impact of fake plays on impeding prediction of personal information [13]. In case multiple services are registered with the neutral recommender, the system can choose the best source for playback, while diversifying sources. Hence, the more sources are available the less concentrated playback information is stored. In the broader context of digital humanism [14], the goals of privacy, transparency, and intervention with platform monopolies are important measures to be taken.

In a later step, the collected user data could be used for service recommendation as well, i.e., based on user preferences and the extrapolated recommendations, the user can be advised which alternative or additional service subscription would provide best value for the customer.

3.3 Advantages for Research

Maintaining a public service for media recommendation can have a positive impact on research as has been demonstrated in the movie domain through the MovieLens project by the GroupLens team of the University of Minnesota [3]. From MovieLens, we can learn that runnign a service does not only provide a source for the release of datasets consisting of snapshots of anonymized user profiles, but also a platform for testing academic recommender algorithms in production. In particular, the currently missing possibility of testing and comparing algorithms developed and optimized using offline data also in an online fashion through A/B testing could boost academic research in music recommendation significantly. In this regard, the MuSe system [9] can serve as a blueprint for our system.

3.4 Implications for Services

For services, implications are not necessarily advantageous. While services make use of the collected data for optimizing user experience (as well as other objectives, cf. sec. 2), they also explore collected data to optimize catalogs and learning about their user base. If services are reduced to their role as content providers and usage data can be recorded only partially or, in a later stage, is concealed by fake play events, service quality might be affected. As a consequence, if such a feature would be developed, dedicated concealment of listening information should not be the default setting but only be an opt-in feature for users. Users should therefore also be given the option of sharing their data with specific services.

4. FURTHER CHALLENGES

Apart from different challenges discussed throughout the paper, two other foreseen aspects should be briefly discussed. A major challenge of such a system would be *cold start*. To combat cold start, upon providing credentials and giving consent, existing data stored in user accounts should be imported into the user profile of the neutral system. By joining multiple user profiles in one central profile, we also pursue the goal of building user profiles that are not biased wrt. one specific service and perpetuating their patterns.

A non-trivial prerequisite for all these steps is to get listings of the available catalogs and find matchings between them. Unambiguous *meta-data matching* has been a challenge since the beginnings of music information retrieval research and is still not satisfactorily solved, especially, if direct access to content for fingerprinting is not possible. Keeping databases constantly updated for every service supported poses further challenges.

5. DISCUSSION

We proposed a music streaming aggregation service that effectively decouples recommendation of music content from its provision and delivery. This concept is not intended as an effort to be accomplished by one group but is inherently envisioned and designed to be carried by a broader initiative, to ultimately open up new opportunities for academic research. Required work does not only comprise the development and maintenance of the service under real-world usage conditions and requirements and management of user profiles, but also approval of requests for testing algorithms and scientific guidance of the process, among others. It should result in a less biased and neutral algorithmic testbed, as well as a music streaming interface providing the state of the art in music recommender systems research, a superior user experience, transparency in algorithms, and control over personal data.

It needs to be noted that a neutral system does not necessarily entail fair recommendations as other effects might impact this aspect, such as data biases and feedback loops. What is considered fair depends on the perspective of the stakeholder, e.g., for providers, fairness might be expressed as equal distribution of record labels in tracks streamed, cf. [2,6]; for artists, e.g., that there is no inherent discrimination based on gender. A definition of fairness from a user's perspective is still to be given. Diversity and novelty might play a role in this.

To conclude, not all undesirable effects of current streaming and recommendation services will be eliminated by the proposed system. However, the objective of being neutral is a prerequisite for mitigating data biases and researching further effects of recommender-driven music listening behaviour, therefore supporting the development and design of improved systems in the future.

6. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments, valuable input for discussion, and additional pointers to related work.

7. REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. arXiv:1905.01986, 2019.

[2] Himan Abdollahpouri and Steve Essinger. Multiple stakeholders in music recommender systems. arXiv:1708.00120, 2017.

[3] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, December 2015.

[4] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 7–10, New York, NY, USA, 2016. ACM.

[5] Iman Kamehkhosh and Dietmar Jannach. User perception of next-track music recommendations. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, pages 113–121, New York, NY, USA, 2017. ACM.

[6] Peter Knees and Moritz Hübler. Towards Uncovering Dataset Biases: Investigating Record Label Diversity in Music Playlists. In *Proceedings of the 1st Workshop on Designing Human-Centric MIR Systems*, Delft, The Netherlands, 2019.

[7] Thomas Krismayer, Markus Schedl, Peter Knees, and Rick Rabiser. Predicting user demographics from music listening information. *Multimedia Tools and Applications*, 78(3):2897–2920, 2019.

[8] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 2243–2251, New York, NY, USA, 2018. ACM.

[9] Martin Przyjaciel-Zablocki, Thomas Hornung, Alexander Schätzle, Sven Gauß, Io Taxidou, and Georg Lausen. MuSe: A music recommendation management system. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 543–548, Taipei, Taiwan, October 2014. ISMIR.

[10] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. Music recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Recommender Systems Handbook*, pages 453–492. Springer US, Boston, MA, 2nd edition, 2015.

[11] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, Jun 2018.

[12] Param Singh, Kamlesh Dutta, Robert Kaye, and Suyash Garg. Music listening history dataset curation and distributed music recommendation engines using collaborative filtering. In Pradeep Kumar Singh, Bijaya Ketan Panigrahi, Nagender Kumar Suryadevara, Sudhir Kumar Sharma, and Amit Prakash Singh, editors, *Proceedings of ICETIT 2019*, pages 623–632, Cham, 2020. Springer International Publishing.

[13] Kosetsu Tsukuda, Satoru Fukayama, and Masataka Goto. Listener Anonymizer: Camouflaging Play Logs to Preserve User's Demographic Anonymity. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 687–694, Paris, France, September 2018. ISMIR.

[14] Hannes Werthner et al. Vienna manifesto on digital humanism. https://www.informatik.tuwien.ac.at/dighum, May 2019.

ARTIST AND STYLE EXPOSURE BIAS IN COLLABORATIVE FILTERING BASED MUSIC RECOMMENDATIONS

Andres Ferraro, Dmitry Bogdanov, Xavier Serra
Music Technology Group -
Universitat Pompeu Fabra
first.lastname@upf.edu

Jason Yoon
Kakao Corp.
jason.yoon@kakaocorp.com

ABSTRACT

Algorithms have an increasing influence on the music that we consume and understanding their behavior is fundamental to make sure they give a fair exposure to all artists across different styles. In this on-going work we contribute to this research direction analyzing the impact of collaborative filtering recommendations from the perspective of artist and music style exposure given by the system. We first analyze the distribution of the recommendations considering the exposure of different styles or genres and compare it to the users' listening behavior. This comparison suggests that the system is reinforcing the popularity of the items. Then, we simulate the effect of the system in the long term with a feedback loop. From this simulation we can see how the system gives less opportunity to the majority of artists, concentrating the users on fewer items. The results of our analysis demonstrate the need for a better evaluation methodology for current music recommendation algorithms, not only limited to user-focused relevance metrics.

1. INTRODUCTION

There are multiple factors that make design decisions for a music recommendation system a complex problem. Some decisions can be related to theoretical aspects of music, while others may have ideological or social connotations, may be subjective, not possible to quantify, or be changing depending on time and context [13].

Collaborative filtering methods are typically used to generate a recommendation by identifying patterns in what people listen from historical information. The drawback of these methods is that since they do not consider any other than information about interactions between users and items, it is not possible to generate recommendations for new items (the cold-start problem). Also the recommendations tend to follow the distribution of popularity of the music [8] with the most popular items being recommended more (the long-tail recommendation problem). Celma and Cano [3] show this by analysing navigation, clustering and connectivity in artist similarity networks built with collaborative filtering data.

With the advances in deep learning, new methods had been proposed for long-tail and cold-start recommendations using audio information and metadata [12, 14], which can learn automatically a representation from the data without the need for manually selecting the features.

Still, these solutions have some issues, in particular related to the fact that they work as black-boxes. For example, it is difficult to explain the results and it is hard to know if different musical styles are well-represented. Also, previous works do not show how robust these methods are to biased datasets and if it is possible to generate recommendations for new styles or genres that are less present in the user-item interactions.

The growth of music streaming services in the last years has increased the importance of music recommender systems, and reducing the choice overload is commonly referred to as one of the advantages of these systems. Therefore, it is important to understand the increasing impact that these systems have to what people listen. They define which song will be the next hit, how much will an artist earn or even which music genres might receive almost zero promotion. This raises some ethical issues that had been discussed in previous works. For example, Holzapfel et al. [6] raise the question if a group of artists that are never recommended by a system can be considered a case of discrimination. As researchers, we have to think about the implications of the systems we develop and the importance of assuring every artist has a fair chance to reach the public [5].

Recently, there have been studies trying to address these issues. Cramer et al. [4] summarizes possible algorithmic biases and highlights that music recommendations for "balanced" not-biased consumption may not necessarily lead to optimal experience for many users. McInerney et al. [10] propose a bandits approach to balance exploration and exploitation in the recommendations for the users, but they do not address its impact on the exposure of different artists or music styles. Mehrotra et al. [11] proposes a way to understand the trade-off between relevance, satisfaction and fairness in music recommendations. In this case, fairness measures the diversity of the level of popularity of recommendations, but it does not capture the overall exposure of the artists or the different musical styles.

Following these studies, we demonstrate preliminary results of our on-going research that gives a better understanding of the influence of music recommendation systems on users' behavior that could affect artists' exposure. We show that the distribution of the recommendations in terms of their artists, styles or genres is different from what the users had listened before. Also, we show that with time the system tends to recommend fewer items, therefore, focusing user interactions on fewer artists, which is not the desired behavior of the system.

2. PROPOSED ANALYSIS

In this work, we use a basic Matrix Factorization [7] algorithm and *Echo Nest Profile Subset* to build a user-track matrix and generate 10 track recommendations for each user. We use the associated tags from *Last.fm Dataset* to analyze how recommendations are distributed across the different musical styles in comparison with listening statistics from our dataset representing the initial preferences of users. We also show a simulation of how these recommendations can affect user behavior in the long term. For this we take the recommendation of the system for each user and increase the counter in the original user-track matrix, simulating that the users listened to all recommendations by the system. We then retrain the model and generate new recommendations. We repeat this process 30 times.

2.1 Datasets

The Million Song Dataset (MSD) [9] is a large dataset of audio features and metadata expanded by the Music Information Retrieval community with additional information including tags, lyrics and other annotations. The Echo Nest Taste Profile Subset [2] provides play counts by 1,019,318 users covering 384,546 songs from MSD, originally gathered from an undisclosed set of applications. For this work we only consider users and items with more than 30 interactions (128,374 tracks by 18,063 artists and 445,067 users), to make sure we have enough information for training and evaluating the model. Additionally, the Last.fm Dataset [2] provides song-level tags extracted from *Last.fm* for a subset of MSD. These tags are crowdsourced and cover genre, instrumentation, moods and eras. One track can have multiple tags.

2.2 Metrics

For a better understanding of system behavior, we need to define metrics that can assess how probable it is for new or less popular artists to be recommended and compare those across different styles. It is also valuable to know to how many different users each artist is recommended.

In this work, we use the *Gini index* to measure the distribution of how many users each artist gets recommended to, but in future works other metrics should be also considered (for example, proposed for multistakeholder recommendation approaches [1]).

We also use *Coverage* to measure the percentage of different artists globally recommended. With this metric we can have an idea of the amount of artists that the system gives zero promotion.

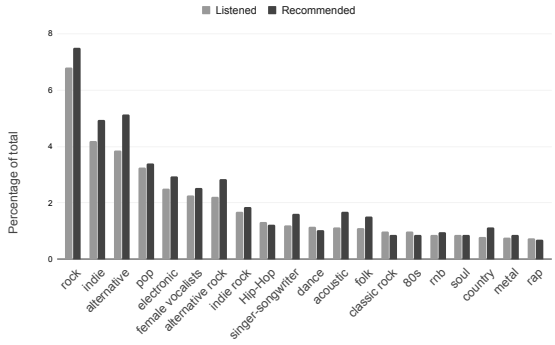


Figure 1: Distribution of recommendations and users listening. Values are average percentages per music style.

3. RESULTS

3.1 Distribution of recommendations

Figure 1 shows the global tag distribution of all user-track recommendations pairs (10 tracks per user) compared to such a distribution for initial user listening behavior for the top 20 tags. [1] For the rest of the tags, the system is recommending 9.4% less compared with what the users listened to. Table 1 similarly reports an average percentage of recommendations and initial user preferences in three tag and artists categories grouped by their popularity in terms of the original play counts.

We can see a clear popularity bias in what users listen to, and this bias is further reinforced by recommendations, which may be not the desired behavior. The system is recommending more top tags and less long-tail tags than what people listened to.

	Tags 1-5	**Tags 5-2k**	**Tag 2k-50k**
Recommended	**4.7807**	**0.0347**	0.0001
Listened	4.1195	0.0327	**0.0003**

(a) Tags

	Artists 1-5	**Artists 5-2k**	**Artists 2k-18k**
Recommended	**1.5672**	**0.0433**	0.0003
Listened	0.6182	0.0370	**0.0014**

(b) Artists

Table 1: Average percentage of recommendations and user play counts for (a) tags and (b) artists with different popularity.

3.2 Simulating feedback loops

Figure 2 shows the results of simulating the feedback loop of the recommendations. We can see how the *Gini index*

[1] Note that there are 51,699 tags and therefore it is not possible to show all of them.

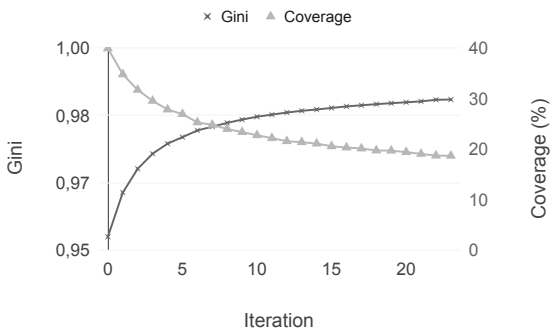


Figure 2: Coverage and Gini index of the recommendations simulating feedback loops.

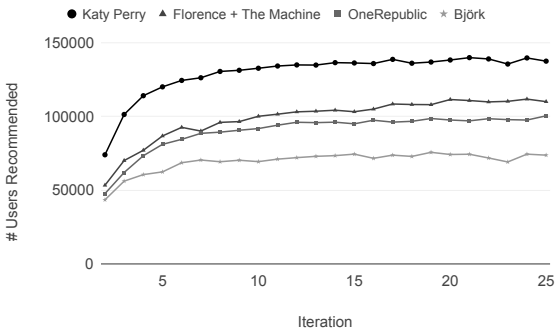


Figure 3: Number of users reached by a song recommendation on the example of four popular songs when simulating the feedback loop.

increases on each iteration, starting in 0.95 and going up to 0.98. A value of 1.0 indicates that the system is recommending the same songs to all users. In the same figure we see the evolution of the *Coverage* of the recommendations. For the first iteration the *Coverage* is 40 % but at the last iteration the *Coverage* is 20 % meaning that 80 % of the songs are not recommended by the system.

In Figure 3 we demonstrate how the four most played songs according to our initial user-track matrix gather even more exposure from recommendations during the feedback loop iterations. These songs have been recommended to between 50,000 and 100,000 users at the first iteration, and ended up being recommended to 100,000 to 135,000 users after 10 iterations, while originally they were listened by around 50,000 users.

It is important to mention that in a real case there will be other interactions between users and items that are not considered here.

4. CONCLUSIONS

In this work, we have considered how the popularity bias is affecting collaborative filtering recommendations based on Matrix Factorization. In our experiments, this algorithm is increasing the exposure of more popular musical styles, while reducing the exposure in the long tail, which may be an undesired behaviour.

The goal of our future research is to expand our analysis on state-of-the-art algorithms proposed for cold-start and long-tail music recommendation, which are still lacking such an evaluation.

5. ACKNOWLEDGEMENTS

This research has been supported by Kakao Corp.

6. REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint arXiv:1905.01986*, 2019.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proc. of the 12th International Society for Music Information Retrieval Conf. (ISMIR)*, 2011.

[3] Òscar Celma and Pedro Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. ACM, 2008.

[4] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conf. on Human Factors in Computing Systems*, 2019.

[5] Andres Ferraro. Music cold-start and long-tail recommendation: bias in deep representations. In *Proceedings of the 13th ACM Conf. on Recommender Systems*, 2019.

[6] Andre Holzapfel, Bob Sturm, and Mark Coeckelbergh. Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, 2018.

[7] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conf. on Data Mining*, 2008.

[8] Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2013.

[9] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. The million song dataset challenge. In *Proc. of the 21st International Conf. on World Wide Web (WWW)*, 2012.

[10] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proc. of the 12th ACM Conf. on Recommender Systems (RecSys)*, 2018.

[11] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proc. of the 27th ACM International Conf. on Information and Knowledge Management*, 2018.

[12] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. A deep multimodal approach for cold-start music recommendation. In *Proc. of the 2nd Workshop on Deep Learning for Recommender Systems*, 2017.

[13] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. Music recommender systems. In *Recommender systems handbook*, pages 453–492. Springer, 2015.

[14] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, 2013.

MUSIC RECOMMENDATION DIVERSITY: A TENTATIVE FRAMEWORK AND PRELIMINARY RESULTS

Lorenzo Porcaro[1], Carlos Castillo [2], Emilia Gomez[1,3]
[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Web Science and Social Computing Group, Universitat Pompeu Fabra, Barcelona, Spain
[3]Joint Research Centre, European Commission, Seville, Spain
{lorenzo.porcaro,carlos.castillo,emilia.gomez}@upf.edu

ABSTRACT

Music recommendations are increasingly part of the listening experience of people all over the world, especially in the context of streaming services. In this scenario, recommender systems' role is to help users in finding music that can fit their interests and tastes. However, Western-centric perspectives in systems' design are often subject to criticisms because of their power of reinforcing already existing cultural bias and therefore potentially impacting negatively on the music distribution mechanisms. In our research proposal, we aim to address the problem of assessing the impact of music recommendation diversity, or the lack thereof. This requires 1) the formalization of a working definition of diversity in the music field 2) the development of evaluation practices for estimating diversity in the context of music recommender systems 3) the observation of emerging impact due to music recommendation diversity 4) the proposal of countermeasures for mitigating negative or reinforcing positive impact observed. Basing on already known consequences of information technologies in political, economic and social areas, our goal is to understand the cultural impact that music recommender systems can have on our society.

1. INTRODUCTION

Our age has been dominated by the advent of streaming services, and music, or more generally audio-visual content, is a cultural product for which the enjoyment and production have been extensively influenced and re-shaped in the last century. As Benjamin at the beginning of the 20th century underlined in his "*Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit*" ("The Work of Art in the Age of Mechanical Reproduction"), the dual identity of music, which influences and is influenced by society, is strictly defined by the historical circumstances, and fundamental are the relationships between the work of art, the medium and human perception [2].

The widespread use and easy accessibility of streaming platforms is an important achievement for today's music listeners, which are one click away from the possibility to enjoy their favourite music and to discover new artists. In this panorama, recommender systems play a key role as bridge between users and the large amount of content accessible in digital repositories. Indeed, to browse the large catalogue of tracks in streaming platforms is a task that may reveal truly complicated for a user [10]. Music recommender systems represent a fraction of the wide family of recommender systems, often used in streaming services for tasks such as playlists continuation or the creation of personalized music radios [24]. Consequently, most of the users' listening behaviours are nowadays subject to different layers of algorithmic-generated interactions, accomplishment that can bring along with it opaque social, economic and cultural issues [12, 33].

In our work, we are interested in assessing how the diversity, which characterizes music from different parts of the world, is represented within recommender systems, and what can be the consequences of diversity in music recommendations, or the lack thereof. We argue that to preserve the richness of musical cultures is undoubtedly a crucial mission of modern technologies, and recommender systems have a critical role in that because of their impact on music distribution [8].

The document is structured as follows. Section 2 provides an overview of the previous work on Recommender Systems (RS) and its relationships with the concept of diversity, and also how this concept has been approached in the Music Information Retrieval (MIR) field. Afterwards, in Section 3 we present the research goals identified in our research, followed by preliminary outcomes in Section 4. Finally, conclusions and future work are discussed in Section 5.

2. RELATED WORK

2.1 Diversity in Recommender Systems

First concerns about the lack of user-centric perspectives in the design of recommendation systems, and related evaluation metrics, start to emerge around two decades ago [19]. In particular, the focus on improving the systems' per-

formance looking exclusively at the accuracy started to be criticized, and alternative evaluation frameworks took hold, introducing the so-called *beyond-accuracy* metrics. In the literature, the most established of these metrics are diversity, serendipity, novelty, and coverage, and we refer to [11] for an extended survey and analysis of these metrics. Following, we focus on diversity, the central topic of our work.

The need to adopt diversification strategies, i.e."*to identify a list of items that are dissimilar with each other, but nonetheless relevant to the user's interests*" [34], is often related to counteract the so-called *portfolio effect*, a situation where recommended items are highly similar to the ones targeted by the user, described by Burke in [4]. At the same time, it justified the emergence of new evaluation metrics, such as the Intra-List Similarity presented by Ziegler et al. in [36], where the items of a list are compared between each other for estimating the general degree of the list dissimilarity. However, when trying to improve the performance of recommender systems in terms of diversity metrics, the accuracy of these systems started to be challenged. This led to the emergence of the *diversity-accuracy dilemma*, hence how to balance the trade-off between diversity and accuracy. Several techniques have been proposed for solving this trade-off, for instance using directed random walks [16], or hybrid algorithms [35].

The multiple facets of diversity have implied the urgency for considering a wide conceptualization and different measurements, therefore different approaches also in the recommender systems literature have tried to tackle the problem of tuning the degree of diversity of recommendation lists. Among the others, in our research line we see links with works on temporal diversity [15], intent-oriented diversity [32], genre diversity [31], and multi-attribute diversity [6]. For an extended review of the approaches to diversity in recommender systems, we refer to [13].

2.2 Diversity in Music Information Retrieval

Our understanding of music diversity is partly based on the analysis of the semiology of music done by Molino in [20]. Indeed, what the author argues is that historical circumstances lead to a constant process of *aesthetic revolution*, making difficult to appear processes of standardization of the music languages. Even if subject to an increasing influence of the mass industry, music symbolic evolution tends to preserve its nature of constant diversification process. In a parallel direction, when analyzing the relationship between Western Music and other musical cultures ("*its Other*"), Born and Hesmondhalgh in [3] underline that 1) the mutual influence that different musical cultures have on each other, and 2) the development of these cultures in a system with specific socio-economic characteristics, are two processes strictly related.

In the field of MIR, diversity has often been examined in relation to musical tastes, hence to the aesthetic domain i.e. the more diverse is the music that you like, the more diverse are your musical taste. From another perspective, diversity, intended as differences in cultural music traditions, has also been at the centre of attention when dealing with information technologies built for extracting knowledge from the poietic domain i.e. the analysis of creative processes. In this direction, the work started by Serra in 2011 in the context of the project CompMusic [1] is an outstanding example of the need for including diverse approaches when tackling the multicultural reality of the music of the world [26].

Music recommender systems research has focused on the problem of understanding and representing diversity at two levels. At an individual level, it has been shown how personal traits and listening behaviours influence users' need for diversified recommendations [18, 29]. From another perspective, by means of cross-country analysis relationships between diversity and musical preferences have been investigated, aiming at evidencing how cultural differences in geographical regions can also be reflected in the listening experiences [7, 17]. These research lines reflect two of the future research directions of music recommender systems research identified by Schedl et al. in [24]: *Psychologically-inspired music recommendation* and *Culture-aware music recommendation*.

3. RESEARCH GOALS

The work of diversity assessment designed is structured in four main goals. In the beginning, the main attention is posed on the development of theoretical instruments for defining and evaluating diversity in the area researched. Subsequently, the focus is shifted in understanding the consequences of music recommendation diversity from a human-centric perspective, as described in the next sections.

3.1 Develop a Framework for Defining and Evaluating Music Recommendation Diversity

The first research goal identified is the creation of a framework for defining and evaluating music recommendation diversity, specific to the MIR and RS fields. A framework that must assure: 1) a solid theoretical background; 2) the development of analysis tools (toolkit, software, etc.). Given the not univocal nature of diversity, as discussed by Stirling in [27], our starting point is to understand how diversity has been approached in different literatures.

3.2 Assess Music Recommendation Diversity

The next step is to research how music recommender systems behave in different scenarios, focusing on the two classic dimensions of recommender systems: user and item. By making explicit the questions that we hypothesize, can we affirm that a recommendation list is more diverse than another one? Can we compare recommendation lists created by different systems, and affirm which system embeds more diversity? Can we compare recommendation lists created for different users, and stating if a list is more diverse than the other? Undoubtedly, the pitfalls contained

[1] https://compmusic.upf.edu

in these questions are several. At the end of this phase, we aim at being able to compare different music recommender systems and to evaluate how tuning systems settings can influence the outcome diversity.

3.3 Understand the Consequences of Music Recommendation Diversity

What we are most interested in is to understand what might be the societal impacts of using recommendation technologies in the context of cultural development. Starting from already-known negative effects on society (such as filter bubbles [21], echo chambers [28], and cyberbalkanization [30]), our objectives are: 1) to understand if consequences found in other areas can be reproduced while analyzing music recommender systems; 2) if others consequences can be found, specifically related to music field.

3.4 Propose Countermeasures for Tuning Music Recommendation Diversity

The final part of this work will target the consequences and impacts found, proposing novel methods for contrasting the counter effects of these technologies on humans. It is difficult to identify *a priori* the techniques to be developed, considering the wide range of scenarios that might derive from the research planned. However, we aim at including a complete spectrum of outcomes, from negative ones, where recommender systems are proven to damage human beings, to positive ones, where on the contrary the use of this technology can improve the well-being and can be used for the social good.

4. PRELIMINARY OUTCOMES

Following the research line previously defined, exploratory experiments have been carried out, thanks to which it has been possible to achieve two initial results. On one hand, in [22] we were able to explore standard diversity measures from the Information Theory literature, applying them for a comparative analysis of playlist datasets. On the other hand, in [23] we made a first attempt of proposing new measures for evaluating the variations of recommendation lists in different scenarios.

Both studies have target items related to recommender systems framework, playlist in [22] and recommendation list in [23], and through the use of information retrieval, statistics, and mathematical modelization techniques, we have evaluated a degree of diversity, statically in the first case whereas dynamically in the second. In these studies, we centered our attention on two dimensions often considered in the MIR literature: popularity, or mainstreamness [1, 5], and semantic information [14].

In [22], characterizing and comparing four playlist datasets created in different historical and technological contexts, we notice the emergence of diverse patterns in the users' grouping choices. Similarly, by means of the comparison of analog and streaming radios, in [23] we started to tackle the limitation of specific metrics for evaluating music recommendation diversity.

5. CONCLUSIONS AND FUTURE WORK

This paper provides an initial formalization of the research work plan and first results on the study of the impact of music recommendation diversity. Until now, few studies in the MIR literature have focused on comprehending what have been the consequences of introducing music recommendation technologies in the actual cultural context. Using the terminology introduced by Selbst et al. in [25], this might cause a *Ripple-Effect Trap*, defined as a "*Failure to understand how the insertion of technology into an existing social system changes the behaviours and embedded values of the pre-existing system*". With our research, we aim at raising awareness in the MIR field for avoiding to fall in this abstraction trap while designing and implementing music recommender systems.

Parallelly, another challenge is to define what diversity can represent, and how it can be represented in the music recommendation field. In 2004, Huron in his article *Issues and Prospects in Studying Cognitive Cultural Diversity* [9], included a call for action where he states:

> We[*music researchers*] should be concerned about the loss of cultural diversity for the same reason that biologists worry about the loss of biodiversity: we don't yet know what the loss will mean, but we do know that the loss will be irreversible.

Even if the comparison between cultural diversity and biodiversity might be challenged in some aspects, the identification as a main problem of the unknown consequences of the loss of diversity is part of the motivation of this work. Indeed, we imagine that the lack of representation or the misrepresentation of diversity potentially could lead to unfair treatment in music distribution, which in the worst scenario might lead to cultural discrimination.

Apart from that, our choice to focus on diversity as sociotechnical concept is also driven by recent debates about the impact of Artificial Intelligence (AI) systems on human behaviour and the related ethical, social, economic and legal issues. Imagining recommender systems as part of a broader field which can be AI, the importance of considering diversity is proven by its inclusion within the list of seven key requirements that AI systems should meet in order to be trustworthy, proposed in the *Ethics Guidelines for Trustworthy Artificial Intelligence*, written by the High-Level Expert Group on Artificial Intelligence [2].

Finally, our motivations are also partly reflected in the less recent but fundamental debate about the importance of the preservation of cultural diversity, which in 2001 has led to the *UNESCO Universal Declaration on Cultural Diversity* [3]. In this document, it is emphasized how cultural diversity is a vehicle by which promoting pluralism, human rights, international solidarity and also, creativity.

[2] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Diversity

[3] http://portal.unesco.org/en/ev.php-URL_ID=13179&URL_DO=DO_TOPIC&URL_SECTION=201.html

6. ACKNOWLEDGMENTS

This work is partially supported by the European Commission under the TROMPA project (H2020 770376).

7. REFERENCES

[1] Christine Bauer and Markus Schedl. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE*, 14(6):1–36, 2019.

[2] Walter Benjamin. *The Work of Art in the Age of Mechanical Reproduction*. Hannah Arendt, ed., Illuminations. London: Fontana., 1968.

[3] Georgina Born and David Hesmondhalgh. Introduction: On Difference, Representation and Appropriation in Music. In *Western Music and Its Others: Difference, Representation, and Appropriation in Music*, pages 1–58. University of California Press, 2000.

[4] Robin Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[5] Óscar Celma. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer-Verlag Berlin Heidelberg, 2010.

[6] Tommaso Di Noia, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. Adaptive multi-attribute diversity for recommender systems. *Information Sciences*, 382-383:234–253, 2017.

[7] Bruce Ferwerda and Markus Schedl. Investigating the relationship between diversity in music consumption behavior and cultural dimensions: A cross-country analysis. In *Proceedings of the 1st Workshop on Surprise, Opposition, and Obstruction in Adaptive and Personalized Systems (SOAP) co-located with 24th Conference on User Modeling, Adaptation, and Personalization (UMAP)*, 2016.

[8] Andre Holzapfel, Bob L. Sturm, and Mark Coeckelbergh. Ethical Dimensions of Music Information Retrieval Technology. *Transactions of the International Society for Music Information Retrieval*, 1:44–55, 2018.

[9] David Huron. Issues and Prospects in Studying Cognitive Cultural Diversity. In *Proceedings of the 8th International Conference on Music Perception & Cognition*, number August, 2004.

[10] Sheena S. Iyengar and Mark R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995–1006, 2000.

[11] Marius Kaminskas and Derek Bridge. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1):1–42, 2016.

[12] Nedim Karakayali, Burc Kostem, and Idil Galip. Recommendation Systems as Technologies of the Self: Algorithmic Control and the Formation of Music Taste. *Theory, Culture & Society*, 35(2):3–24, 2018.

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

[14] Paul Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[15] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, page 210, 2010.

[16] Jian Guo Liu, Kerui Shi, and Qiang Guo. Solving the accuracy-diversity dilemma via directed random walks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(1), 2012.

[17] Meijun Liu, Xiao Hu, and Markus Schedl. Artist Preferences and Cultural, Socio-Economic Distances Across Countries: a Big Data Perspective. In *Proceedings of the 18th International Conference on Music Information Retrieval, ISMIR 2017*, 2017.

[18] Feng Lu and Nava Tintarev. A diversity adjusting strategy with personality for music recommendation. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2018)*, pages 7–14, 2018.

[19] Sean M. McNee, John Riedl, and Joseph A. Konstan. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2006)*, pages 1097 – 1101, 2006.

[20] Jean Molino and Craig Ayrey. Musical Fact and the Semiology of Music. *Music Analysis*, 9(2):105–111;113–156, 1990.

[21] Eli Parisier. *The filter bubble: What the Internet is hiding from you*. Penguin Press, New York, 2011.

[22] Lorenzo Porcaro and Emilia Gómez. 20 Years of Playlists: A Statistical Analysis on Popularity and Diversity. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR'19)*, 2019.

[23] Lorenzo Porcaro and Emilia Gómez. A Model for Evaluating Popularity and Semantic Information Variations in Radio Listening Sessions. In *Proceedings of the 1st Workshop on the Impact of Recommender Systems (ImpactRS), at the 13th ACM Conference on Recommender Systems (RecSys 2019)*, 2019.

[24] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, 2018.

[25] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, volume 1, pages 59–68, 2018.

[26] Xavier Serra. A Multicultural Approach in Music Information Research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11)*, pages 151–156, 2011.

[27] Andy Stirling. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, 2007.

[28] Cass Sunstein. *Echo Chambers: Bush v. Gore Impeachment, and Beyond*. Princeton University Press, 2001.

[29] Nava Tintarev, Christoph Lofi, and Cynthia C.S. Liem. Sequences of Diverse Song Recommendations: An exploratory study in a commercial system. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 391–392, 2017.

[30] Marshall Van Alstyne and Erik Brynjolfsson. Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. *Management Science*, 51(6):851–868, 2005.

[31] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. Coverage, redundancy and size-awareness in genre diversity for recommender systems. *Proceedings of the 8th ACM Conference on Recommender systems*, (October):209–216, 2014.

[32] Saúl Vargas, Pablo Castells, and David Vallet. Intent-Oriented Diversity in Recommender Systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pages 1–2, 2011.

[33] Patrick Vonderau. The Spotify Effect: Digital Distribution and Financial Growth. *Television and New Media*, 20(1):3–19, 2019.

[34] Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. It takes variety to make a world: Diversification in recommender systems. In *Proceedings of International Conference on Extending Database Technologies (EDBT) 2009*, number June 2014, page 368, 2009.

[35] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[36] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *WWW'15 Conference*, page 22, 2005.

ALLOWING FOR EQUAL OPPORTUNITIES FOR ARTISTS IN MUSIC RECOMMENDATION: A POSITION PAPER

Christine Bauer
Johannes Kepler University Linz
Institute of Computational Perception
christine.bauer@jku.at

ABSTRACT

Promoting diversity in the music sector is widely discussed on the media. While the major problem may lie deep in our society, music information retrieval contributes to promoting diversity or may create unequal opportunities for artists. For example, considering the known problem of popularity bias in music recommendation, it is important to investigate whether the short head of popular music artists and the long tail of less popular ones show similar patterns of diversity—in terms of, for example, age, gender, or ethnic origin—or the popularity bias amplifies a positive or negative effect.

I advocate for reasonable opportunities for artists—for (currently) popular artists and artists in the long-tail alike—in music recommender systems. In this work, I represent the position that we need to develop a deep understanding of the biases and inequalities because it is the essential basis to design approaches for music recommendation that provide reasonable opportunities. Thus, research needs to investigate the various reasons that hinder equal opportunity and diversity in music recommendation.

1. INTRODUCTION

Creating and maintaining diversity is an important and widely discussed topic in our society [27]. Thereby the debates on diversity are dominated by the challenges in promoting diversity as our society is prone to lay ground for unequal opportunities with respect to, for example, age, disability, gender, ethnic origin, religion, or sexual orientation throughout our society.

The issue of unequal opportunities is also relevant and a highly topical subject in the music sector. Some people voiced their concerns that there is a general discrimination of female artists [3, 16, 20, 24]. A similar inequality problem exists with respect to the little representation of black artists (especially black female artists) in high-popularity playlists on online music platforms [19, 20].

While the major problem may lie far beneath online music platforms or the music sector at large, the vast possibilities of music information retrieval and recommendation may contribute tremendously in promoting diversity, inclusion, and equity—but may also be used to (intentionally or unintentionally) create unreasonable imbalances.

For instance, it is widely known that algorithms used for music recommendation are frequently prone to popularity bias [12]. This is a burden to inclusion, as such algorithms prioritize popular items and almost disregard the long tail of less popular items. In other words, the spectrum of suggested items is limited to a proportionally small set of items. As popularity bias is a common phenomenon in algorithmic filtering, research came up with diversity measures [15, 23] and there are various attempts to introduce diversity to recommendation algorithms [5, 6]. Studies (e.g., [9]) have shown that an increase in diversity has a positive effect on user experience, while the ideal degree of diversity may depend on user characteristics [10, 13, 21].

I postulate that we need to develop a deep understanding of the biases and inequalities because it is the essential basis to design approaches for music recommendation that are free from undesired biases and inequalities.

When we take a human-centric approach to music information retrieval (MIR), we need to consider all kinds of roles involved in MIR—not just the user. In this work, I put the—previously neglected—artists' perspective in the loop. With the goal to provide reasonable opportunities for artists—for (currently) popular artists and artists in the long-tail alike—in music recommender systems, I take the position that research needs to investigate the various reasons that hinder equal opportunity and diversity in music recommendation.

This position paper is structured as follows: Section 2 presents the complexity of bias in music recommendation. Section 3 puts the artists' perspective into the loop. Section 4 presents the fundamental research questions that have to be addressed to allow for equal opportunity for artists and promoting diversity in music consumption.

2. THE COMPLEXITY OF BIAS

Music recommendation relies on algorithmic decision-making. And an emerging body of literature has shown that algorithmic decision-making can go wrong in multiple ways [25], due to algorithmic problems, data sparsity, or actors gaming the system (e.g., via click manipulation). Typical problems include popularity bias, cold start prob-

lem, shilling attacks, grey-sheep problem, synonymy, as well as scalability and latency problems [14]. This leads to severe problems for society—from filter bubbles [18] to the reproduction and amplification of stereotypes and discrimination [22] to cognitive bias and humans' overconfidence in algorithmic results [11]. Addressing these problems, there is a growing body of literature on fairness, accountability, and transparency in machine learning and artificial intelligence [2, 7, 17].

Still, while some aspects of bias in data and algorithms are subject of interest in research and draw attention on the media (e.g., filter bubble and popularity bias), other biases are not addressed or may even not have been identified yet.

3. THE NEGLECTED ARTIST IN MUSIC INFORMATION RETRIEVAL

In the music information retrieval (MIR) community (and related communities), research on diversity typically takes the perspective of the system—for instance, to mitigate the cold-start problem [4]—or the user (here: music consumer)—to better meet user preferences [13, 26]. The perspective of the item suppliers is considered only occasionally. For instance, Reference [1] raise awareness that recommender systems in multi-stakeholder environments may be fair for one stakeholder while being unfair for other stakeholders. Reference [8] proposes an approach with the goal to provide all artists in a collection with the opportunity of being listened in recommendations. Taking a human-centric approach to MIR systems, the goal is to include the artists' perspective in MIR research.

4. RESEARCH DIRECTIONS

Taking a human-centric perspective with the aim to allow for equal opportunity for artists and promoting diversity in music consumption, requires to address fundamental research questions concerning potential bias in current systems and, generally, in music consumption. For instance:

Research Question 1. *How is diversity in terms of, for example, age, disability, gender, ethnic origin, religion, or sexual orientation of artists represented in the long tail of the popularity distribution?*
How is diversity represented in the short head of popular artists?
How does the diversity in the long tail and the short head relate to each other, and to the entire population?

Research Question 2. *How does the popularity of music items reflect inherent user taste?*
How is the popularity of music items affected by what is offered on online music platforms, on playlist, in recommendations, in advertising, etc.?

Understanding bias is a prerequisite to address its various facets and mitigate them. One concrete research question could be formulated as follows:

Research Question 3. *What is the influence of using timbre of the singing voice for music recommendation on the artist gender distribution in recommended items?*
If recommendations allow for little diversity in timbre, items will likely be sung by same-gender singers.

Overall, the goal of future work is to investigate the various facets reasons that hinder equal opportunity and diversity in music recommendation. A deep understanding of the biases and inequalities is the essential basis to design approaches for music recommendation that provide reasonable opportunities for artists—for (currently) popular artists and artists in the long-tail alike.

5. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF): V579.

6. REFERENCES

[1] Himan Abdollahpouri and Robin Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness. In Robin Burke, Himan Abdollahpouri, Edward Malthouse, KP Thai, and Yongfeng Zhang, editors, *Proceedings of the Workshop on Recommendation in Multistakeholder Environments (RMSE 2019)*, number 2440 in CEUR Workshop Proceedings, Aachen, Germany, 2019. http://ceur-ws.org/Vol-2440/paper3.pdf.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *2018 CHI Conference on Human Factors in Computing Systems*, CHI'18, New York, NY, USA, 2018. ACM.

[3] Pauwke Berkers and Julian Schaap. Gender inequality in metal music production. *Emerald Studies in Metal Music and Culture*, pages 145–149, 2018.

[4] K. R. Bindu, Rhama Lalgudi Visweswaran, P. C. Sachin, Kundavai Devi Solai, and Soundarya Gunasekaran. Reducing the cold-user and cold-item problem in recommender system by reducing the sparsity of the sparse matrix and addressing the diversity-accuracy problem. In Nilesh Modi, Pramode Verma, and Bhushan Trivedi, editors, *Proceedings of International Conference on Communication and Networks*, pages 561–570, Singapore, 2017. Springer.

[5] Keith Bradley and Barry Smyth. Improving recommendation diversity. In *Proceedings of the 12th National Conference in Artificial Intelligence and Cognitive Science*, AICS'01, pages 75–84, 2001.

[6] Jinpeng Chen, Yu Liu, Jun Hu, Wei He, and Deyi Li. A novel framework for improving recommender diversity. In *Behavior and Social Computing*, pages 129–138, Cham, Germany, 2013. Springer.

[7] Amitai Etzioni and Oren Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4):403–418, 2017.

[8] Andres Ferraro. Music cold-start and long-tail recommendation: Bias in deep representations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys'19, pages 586–590, New York, NY, USA, 2019. ACM.

[9] Bruce Ferwerda, Mark P. Graus, Andreu Vall, Marko Tkalcic, and Markus Schedl. How item discovery enabled by diversity leads to increased recommendation list attractiveness. In *Proceedings of the Symposium on Applied Computing*, SAC'17, pages 1693–1696, New York, NY, USA, 2017. ACM.

[10] Bruce Ferwerda, Andreu Vall, Marko Tkalcic, and Markus Schedl. Exploring music diversity needs across countries. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP'16, pages 287–288, New York, NY, USA, 2016. ACM.

[11] George Hurlburt. How much to trust artificial intelligence? *IT Professional*, 19(4):7–11, 2017.

[12] Dietmar Jannach, Lukas Lerche, Fatih Gedikli, and Geoffray Bonnin. What recommenders recommend: An analysis of accuracy, popularity, and sales diversity effects. In *21st International Conference on User Modeling, Adaptation, and Personalization*, UMAP'13, pages 25–37, Berlin Heidelberg, Germany, 2013. Springer.

[13] Yucheng Jin, Nava Tintarev, and Katrien Verbert. Effects of individual traits on diversity-aware music recommender user interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP'18, pages 291–299, New York, NY, USA, 2018. ACM.

[14] Shah Khusro, Zafar Ali, and Irfan Ullah. Recommender systems: Issues, challenges, and research opportunities. In *Information Science and Applications*, ICISA'16, pages 1179–1189, Singapore, 2016. Springer.

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

[16] Naomi Larsson. Live music acts are mostly male-only. what's holding women back? *The Guardian*, October 2017. https://www.theguardian.com/inequality/2017/oct/12/tonights-live-music-acts-will-mostly-be-male-only-whats-holding-women-back.

[17] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[18] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, London, United Kingdom, 2011.

[19] Vaughn Schmutz and Alison Faupel. Gender and cultural consecration in popular music. *Social Forces*, 89(2):685–707, 2010.

[20] Stacy L. Smith, Marc Choueiti, and Katherine Pieper. Inclusion in the recording studio?: Gender and race/ethnicity of artists, songwriters & producers across 600 popular songs from 2012–2017. Report, Annenberg Inclusion Initiative, 2018. http://assets.uscannenberg.org/docs/inclusion-in-the-recording-studio.pdf.

[21] Nava Tintarev, Matt Dennis, and Judith Masthoff. Adapting recommendation diversity to openness to experience: A study of human behaviour. In Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro, editors, *User Modeling, Adaptation, and Personalization*, UMAP'13, pages 190–202, Berlin Heidelberg, Germany, 2013. Springer.

[22] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260, 2018.

[23] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys'11, pages 109–116, New York, NY, USA, 2011. ACM.

[24] Yixue Wang and Emőke-Ágnes HorvÁt. Gender differences in the global music industry: Evidence from musicbrainz and the echo nest. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13 of *ICWSM'19*, pages 517–526. AAAI, 2019.

[25] Christine T. Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R. Brubaker. The changing contours of "participation" in data-driven, algorithmic ecosystems: Challenges, tactics, and an agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 377–384, New York, NY, USA, 2018. ACM.

[26] Wen Wu, Li Chen, and Yu Zhao. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction*, 28(3):237–276, August 2018.

[27] Patrizia Zanoni, Maddy Janssens, Yvonne Benschop, and Stella Nkomo. Guest editorial: Unpacking diversity, grasping inequality: Rethinking difference through critical perspectives. *Organization*, 17(1):9–29, 2010.

TOWARDS UNCOVERING DATASET BIASES: INVESTIGATING RECORD LABEL DIVERSITY IN MUSIC PLAYLISTS

Peter Knees
Faculty of Informatics
TU Wien
Vienna, Austria
peter.knees@tuwien.ac.at

Moritz Hübler
Faculty of Informatics
TU Wien
Vienna, Austria
e1426077@student.tuwien.ac.at

ABSTRACT

Music recommender models are predominantly built upon the assumption that historic listening data is the result of the interaction between a user and an item and that overall interaction patterns can be extrapolated for making future recommendations (collaborative filtering). In this paper, we argue that listening logs are not only the result of users interacting with items but users interacting with items through a listening service. As such, the service has an impact on the recommendations made and the data created, consequently also introducing biases to datasets used for model training and evaluation.

We investigate the case of a large dataset of Spotify playlists. In order to uncover patterns in the data, we augment the dataset with record label information crawled from the web. Subsequent first analyses of record label diversity within the playlists reveal unequal distributions and higher consistency of the most popular label especially in short playlists with few albums. We discuss possible reasons causing these patterns as well as potential algorithmic biases of the approach taken.

1. MOTIVATION

With the establishment of commercial online music streaming services, with virtually unlimited access to music, music recommendation has become a commodity in music listening and discovery. Models for recommendation are frequently based on collaborative filtering on historic listening data and/or manually generated playlists, cf. [14, 16]. An underlying assumption of these models, specifically matrix factorization models, is that the observed listening events are the result of the interaction between users and items. This assumption neglects, however, the role that the listening platform, i.e. the service delivering the music to the listeners, plays a crucial role in this process. In fact, music recommender systems operate in multistakeholder environments, serving the interests not only of the customers, but also the producers and the system itself, among others, e.g., by maximizing revenue, cf. [1, 2, 6]. This setting impacts therefore not only the recommendations made, but also music recommender systems research, as the data generated might exhibit biases of various nature (popularity bias, selection bias, etc.), affecting model selection and evaluation and potentially leading to ethical issues in MIR and recommender research, cf. [5].

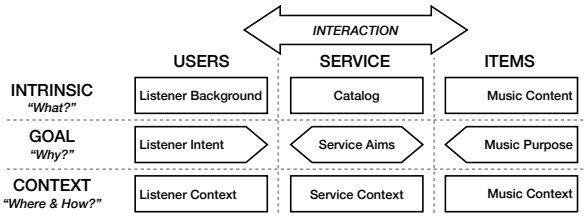


Figure 1. The service, i.e. the system itself, as a factor in recommender systems and influences the interaction between users and items, therefore the data logged. Schematic extended from the user-item interaction factor model in [9].

In light of these considerations, in the following, we aim at exploring possible effects of the multistakeholder setting on data used for music recommender systems research by considering record label information. After discussing the service as a relevant factor in more detail in Sec. 2, in Sec. 3 we present an investigation of record label diversity in a large dataset of playlists and report on initial findings. We conclude the paper with a discussion on possible reasons, shortcomings, and algorithmic biases of the presented study in Sec. 4 and future work in Sec. 5.

2. THE SERVICE AS A FACTOR

As depicted in Fig. 1, the service (i.e. the system, in the terminology of the multistakeholder framework [1]) acts as intermediary between users and items, and, just like these, has different characteristic dimensions. *Intrinsic* properties comprise the catalog, i.e. which content is provided and can be recommended. As an example, consider Soundcloud, which intrinsically recommends different music content than, e.g., Spotify. *Goal* refers to the purpose and aim of the service (cf. the utility of the sys-

tem in the multistakeholder framework): Why is this service in place? What is the served market niche? What are the identified use cases, such as a focus on music discovery? Furthermore, is delivery of specific content, such as content owned by the service a priority, as is increasingly the case in the movie streaming domain (cf. Netflix, Amazon)? This information, ultimately, should facilitate answering the central question why recommendations are (made) the way they are, and fostering transparency of music recommender systems, cf. [7]. Finally, the *Context* refers to the circumstances under which a service is operating and how this affects, e.g., the available catalog through licensing restrictions in particular countries. Also the question of whether the service itself performs context-aware recommendations, e.g., by serving different content based on the used platform (app vs. desktop version), becomes relevant under this property and affects the data created and the recommendations made in consequence.

3. PLAYLIST DATASET INVESTIGATION

When analyzing patterns in listening and playlist data, common aspects considered are artist, album, genre, emotion, and listening context of the contained tracks [14]. Recent work by Porcaro and Gómez [13] investigates playlist diversity wrt. popularity and social tag categories in different datasets.

A so far largely neglected aspect of music listening data analysis in music information retrieval and recommendation research is information on the publisher, i.e. the record label, which however is a key stakeholder in music recommendation. A record label or record company is defined as a company which either published an album itself or is distributing albums of other record labels (distributor). Usually, an album is published by one or more record labels and each record label may have multiple distributors. For instance, the album "Stoney" of rapper Post Malone was published by the record label Republic, which is distributed by Universal Music Group, Virgin EMI Records, Island Records and Universal Music Enterprises. The album "Coloring Book" by Chance the Rapper, on the other hand, has neither a record label nor a distributor and is therefore considered as being released independently. As record labels of interest, we focus on the "big three" major global players of the record industry, namely *Universal Music Group*, *Sony Music*, and *Warner Music Group*, as well as independent releases.

3.1 Dataset

For our initial study, we investigated the Million Playlist Dataset (MPD), a dataset of one million hand-crafted playlists released by Spotify for the purpose of the RecSys Challenge'18 [3]. [1] As described in the documentation of the dataset, included playlists were selected based on several requirements, including that each playlist creator is from the US and older than 13 years; each playlist was public when the MPD was generated; each playlist was created between the years 2010 and 2017; each playlist has at least one follower, who is not the creator; each playlist contains between 5 and 250 tracks; and—most interesting wrt. to the findings presented—each playlist contains at least 3 unique artists and 2 unique albums. The million playlists contain a total of 66.346.428 track entries (2.262.292 unique tracks from 825.245 unique albums by 295.860 unique artists).

No information on whether a recommender system was supporting the manual creation of the playlists, e.g., by providing suggestions for continuation, is given. As no record label information was provided with the dataset, we attempted to automatically augment the dataset with label information for each contained album by developing a web crawler as described next.

3.2 Crawling Record Labels

To gain additional data on music, web information extraction has shown to be a useful approach, cf. [8, 12]. For looking up label information for an album, first, the title of the album (and the additional constraint *(album)*, optionally in combination with the artist name) is sent to Google, and the first returned result from Wikipedia is searched for label information in the free text as well as in the Wikipedia info box. For the free text, plain occurrences of the keywords *sony*, *universal*, *warner*, and *independent* in the free text are counted. For the info box, the section "Record labels" is searched for links to entries of record labels. Info boxes of pages of record labels are checked for entries in the section "Parent company", recursively pointing to other record labels until no further parent companies can be identified. All pages of record labels appearing in this hierarchy are checked for the occurrence of the above keywords, their frequencies weighted by a factor decreasing exponentially with the number of hops from the album page. Finally, overall weighted frequencies per keyword are summed up and the album assigned to the label with the most frequent occurrence. If no keyword occurrence is significantly high (after thresholding), the album is assigned to the label class *unknown*.

Due to time constraints, lookup was restricted to the 147,883 most frequent albums (17.92% of the 825,245 unique albums), of which 117,879 albums (14.28% of the overall contained albums) could be classified. Because of the skewed distribution of albums, the classified 14.28% of albums already cover 92.1% of the occurring tracks of the MPD. When adding looked up but unclassified albums, the coverage increases to 96.59%.

3.3 Playlist Diversity

To quantify the diversity of a single playlists of the dataset wrt. record label, the Simpson index is introduced [15]. The Simpson index measures the probability that two items taken from a set belong to the same class—here, that two tracks of a playlist belong to the same record label. It is calculated as $\lambda = \sum_{i=1}^{R} p_i^2$, where R is the number of label classes, i.e. 5, and p_i the probability of each record

[1] Note that data enhancement and experiments were carried out in preparation of a submission to the creative track.

label to be drawn from the set, i.e. the number of tracks belonging to a label divided by the total number of tracks of the playlist. A low λ, therefore, stands for a high diversity while a high λ indicates a low diversity. Fig. 2 shows the distribution of λ over all playlists in the set.

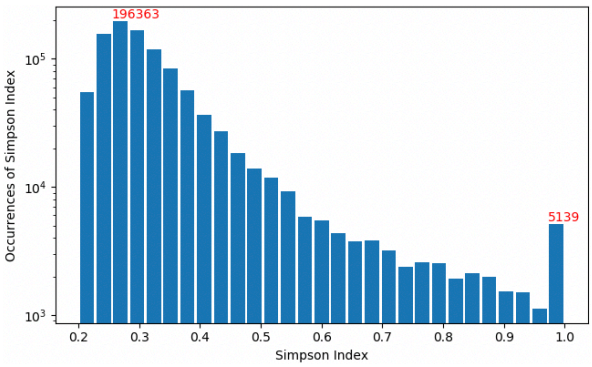


Figure 2. Distribution of Simpson index over all playlists.

3.4 Findings

Due to the high popularity of few tracks and albums in the dataset, the distribution of albums is highly skewed. Fig. 3 orders albums based on frequency in the dataset and shows the label distribution as a histogram where each bucket contains the same amount of album occurrences. Note that bucket 1 (most frequent) comprises only 33 unique albums, whereas bucket 20 comprises 718,836 unique albums.

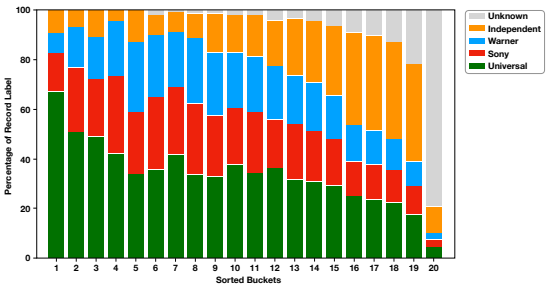


Figure 3. Distribution of labels sorted by album popularity (equal number of occurrences in dataset per histogram bucket).

It can be seen that Universal is publishing most of the tracks of the MPD (see also Fig. 4) and predominates a vast majority of popular albums. It can also be seen that rare and unpopular albums are more likely to be independent releases or to have an unknown label (the latter also caused by the restricted lookup).

The dominance of Universal can be seen more clearly in Fig. 4, where the overall occurrence of labels among all tracks is compared to the occurrence in highly homogeneous playlists. While Universal accounts for 34% of the tracks in the overall dataset (followed by Sony with 20.8%, Independent with 18.8%, and Warner with 18.4%), for playlists with low label diversity (high λ), the share of Universal increases substantially (likewise other known labels

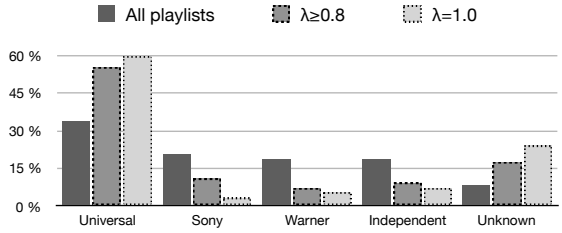


Figure 4. Focus on percentage of labels in playlists (all playlists *vs* playlists with Simpson index $\lambda \geq 0.8$ and $\lambda = 1.0$, resp.).

decrease). Analysis of playlists with $\lambda \geq 0.8$ shows that they contain on average a smaller number of tracks (45.5) and unique albums (12.0) than the global average (66.3 and 48.6, resp.). For playlists with $\lambda = 1.0$, i.e. playlists containing tracks from only one label, the average number of tracks is 29.3, the average number of unique albums 6.8.

4. DISCUSSION

Understanding the role of the recommendation service (cf. Sec. 2) and looking into the record label as an additional source for playlist and listening dataset analysis seem to be worthwhile efforts. While this information might not be of primary interest to most consumers, it can help in revealing otherwise non-obvious patterns in the data.

From the collected information on the record labels it can be concluded that (a) the original choice of investigating the "big three" record labels was justified as 73.2% of all tracks could be assigned to one of them, and (b) that labels are not distributed evenly in the dataset, with Universal Music Group having a predominant position. A possible explanation is that popularity is the dominating effect leading to this behavior. This can have an impact regardless of whether playlists are chosen purely manually or with the help of a data-driven recommender system.

Whether the dominance of Universal in Spotify data is leading to feedback loops or has other, possibly strategic, reasons remains subject to speculation. It needs to be kept in mind that the MPD might be a non-representative dataset, and that observed effects could relate to filtering performed specifically for the RecSys Challenge'18. Generally, Spotify seems to put a focus on its role as a tool for discovery, therefore aiming at exploring user preferences by delivering diverse recommendations [10] and aiming at providing fairness for suppliers [11].

When interpreting our findings on the MPD, one has to be aware that the methodological process chosen itself can introduce errors and biases. Such algorithmic biases, cf. [4], could be connected to the assignment of a record label. Counting keyword occurrences is a simple heuristic without check of plausibility or assessment of context. Especially when using keywords such as *universal* and *independent* higher occurrences can be expected than for proper names like *sony* and *warner*. For the task at hand, a connection can likely be modelled in this fashion, however, correctness of the output is not guaranteed.

5. FUTURE WORK

Our initial findings require further analysis and deeper investigation of the identified effects. An important step towards this is to reconsider our design choices and data extraction methods. Instead of parsing Wikipedia, we could resort to better structured sources of editorial metadata, such as MusicBrainz, Discogs, allmusic, or Bandcamp. Deeper investigation also comprises inspection of other music recommendation datasets and comparison of findings. Another logical step will be to complement our investigations wrt. record label with other dimensions of diversity (style, genre, tags, etc. [13] to identify potential effects within sub-communities and use cases.

With regard to the consequences of possible biases in the data, future work will consist in measuring the effect of modeling label information within recommendation approaches on prediction accuracy. For the MPD, this could be simulated upon release of the dataset for research beyond the RecSys Challenge'18. Also the identification of other metrics that measure the impact of these effects, potentially in relation to societal values could be addressed.

Another direction for future research could be to extend our approach to other domains and investigate, e.g. movie recommendation datasets based on movie studios.

6. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback and valuable input for future work.

7. REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. arXiv:1905.01986, 2019.

[2] Himan Abdollahpouri and Steve Essinger. Multiple stakeholders in music recommender systems. arXiv:1708.00120, 2017.

[3] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, 2018.

[4] Arthur Flexer, Monika Dörfler, Jan Schlüter, and Thomas Grill. Hubness as a case of technical algorithmic bias in music recommendation. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2018.

[5] Andre Holzapfel, Bob L. Sturm, and Mark Coeckelbergh. Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, 1(1):44–55, 2018.

[6] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 2016.

[7] Peter Knees. A Proposal for a Neutral Music Recommender System. In *Proceedings of the 1st Workshop on Designing Human-Centric MIR Systems*, 2019.

[8] Peter Knees and Markus Schedl. Towards Semantic Music Information Extraction from the Web Using Rule Patterns and Supervised Learning. In *Proceedings of the 2nd Workshop on Music Recommendation and Discovery (WOMRAD)*, 2011.

[9] Peter Knees, Markus Schedl, Bruce Ferwerda, and Audrey Laplante. User awareness in music recommender systems. In M. Augstein, E. Herder, and W. Wörndl, editors, *Personalized Human-Computer Interaction*, pages 223–252. De Gruyter, 2019.

[10] Malte Ludewig and Dietmar Jannach. User-centric evaluation of session-based recommendations for an automated radio station. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*, 2019.

[11] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, 2018.

[12] Sergio Oramas, Luis Espinosa-Anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. Information extraction for knowledge base construction in the music domain. *Data & Knowledge Engineering*, 106:70 – 83, 2016.

[13] Lorenzo Porcaro and Emilia Gómez. 20 years of playlists: A statistical analysis on popularity and diversity. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[14] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. Music recommender systems. In F. Ricci, L. Rokach, and B. Shapira, editors, *Recommender Systems Handbook*, pages 453–492. Springer, 2nd ed., 2015.

[15] Edward H. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.

[16] Hamed Zamani, Markus Schedl, Paul Lamere, and Ching-Wei Chen. An analysis of approaches taken in the ACM RecSys Challenge 2018 for automatic music playlist continuation. *ACM Transactions on Intelligent Systems and Technology*, 10(5):57:1–57:21, Sept 2019.

HUMAN SUBTRACTED: SOCIAL DISTORTION OF MUSIC TECHNOLOGY

Finn Upham
McGill University
finn.upham@mail.mcgill.ca

ABSTRACT

The social functions of music have been broken by successive music technology advances, bringing us to the current "boundless surfeit of music" (Schoenberg) navigated with only the faintest traces of common interests retained in personalised music recommendation systems. This paper recounts the desocialisation of music through sound recording, private listening, and automated recommendation, and considers the consequences of music's persistent cultural and interpersonal power through this changing use.

1. WHAT MUSIC HAS BEEN

Humanity developed and developed with musical behaviours when these sounds had to come from people in physical and social proximity. Today, much if not most of our musical experiences involves listening alone to sounds constructed in the past by people we will never meet [12], sounds often chosen for us according to inferred individual preference. Consumer behaviour demonstrates that this change is easy to adopt, but convenience does not guaranty the shifts are benign. Despite the impacts of technology, cross-cultural studies of modern musical practices show that music continues to carry social weight in a number of ways [18]. From a few claims about music before recording technology, we can contextualise their impacts on our current listening cultures. For most of our species history, the following held true:

1. **Proximity to source**: Heard music is made by nearby humans, people known to the hearer either personally or by a role justifying their physical proximity.

2. **Open broadcast signal**: This music is also heard by everyone else within earshot.

3. **Effortful sound**: Music is present when it is worth the physical effort of producing it, whether for lullabies, group entertainment, solitary distraction, intimidation, etc.

4. **Cultural affinity**: Most music heard is by members of the hearer's culture and it expresses that shared identity with familiar sound and structures.

5. **Social interpretability**: The hearer easily interprets the performers' purpose from their sounds: to calm, play, mourn, etc.

6. **Group distinction**: Music that sounds different and is hard to interpret is by people from a different group or culture.

Constrained by acoustics and mobility, music has predominantly been an insular social practice. Sound recording and reproduction technology broken the requirement of physical proximity between music producers and listeners, personal playback devices divided listeners from each other, and personalised music recommendation is loosening the last cultural/social constraints on musical exposure in pursuit of preference within a narrow range of uses.

2. SEPARATING MUSICIANS FROM LISTENERS

Separating sound and source has not removed the social and cultural relationships previously associated with musical contact. Instead the identities of musicians are amplified and opened up to interpretation without practical constraints like physical proximity and voluntary interaction. Only a speaker away, they never refuse to "Play it again!"

Creators of favourite and famous music engage our attention and care because we are free to "know" them through a medium that articulates cultural belonging, mutual trust, and intentional engagement. Repeated exposure to specific tracks extends familiarity from baseline social interpretability to the intimacy of co-performers. This sensitivity permits dedicated listeners to hear recording artists as friends, peers, family, developing deep parasocial attachments. When a performer is perceived to contradict the image their followers have inferred, whether on grounds of musical skill [2] or social failings, betrayal of these unidirectional bonds challenge listeners appreciation of their works. Social factors define the value of recorded music.

Ease of distribution has exposed listeners to a greater diversity of music styles, crossing the boundaries of time, geography, and socio-economic stratification. While music can promote cross-cultural understanding and respect, exposure to new genres and artists exclusively via recordings may be having undesirable consequences. Humans

infer rules of musics quickly by ear, and repeated exposure to recordings shift what is heard as "foreign" to sounds listeners claim as part of their own cultural practice. This intuitive appropriation encourages audiences to feel entitled to the cultural work and products of other communities without acknowledging all the differences (and disparities) between them. A genre enthusiast can feel vague empathy and affiliation for those making the music and yet never confront conflicting bigotries such as racism [15]. Beyond the problem of enjoying music without respecting the musicians, this pattern of appropriation has financial costs to the communities from whom musical styles and works have been stolen. Musical genres and works originating from Black musicians in North America have repeatedly been taken up by White musicians who go on to have hugely impactful and profitable recording careers [9].

3. LISTENING ALONE

Solitary listening has become common practice, for some the most common context for music listening [12], contrary to historical acoustic conditions and presumed uses for group bonding and coordination. Listening to music over headphones is a convenient way to isolate a listener from their environment [11], to find entertainment without bothering others. Besides discouraging interpersonal interactions, tailoring music to one person's interests has facilitated substantial changes in musical use.

Musical subcultures within larger communities are not new phenomena, but technology and solitary listening practices has shifted the membership from the people gathered for live events to individuals picking up associations independent of their predominant cultural environment. Intergenerational conflict over musical taste is cliche, but the contrast in preference is exaggerated by uneven exposure to new genres. Family and neighbours can grow deep cultural investment in musical styles without allowing each other to develop even superficial understanding through passive exposure.

Self-actualisation is a notable aspects of teenage music consumption choices [17], using this medium to articulate personality and identity against the norms of their immediate social environment. Like other cultural signifiers, genres carry stereotypes about their listeners [14], and peer opinions seem to have more weight in determining listening preferences for students than many structural qualities [8]. And yet, private listening also allows people access to music they'd rather not admit enjoying, for fear of being judged by association with the musicians or culture [5]. When music is a mechanism for defining ourselves as well as our community, the implications of association become personal.

Many of today's recorded music consumers report selecting tracks strategically, to change their mood or explore and resolve feelings [10]. Independent music consumption allows individuals to focus music's power to move a crowd on themselves, using dance music to stay awake or be inspired by a favourite love song, free of having to consider what might be overheard or the musicians own goals in performing. For some, music has become a tool to optimise their behaviour and feelings, compensating for undesirable emotional challenges [1]. Although physically removed from musicians and other listeners, music can still carry the feeling of company. Like other forms of socially-loaded media, music is often used as a social surrogate to sooth loneliness, reminding people of community, identity, and past interpersonal connections [16].

4. PASSIVE EXPOSURE

It's always been common for music we hear to be chosen by others. Most musical sounds heard would have been culturally familiar and socially situated in who was making them and why. Some circumstance would oblige engagement, but if the music was not in someway intended for the hearer, they would be free to ignore the musicians' efforts. While the range of relationships to music overheard is much the same, technology has changed the reasons for it to be in our environment, including who is responsible for music's presence.

With recorded tracks came new cultural contexts for the introduction of new music. DJs with cultural authority program offering with information about the pieces or artist while expressing personal assurances of the music's quality. Videos present music with extra-auditory narratives. And friend or expert mixtapes became playlists shared online with social value informing consumers' relationship to the sound before hearing it [6].

Automated music recommendation systems cut away this last layer of social context from new works. Tracks are offered mysteriously, anonymously, presenting the illusion of understanding through personalization without a story as to why the music is to be heard here and now. And without interpersonal pressure to pay attention and use one's reactions, these sounds are easy to ignore. Some tracks may catch our attention with novelty, but many will be overlooked as comfortably interpretable but not special unless a social connection gives it worth. Instead of investing in musical works to forge lasting affective meaning, services like Spotify Discover parade an array of new pieces to be heard without context, hassle free. As one user reported: "... Spotify has changed the way I listen to music. When previously I would stick to the music I had always listened to due to the high level of work required to source new music that I like, I now enjoy large varieties of music and get bored quickly of the same music over and over." [4]

Casual listening to fresh material is fine for some purposes, but it is not an efficient path experiencing the powerful emotions many consumers look for in music [13]. Comparisons of self-selected vs expert-selected music consistently shows that music people choose themselves have stronger impacts on how they feel [19]. Personalised recommendation may try to give listeners music that fits their cultural affiliations and general mood, but without social emphasis, it may be hard for consumers to build the associations so useful for triggering stronger feelings.

If novelty is easier to serve than emotional impacts or

familiarity, users of music recommendation systems are at risk of all the concerns raised so far. These services encourage cultural wandering, helping users to appropriate genres without understanding the originating peoples and cultures. They discourage the attentional investments needed to develop strong emotional ties. And by tempting music consumers with personalised offerings, users are further pressured to allow their proximal social networks to decay.

5. CONCLUSIONS

Sound recording and subsequent technologies have utterly transformed how we use music today and yet the acoustic, social, and cultural constraints of music practices past still define its impacts on music consumers. When commercial recording was just starting, many musicians of the day were concerned by the disruptions they anticipated. In our present effortless consumption of recorded music selected to suit to our personal pallet, we have reached the dreaded "domestication of sound" (Debussy) that allows us to "listen lazily" (Stravinsky) and loose "our powers of musical concentration" (Keller) in this "boundless surfeit of music" (Schoenberg) [3] [7, p. 45]. But the greatest loss in experience may be through the de-socialisation of music, as we overlook where it comes from and what it means when that information is stripped away by the dominant means of dissemination.

6. REFERENCES

[1] Margarida Baltazar and Suvi Saarikallio. Strategies and mechanisms in musical affect self-regulation: A new model. *Musicae Scientiae*, 23(2):177–195, 2019.

[2] William Cheng. So you've been musically shamed. *Journal of Popular Music Studies*, 30(3):63–98, 2018.

[3] Eric F Clarke. The impact of recording on listening. *twentieth-century music*, 4(1):47, 2007.

[4] Sally Jo Cunningham, David Bainbridge, and Annette Bainbridge. Exploring personal music collection behavior. In *International Conference on Asian Digital Libraries*, pages 295–306. Springer, 2017.

[5] Sally Jo Cunningham, David Bainbridge, and Annette Falconer. "more of an art than a science": Supporting the creation of playlists and mixes. In *7th International Society for Music Information Retrieval Conference*. University of Victoria, 2006.

[6] Sally Jo Cunningham, Nina Reeves, and Matthew Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 5–16. IEEE Computer Society, 2003.

[7] Evan Eisenberg. *The recording angel: music, records and culture from Aristotle to Zappa*. Yale University Press, 2005.

[8] Leif Finnäs. How can musical preferences be modified? a research review. *Bulletin of the Council for Research in Music Education*, (102):1–58, 1989.

[9] Kevin J Greene. Copyright, culture & (and) black music: A legacy of unequal protection. *Hastings Comm. & Ent. LJ*, 21:339, 1998.

[10] Patrik N Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

[11] Miracle J.Y.J. Lim and PerMagnus Lindborg. How much does earphone quality matter while listening to music on buses and trains? In Geoff Luck and Olivier Brabant, editors, *Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland*, 2013.

[12] Adrian C North, David J Hargreaves, and Jon J Hargreaves. Uses of music in everyday life. *Music Perception: An Interdisciplinary Journal*, 22(1):41–77, 2004.

[13] William M Randall, Nikki S Rickard, and Dianne A Vella-Brodrick. Emotional outcomes of regulation strategies used during personal music listening: A mobile experience sampling study. *Musicae Scientiae*, 18(3):275–291, 2014.

[14] Peter J Rentfrow, Jennifer A McDonald, and Julian A Oldmeadow. You are what you listen to: Young people's stereotypes about music fans. *Group Processes & Intergroup Relations*, 12(3):329–344, 2009.

[15] Jason Rodriquez. Color-blind ideology and the cultural appropriation of hip-hop. *Journal of Contemporary Ethnography*, 35(6):645–668, 2006.

[16] Katharina Schäfer and Tuomas Eerola. How listening to music and engagement with other media provide a sense of belonging: An exploratory study of social surrogacy. *Psychology of Music*, page 0305735618795036, 2018.

[17] Mark Tarrant, Adrian C North, and David J Hargreaves. English and american adolescents' reasons for listening to music. *Psychology of Music*, 28(2):166–173, 2000.

[18] Sandra E. Trehub, Judith Becker, and Iain Morley. Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140096, 2015.

[19] Juliane Völker. Personalising music for more effective mood induction: Exploring activation, underlying mechanisms, emotional intelligence, and motives in mood regulation. *Musicae Scientiae*, 2019.

THE RELATIONSHIP BETWEEN THE CONSISTENCY OF USERS' RATINGS AND RECOMMENDATION CALIBRATION

Masoud Mansoury[1] **Himan Abdollahpouri**[2] **Joris Rombouts**[1] **Mykola Pechenizkiy**[1]

[1] Eindhoven University of Technology, the Netherlands
[2] University of Colorado Boulder, USA

m.mansoury@tue.nl, himan.abdollahpouri@colorado.edu, j.c.rombouts@student.tue.nl, m.pechenizkiy@tue.nl

ABSTRACT

Fairness in recommender systems has recently received attention from researchers. Unfair recommendations have negative impact on the effectiveness of recommender systems as it may degrade users' satisfaction, loyalty, and at worst, it can lead to or perpetuate undesirable social dynamics. One of the factors that may impact fairness is *calibration*, the degree to which users' preferences on various item categories are reflected in the recommendations they receive.

The ability of a recommendation algorithm for generating effective recommendations may depend on the meaningfulness of the input data and the amount of information available in users' profile. In this paper, we aim to explore the relationship between the consistency of users' ratings behavior and the degree of calibrated recommendations they receive. We conduct our analysis on different groups of users based on the consistency of their ratings. Our experimental results on a movie dataset and several recommendation algorithms show that there is a positive correlation between the consistency of users' ratings behavior and the degree of calibration in their recommendations, meaning that user groups with higher inconsistency in their ratings receive less calibrated recommendations.

1. INTRODUCTION

Recommender systems are powerful tools for predicting users' preferences and generating personalized recommendations. These systems, while effective, often suffer from lack of fairness in recommendation results, meaning that the outputs of recommendation algorithms are, in some cases, biased against some protected groups [4]. As a result, this discrimination among users will negatively affect users' satisfaction, loyalty, and overall effectiveness of the system.

Unfair recommendation is often defined as the situation that a recommendation algorithm behaves differently when generating recommendations for different groups of users (i.e., protected and unprotected groups). As an example, when users who belong to the unprotected group receive more accurate recommendations than the users in the protected group, we say there is discrimination against the protected group. This unfair behavior can originate from either the underlying biases in the input data used for training [1,3,11] or the result of recommendation algorithms [12].

Abdollahpouri et al. in [2] showed that popularity bias has a negative impact on the fairness of recommendation outputs. In that work, authors showed that the recommendations generated for the majority of users are concentrated on popular items even for those who are interested in long-tail and non-popular items. A more similar analysis to our work is done in [1] where authors showed how popularity bias is correlated with the miscalibration of the recommendations and how different user groups with varying degree of interest in popular items experience different levels of miscalibration.

In this paper, we aim to do more exploration on the possible reasons for discrimination in recommendation results. Our hypothesis is that the richness of a user's profile might have impact on how the algorithm performs for that user. To explore this, we analyze users' profile and investigate the relationship between the consistency of users' ratings and the degree of calibrated recommendations. We believe that the lack of consistency in user's profile can be one possible reason for miscalibrated recommendations as recommender system is unable to correctly predict user's preferences. We discuss the approach for measuring profile consistency in next section.

2. PROFILE CONSISTENCY

We define a rating to be consistent if it is in agreement with the ratings given by other users. For instance, if a user has given 5 to an item with the average rating of 2, it means his rating has an inconsistency of degree 3. Profile consistency refers to the fact that how similar a user rates an item compared to the majority of other users who have rated that item. This has been referred to the gray sheep problem in the literature [5]. Since collaborative filtering approaches use opinions of other users (e.g., similar users) for generating recommendations for a target user, it is highly possible that inconsistent profiles do not receive effective recommendations. Given a target user, u, and I_u as all items rated by u, inconsistency of u can be calculated as:

$$inconsistency_u = \frac{\sum_{i \in I_u} |r_{u,i} - \overline{r_i}|}{N_u} \quad (1)$$

Table 1: Accuracy of recommendation algorithms

algorithm	`UserkNN`	`ItemkNN`	`SVD++`	`ListRankMF`
precision@10	0.214	0.223	0.122	0.148

where $r_{u,i}$ is the rating provided by u on item i, $\overline{r_i}$ is average of ratings assigned to item i, and N_u is the number of items rated by u.

3. CALIBRATION MEASURE

Measuring fairness of recommendation results is a complex task. Several metrics have been recently proposed for measuring the equity of recommendation results [3,11,12]. Bias disparity [9, 11] is one of those metrics that measures how much an individual's recommendation list deviates from his or her original preferences in the training set across an item's category. The issue with bias disparity is that it calculates the bias value for a group of users on a specific item category and does not return the overall bias value for a group of users across all item categories.

Calibration of recommendations is another factor that affects fairness in recommender systems [10]. Calibration measures the distance between users' preferences in training data and the predicted preferences in recommendation lists. Distance equals to zero indicates perfect calibration, while distance larger than zero indicates a degree of *miscalibration*. For the rest of the paper, we use the term miscalibration to refer to this distance value.

Original preferences in train set and predicted preferences in recommendation lists are represented as distributions across item categories and the distance between these two distributions shows the degree of miscalibration. The main incentive behind having calibrated recommendation is the fact that recommendation lists should appropriately represent users' profile/interest in train data. Assume a user's profile consists of 70% action movies and 30% adventure movies. Then, it is expected that the recommendation list for this user also contains the same proportion of each genre.

For calculating the miscalibration, we follow the equations introduced in [10]. Given the distribution of items' category in user u's profile as p and the distribution of items' category in recommendation list generated for user u as q, we use Kullback-Leibler divergence measure for calculating the distance between these two distributions for user u as follow:

$$KL_u(p|\widetilde{q}) = \sum_{c \in C} p_c log \frac{p_c}{\widetilde{q}_c} \tag{2}$$

where C is item categories (e.g., genres in movie recommendations) and $\widetilde{q}$ is approximately similar to q calculated as:

$$\widetilde{q}_c = (1-\alpha).q_c + \alpha.p_c \tag{3}$$

The purpose of $\widetilde{q}$ is to overcome the issue of zero values for some categories in q. Small value for $\alpha > 0$ guarantees $\widetilde{q} \approx q$. In our experiments, we use $\alpha = 0.01$ as suggested in [10].

4. EXPERIMENTS

For experiments, we use MovieLens 1M (ML1M) dataset which is a movie rating data collected by the MovieLens [1] research group. In this dataset, 6,040 users provided 1,000,209 ratings on 3,706 movies. The ratings are in the range of 1-5 and the density of the dataset is 4.468%. Also, each movie is assigned several genres. Overall, there are 18 genres in this dataset.

For performing experiments, we divided the dataset into train and test sets as 80% and 20%, respectively. The train set is used for building the model, and in the test condition, we generate recommendation lists of size 10 for each user.

After recommendation generation, for each user, we calculate a value for inconsistency of profile and a value for miscalibration. We measure inconsistency of profile using equation 1 and miscalibration of recommendations generated for a user using equation 2. For the purpose of presentation, we sort users based on their profile inconsistency and then group them into several groups with the same range. Finally, for each group we calculate the average of profile inconsistency and miscalibration.

Our experiments includes user-based collaborative filtering (`UserKNN`), item-based collaborative filtering (`ItemKNN`), singular value decomposition (`SVD++`), and list-wise matrix factorization (`ListRankMF`). All recommendation models are optimised using Grid Search over hyperparameters and best results in terms of precision are reported here. Table 1 shows the accuracy of those recommendation algorithms. We used *librec-auto* and LibRec 2.0 for all experiments [6, 8].

4.1 Experimental results

Figure 1 shows the relationship between inconsistency in users' profiles and the miscalibration of the recommendations for each group. For all recommendation algorithms, there is a positive correlation between inconsistency of the ratings in the profile and miscalibration: as inconsistency increases, miscalibration will also increase. Except for `SVD++`, there is a strong correlation for all other recommendation algorithms.

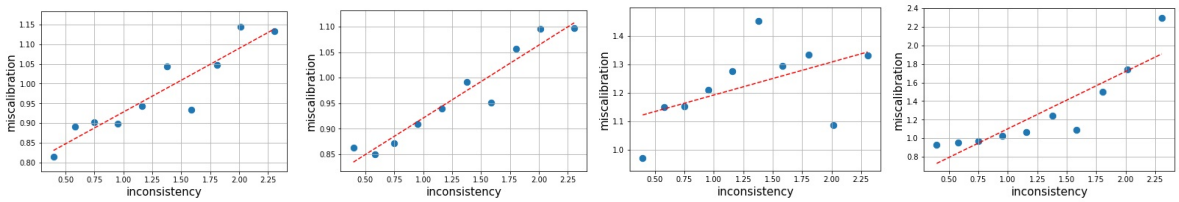
The correlation coefficient for `UserKNN` is 93%, for `ItemKNN` is 96%, for `SVD++` is 53%, and for `ListRankMF` is 88% which are indicative of strong correlation between inconsistency of profile and miscalibration, except for `SVD++`.

These are interesting results as they show that users who provide inconsistent ratings will less likely receive calibrated recommendations. This can increase unfair situation such that different users will receive different level of calibration in their recommendation lists. Therefore, taking into account the inconsistency of users' profile when generating recommendations can alleviate unfairness of recommendation outputs.

5. DISCUSSION

Although in this paper we considered consistency of users' profile as a factor that has positive impact on the effectiveness of recommender systems, there might be other factors that also contribute to the effectiveness of these systems.

[1] https://grouplens.org/datasets/movielens/



(a) `UserKNN` (b) `ItemKNN` (c) `SVD++` (d) `ListRankMF`

Figure 1: Relationship between inconsistency of users' profile and miscalibration of recommendations generated for those users.

Profile size can be one of the factors for generating successful recommendations and may affect the performance of recommender systems. Users with low profile size or insufficient number of ratings are often known as *cold-start users*. It has been long noted that these profiles are the source of concern for recommender systems as recommendation algorithms are unable to accurately predict their preferences [7].

Information gain (i.e. entropy) is one form of measuring informativeness of a profile and another factor that may affect the performance of recommender systems. A Profile with high entropy is the one where the user has provided ratings to a wide range of items from least preferred to most preferred ones. These profiles are informative because they provide both positive and negative feedback and recommender system will better learn to what recommend and what not recommend. We will consider aforementioned metrics (or combination of those metrics) for measuring informativeness or richness of a profile as our future work.

Our experiments in this paper are performed on a user-item rating data in movie domain. However, it can be extended to other datasets from different domains. In particular, as a future work, we intend to extend this work to music recommendation. We are interested in investigating whether inconsistency of a user's profile has any connection with the fact that some users have a niche taste and they might rate some popular songs differently from the the majority of other users.

6. CONCLUSION

In this paper, we explored the relationship between the consistency of users' profile and calibration of recommendations. Our experimental results showed that recommendation algorithms generate more calibrated recommendations for consistent profiles. As a future work, we aim to further explore the relationship between profile richness and recommendation calibration by taking into account other metrics like profile size and entropy.

7. REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755*, 2019.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *In Workshop on Recommendation in Multistakeholder Environments (RMSE)*, 2019.

[3] Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. Balanced neighborhoods for fairness-aware collaborative recommendation. In *RecSys workshop on Fairness, Accountability and Transparency in Recommender Systems*, 2017.

[4] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *In Conference on Fairness, Accountability and Transparency*, pages 172–186, 2018.

[5] Mustansar Ali Ghazanfar and Adam Prügel-Bennett. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, 41(7):3261–3275, 2014.

[6] Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. Librec: A java library for recommender systems. In *UMAP Workshops*, 2015.

[7] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073, 2014.

[8] Masoud Mansoury, Robin Burke, Aldo Ordonez-Gauger, and Xavier Sepulveda. Automating recommender systems experimentation with librec-auto. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 500–501. ACM, 2018.

[9] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. In *RecSys workshop on Recommendation in Multi-Stakeholder Environments*, 2019.

[10] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 154–162. ACM, 2018.

[11] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. Bias disparity in recommendation systems. *CoRR*, abs/1811.01461, 2018.

[12] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *In Advances in Neural Information Processing Systems*, pages 2921–2930, 2017.