https://doi.org/10.1007/s41109-022-00530-7

Applied Network Science

RESEARCH **Open Access**

Measuring the effect of collaborative filtering on the diversity of users' attention

Augustin Godinot* and Fabien Tarissan

*Correspondence: augustin.godinot@ens-paris-saclay.fr

Ecole normale supérieure Paris-Saclay, Gif-Sur-Yvette, France

Abstract

While the ever-increasing emergence of online services has led to a growing interest in the development of recommender systems, the algorithms underpinning such systems have begun to be criticized for their role in limiting the variety of content exposed to users. In this context, the notion of *diversity* has been proposed as a way of mitigating the side effects resulting from the specialization of recommender systems. In this paper, using a well-known recommender system that makes use of collaborative filtering in the context of musical content, we analyze the diversity of recommendations generated through the lens of the recently proposed *information network diversity measure*. The results of our study offer significant insights into the effect of algorithmic recommendations. On the one hand, we show that the musical selections of a large proportion of users are diversified as a result of the recommendations. On the other hand, however, such improvements do not benefit all users. They are in fact mainly restricted to users with a low level of activity or whose past musical listening selections are very narrow. Through more in-depth investigations, we also discovered that while recommendations generally increase the *variety* of the songs recommended to users, they nonetheless fail to provide a *balanced* exposure to the different related categories.

Keywords: Network science, Multi-partite graphs, Random walk, Diversity, Recommender systems, Collaborative filtering, Filter Bubbles

Introduction

The ever-growing quantity of information and data available on online platforms has led to the need for efficient and reliable methods to filter this information. To this end, recommender systems have been introduced in different contexts, ranging from email filtering (Goldberg et al. 1992) to news exposure on social media platforms (Karimi et al. 2018), purchasing recommendations on online stores (Pradel et al. 2011), and generating playlists in streaming services (Hansen et al. 2020). The popularity of such systems lies in their ability to efficiently filter a high number of items to ensure that users are only presented with a few relevant ones.

While there has been a growing interest in developing such systems, the algorithms underpinning them have begun to face challenges, partly due to the over-personalization of their recommendations. Concerns have also been raised regarding the impact of algorithmic recommendations on users' behavior. Recently, for instance, news recommender systems have been criticized for their role in the appearance of echo chambers and the spread of fake news (Helberger 2019). In this context, *diversity* has been proposed as a way of mitigating the side effects resulting from the specialization of recommender systems (Helberger et al. 2018). However, while the scientific community usually agrees when it comes to the usefulness of diversity, little is known about how classic recommendation paradigms relate to the diversity of the content exposed to users.

This paper makes a significant step in this direction by analyzing the impact of a classic recommendation approach, namely the *Collaborative Filtering via Matrix Factorization for Implicit Datasets* (Hu et al. 2008), in relation to diversity. By conducting random walks on a tripartite graph that describes the relation between users, products, and categories, we were able to quantify the diversity of users' attention (Ramaciotti Morales et al. 2021) in relation to the categories before and after recommendations. We applied this approach to a dataset that records user activity on an online platform featuring musical content (Bertin-Mahieux et al. 2011) and study the impact the parameters of the model have on different aspects of the diversity of recommendations.

Our results show that the relation between algorithmic recommendations and diversity is complex. On the one hand, and in contrast with claims about the effect of algorithmic recommendations on limiting the diversity of content exposed to users (Pariser 2011), we show that recommendations do not necessarily limit diversity. On the contrary, the musical selections of a large proportion of users are in fact diversified by algorithmic recommendations. Moreover, we observe that diversity also increases with the number of items recommended.

On the other hand, however, such positive outcomes are mitigated by the fact that this trend depends strongly on the user's profile. First, diversity mostly increases for users with a low level of activity on the platform or whose past musical listening records are very narrow in scope. Second, by investigating in greater depth the recommendations exposed to users for whom diversity is increased, we were able to reveal which facet of the diversity is improved by collaborative filtering. If the recommendations generally increase the *variety* of the songs recommended to users, they nonetheless fail to provide a *balanced* exposure to the different related categories. When it comes to diversity, the extent to which recommendations fit in with a user's musical habits actually proves more important than the diversity of the recommendations itself.

We believe that our proposed method, the practical investigation we carried out on a collaborative filtering approach and the results we obtained on a real dataset all serve to offer a new perspective on how researchers can leverage network structures to examine the ethical effects of recommendation algorithms. It is worth noting here that we tested this approach on music exposure but we believe it could be applied to any other context with a similar network structure, thus paving the way for more general and systematic studies that would shed light on the effect of recommender systems on diversity.

Paper outline

After reviewing in Sect. Related work the existing literature on recommender systems and their relation to diversity measures, in Sect. Diversity in multi-partite graphs and recommender systems we provide details about the recommendation system we studied in this paper and the network science approach we used to measure its effect on diversity. In Sect. Analysing the effect of collaborative filtering in terms of diversity, we apply this approach to a specific dataset which records user activity on an online platform featuring musical content, before demonstrating the results achieved. Finally, in Sect. Conclusion and perspectives we conclude the paper and propose possible avenues for future work.

Related work

The question addressed in this paper falls within two independent lines of research: defining the notion of diversity in order to derive diversity measures, and analyzing recommender systems. Our approach relies on recent advances in both fields.

The concept of diversity

Independently from the analysis of the effect of recommender systems, diversity has been the subject of a wide variety of research studies in many different contexts and the question of which diversity measure should be used is a recurrent debate (Sakai and Zeng 2019). These range from ecology (McCann 2000) to economics (Gini 1921) and information theory (Rényi et al. 1961), to name just a few. In this line of research, there is a long tradition of proposing different indexes to quantify the diversity of a system. We could cite, for instance, the well-known Shannon and Rényi entropy, the Gini coefficient or the Hirschman-Herfindahl index.

In his seminal paper (Stirling 2007), Stirling observed that although the concept of diversity depends largely on the system being studied, common traits can be identified. Diversity is an irreducible property of a system (and not only of its parts) that can be expressed as a combination of variety, balance, and disparity. In Stirling's own words, *variety* is the number of categories into which system elements are apportioned; *balance* is a function of the pattern of apportionment of elements across categories; and *disparity* refers to the way and extent to which the elements may be distinguished.

Continuing from Stirling's work, the authors of Ramaciotti Morales et al. (2021) developed a theory of diversity measures and introduced a general methodology to formally quantify the different aspects of diversity as soon as the system under consideration can be represented as a network. This method relies on the distribution generated by random walks on a network that captures how the nodes from one layer (typically users) are related to the nodes from another layer (such as categories of products). By measuring the extent to which such a distribution differs from a uniform distribution, the authors define the α-diversity as a measure of the diversity of a system at different orders (see Sect. A formalism to analyse diversity for more details).

Interestingly, the different orders of the diversity directly refer to two of the three facets highlighted by Stirling. In particular, the 0-diversity captures the variety of a system exactly, while the ∞-diversity captures its balance. Any other value of α is then an attempt to take those two dimensions into consideration in the measure. For this reason, we chose to follow this approach by using the 0, 2 and ∞-diversity in the rest of our study.

Recommender systems and diversity

When it comes to recommender systems, the technique most commonly used to measure diversity is based on intra-list dissimilarity (Silveira et al. 2019), in particular in the context of the diversity of music recommendations (Zhang et al. 2012; Schedl et al. 2017), which is the context studied in this paper. Given a set of items $i \in I_u$ listened to by a user u, a dissimilarity metric $d(i, i')$ between items i and i' is constructed. When evaluating recommendations, it is then common to derive d from the cosine similarity between the matrix factorization latent vectors of i and i' (Cheng et al. 2017) and to take the average over all pairs (i, i'). Other metrics can be used, however, such as the inverse Pearson correlation (Vargas and Castells 2011) or the hamming distance (Kelly and Bridge 2006).

As an alternative, when matrix factorization is not used, the authors of Waller and Anderson (2019) used a community embedding model to obtain such vectors. Likewise, instead of taking the mean of the dissimilarities, one could use the maximum over any I_u k-subset of the minimum pairwise distance (Abbar et al. 2013). Another diversity measure, used in particular when metadata (such as tags or categories) are available, is the topic coverage (Li et al. 2020): the diversity relates to the number of topics reached by a user through the items he/she listens to. Dissimilarity metrics and topic coverage usually measure the *individual diversity* of each user, but they can also be computed over all the items listened to by all the users, leading to a measure of the *collective* (or *aggregate*) *diversity* (Shi 2013).

Most studies have focused on dissimilarity to incorporate this notion into recommendations. However, when it comes to the notion of balance or the use of such properties to analyze the effect of recommendations, little is known. One of the few works to have leveraged the volume of user-item interactions and item-tag strengths to capture the notion of balance is Vargas et al. (2014), while (Paudel et al. 2017) is the only recent work to have studied the effect of the parameters of a model on the diversity of recommendations.

To the best of our knowledge, the work presented in this paper is the first attempt to build on this extensive literature on diversity in recommender systems in the light of the new network science approach introduced in Ramaciotti Morales et al. (2021), in order to analyze the effect of a recommender system (and particularly its parameters) on both the variety and the balance of the content exposed to users.

Diversity in multi-partite graphs and recommender systems

When studied in fields such as ecology or economics, diversity metrics have historically been derived from probability distributions. In certain situations, such as the distribution of income or species, obtaining such distributions can be straightforward. In most cases, however, it requires a detailed understanding of the field and choosing between different possibilities remains largely subjective. In the case of musical content, for instance, would it be more relevant to analyze diversity in relation to the distribution of the songs listened to by a user, the musical categories to which they belong, the dissimilarity of the songs, or another aspect?

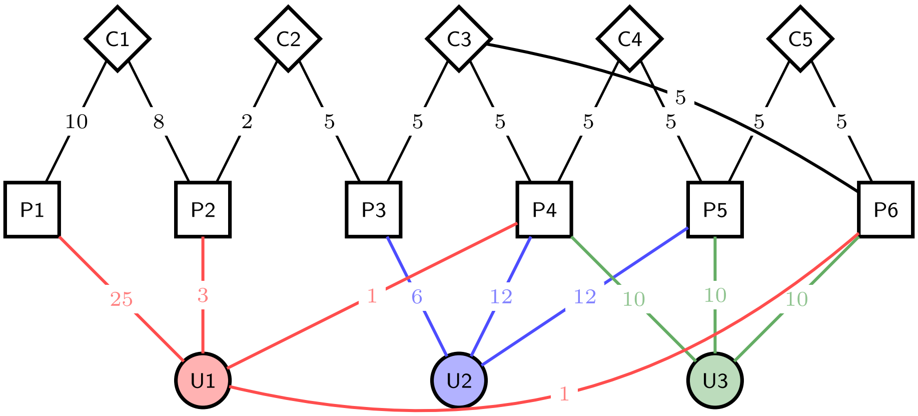


Fig. 1 Example of a tripartite graph

The work set out in Ramaciotti Morales et al. (2021) provides a framework for guiding such choices when the data can be represented as a *Heterogeneous Information Network*. In this section, we introduce this formalism (Sect. A formalism to analyse diversity), before then describing the particular recommender system that will be the focus of our study, namely the *Collaborative Filtering via Matrix Factorization for Implicit Datasets* (Hu et al. 2008) (Sect. Collaborative filtering).

A formalism to analyse diversity

Tripartite graphs and random walks

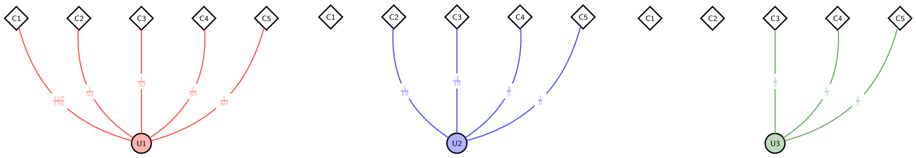
A tripartite graph is a graph whose nodes can be divided into three disjoint sets such that two nodes of one set cannot be connected by an edge. In perspective of using tripartite graphs in the context of users (U) listening to songs (I) related to musical categories (C), we restrict tripartite graphs to the case $\mathbb{T} = \left(U, I, C, E_U^I, E_I^C\right)$, with $E_U^I \subseteq U \times I$ and $E_I^C \subseteq I \times C$[1]. In addition, weights can be assigned to edges by defining the associated weight functions: $w_U^I : E_U^I \rightarrow \mathbb{R}_+$ (the number of times a user has listened to a song) and $w_I^C : E_I^C \rightarrow \mathbb{R}_+$ (the strength of the relation between a song and a musical category). For any node v we denote by $N(v)$ the set of its neighbors, $d(v)$ its degree and $d_w(v)$ its weighted degree[2]. An example of such a tripartite graph is given in Fig. 1.

Once represented as a tripartite graph, one would like to analyze how the induced relation between bottom (users) and top (categories) nodes are distributed. To do so, we rely on random walks on the tripartite structure. For every node v, we define the probability to reach a neighbor $z \in N(v)$ as $p_{v \rightarrow z} = \frac{w(v,z)}{d_w(v)}$. Then, for each bottom node $u \in U$ and top node $t \in C$, we define the probability $p_{u \rightarrow t}$ to reach t from u through I as:

$$p_{u \rightarrow t} = \sum_{i \in N(u) \cap N(t)} p_{u \rightarrow i} p_{i \rightarrow t} \tag{1}$$

[1] In general, a tripartite graph could have a set $E_U^C \subseteq U \times C$.

[2] Formally, for $v \in I$, we distinguish the set of C neighbors from the set of U ones.



(a) case of U_1 (b) case of U_2 (c) case of U_3

Fig. 2 Example of a bipartite graph projection issued from Fig. 1 with the transition probabilities for different users

If we repeat this process for each bottom node, we obtain the *bipartite projection* $\Pr(\mathbb{T}) = \left(U, C, E_U^C, w_{E_U^C}\right)$ of the tripartite graph into a bipartite graph where E_U^C is the set of edges between bottom nodes u and top nodes t such that there exists a *path* from u to t trough a middle node $v \in I$. The weights of the resulting edges are the transition probabilities $p_{u \to t}$ (see Fig. 2 for the induced projections and transition probabilities of users U_1, U_2 and U_3 from Fig. 1).

One of the strengths of this approach is that it provides a sound and interpretable way to extract different probability distributions from a user–song–category tripartite graph. In this paper, we consider a random walk restricted to the paths from U nodes to C ones, the aim being to obtain the distribution of the categories related to a user through his/her listening habits (weighted by the play count). This is similar to the work conducted in Vargas et al. (2014). Doing so, we measure the *individual diversity* of a user but the framework could also be used to extract other distributions that reflect on other aspects of diversity. For example, it could be used to measure the *collective* (or *aggregate*) *diversity* of a set of users by initiating the random walk from the (weighted) set of users instead of an individual one (see Ramaciotti Morales et al. (2021) for a complete review of the possible uses of this framework).

Diversity of users' attention

We base our diversity index on the *true diversity of order* α (or Hill Number) $D_\alpha(p)$ (Hill 1973; Jost 2006). The aim is to distinguish between a perfect situation in which all categories are reached uniformly by the users (highest diversity) and a worst situation in which few categories capture all the links (lowest diversity). More formally, for any probability vector p ($p_i \in [0, 1]$, $\sum_i p_i = 1$) and positive α, we define $D_\alpha(p)$ as:

$$D_\alpha(p) = \left(\sum_{i=1}^{k} p_i^\alpha\right)^{\frac{1}{1-\alpha}} \text{ if } \alpha \neq 1 \quad \text{and} \quad D_1(p) = \left(\prod_{i=1}^{k} p_i^{p_i}\right)^{-1} \text{ if } \alpha \to 1 \tag{2}$$

We then derive the α-diversity of a user's attention $u \in U$ as:

$$\alpha\text{−diversity}(u) = D_\alpha((p_{u \to t})_{t \in C}) \tag{3}$$

A straightforward interpretation of the α-diversity is that a high value indicates that a user reaches a wide range of categories almost uniformly, while a low value indicates a concentration of his interest towards a small and unbalanced number of categories.

However, this interpretation also depends on the order at which the diversity if measured. Interestingly, depending on the values of α, this diversity index expresses well-known diversity measures:

- $\alpha = 0$ is exactly the richness diversity (MacArthur 1965; Gotelli and Colwell 2011), which captures the notion of *variety* as defined by Stirling (Stirling 2007).
- $\alpha = 1$ is related to the Shannon entropy (Shannon 1948; Shannon et al. 1963) ($H(p)$):

$$H(p) = \log_2(D_1(p))$$

- $\alpha = 2$ is related to the Herfindal-Hirschman diversity (Rhoades 1993) ($HHI(p)$):

$$HHI(p) = \frac{1}{D_2(p)}$$

- $\alpha = \infty$ is related to the Berger-Parker diversity ($BBI(p)$) (Berger and Parker 1970) and captures the notion of *balance*:

$$BPI(p) = \frac{1}{D_\infty(p)}$$

Going back to the example of Figs. 1 and 2, one can see the interest of the different orders of diversity in order to differentiate between variety and balance. If one focuses on user U_1 for instance (Fig. 2a), one could clearly state that this is the most diverse user as he reaches all five categories of the tripartite graphs. This is captured by the 0-diversity ($D_0(U_1) = 5$) which is the highest value in the example ($D_0(U_2) = 4$ and $D_0(U_3) = 3$).

However, if one is interested in a balanced (or uniform) distribution of the relation between users and categories, one can clearly observe that U_1 is far from diverse since C_1 attracts almost all the paths starting from this user. This is clearly captured by the ∞ -diversity which is close to the lowest value 1 ($D_\infty(U_1) \sim 1.09$). To that regard, although U_3 reaches only 3 categories (Fig. 2c), the induced transition probabilities are completely uniform. This results in a ∞-diversity which is clearly higher ($D_\infty(U_3) = 3$) and U_3 is actually the user that presents the most balanced profile in the example ($D_\infty(U_2) = 2.5$ in comparison).

Finally, U_2 (Fig. 2b) is the user that presents the best profile if one wants to take into consideration both variety and balance in the diversity measure. Although the user is related to less categories than U_1 and that the distribution probabilities issued from the random walks are not as uniformly distributed than the ones issued from U_3, his profile is sufficiently wide and balanced for obtaining the highest 2-diversity of the example: $D_2(U_2) \sim 3.33$ while $D_2(U_1) \sim 1.20$ and $D_2(U_3) = 3$.

Having these observations in mind, in the rest of the paper we will systematically study the α-diversity for $\alpha = 0$ (variety), $\alpha = \infty$ (balance) and $\alpha = 2$ which is a way to take into account the two dimensions. This will provide a more comprehensive picture of the impact of recommender systems on those different facets of diversity.

Collaborative filtering

The model

In addition to the set U and I, let $\mathcal{R} = \left\{ r_{ui} \mid u \,\text{rated}\, i \right\}$ be the matrix describing the preferences of users as regards the items. When an item i has never been listened to by a user u, we assume $r_{ui} = 0$, otherwise we take the play count. We define c_{ui} as the confidence the model has in the proposition "the user u likes the item i". Because there is no canonical relation between r_{ui} and c_{ui} we will use a simple model as suggested in Hu et al. (2008). This model is defined in Equation 4 (with $\mu \geq 0$) where the variable p_{ui} is introduced to describe whether a user likes an item or not. We consider that as soon as a user listens to a song, he/she is prone to like it, therefore $p_{ui} = 1$.

$$c_{ui} = 1 + \mu r_{ui} \quad \text{and} \quad p_{ui} = \begin{cases} 1 \text{ if } r_{ui} > 0 \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

As with classic matrix factorization, we assume that each user u (resp. each item i) can be represented by a column vector of *latent factors* x_u (resp. y_i). We name $x = (x_{u_1} \ldots x_{u_n})$ (resp. $y = (y_{i_1} \ldots y_{i_m})$) the matrix of user factors (resp. item factors) and $\widehat{p}_{ui} = x_u^T y_i$ the estimator of p_{ui}. The values x^* and y^* of the *latent factors* are computed by minimizing the weighted mean squared error between p_{ui} and its estimator $\widehat{p}_{ui}$.

$$x^*, y^* = \min_{x,y} \sum_{u,i} c_{ui} \left(p_{ui} - x_u^T y_i \right)^2 + \lambda \left(\|x\|_2^2 + \|y\|_2^2 \right) \tag{5}$$

In order to prevent the model from overfitting the training data, we add a regularization term with $\lambda \geq 0$ and $\|x\|_2 = \sqrt{\sum_k \sum_l x_{kl}^2}$ the Frobenius norm. To recommend a set of items to a user u, we first select a set of candidates I_u. This set contains all the items i a user did not listened to: $I_u = \{i \in I \mid r_{ui} = 0\}$. These items are then sorted by decreasing order of $\widehat{p}_{ui}$. Finally, the recommended items are the k-best items in the sorted candidates list I_u.

Training and evaluation

The model in 5 is optimized with a *Regularized Alternative Least Squares* method as described in Hu et al. (2008). However, each least square problem (which is equivalent to solving a linear system) is solved via *Conjugate Gradient Descent* (Takács et al. 2011).

Finally, the evaluation is conducted as follow. We first randomly select a proportion β of users in U, and name the resulting set U_{sel}. Then, for each user $u \in U_{\text{sel}}$, we randomly select a proportion of items previously listened to by u (*ie.* with rating $r_{ui} > 0$) and add the corresponding ratings in the global user–item test set $\mathcal{R}_{\text{test}}$. All the remaining ratings are used to create the training set $\mathcal{R}_{\text{train}} = \mathcal{R} \backslash \mathcal{R}_{\text{test}}$. Finally, we make sure that while sampling items listened to by a user u, at least one item stays in the train set. This results in a testing set $\mathcal{R}_{\text{test}}$ in which all the users have been encountered during training, thus eliminating the cold start problem. Assuming that the goal of a recommender system is to produce ordered lists that best match the preference of a user, we assess the performance of the model with the *Mean Normalized Discounted*

Cumulative Gain. Given a recommended list of items $L_u = (i_1, \ldots, i_{|L|})$, ordered by their score $\widehat{p}_{ui}$, we define the *Discounted Cumulative Gain* (DCG) in 6. To obtain the *Normalized Discounted Cumulative Gain* (NDCG), we generate an ideal recommendation list $L_{u,\text{ideal}}$ with the test data (the user's most listened items sorted by decreasing play count).

$$\mathrm{DCG}(L_u, u) = \sum_{k=1}^{|L|} \frac{r_{ui_k}}{\log_2(i)} \quad \text{and} \quad \mathrm{NDCG}(L_u, u) = \frac{\mathrm{DCG}(L_u, u)}{\mathrm{DCG}(L_{u,\text{ideal}}, u)} \tag{6}$$

Analysing the effect of collaborative filtering in terms of diversity

This section presents the results we obtained by using the approach presented in Sect. Diversity in multi-partite graphs and recommender systems in the context of a dataset recording user activity on an online platform featuring musical content. We first present the general setting (Sect. Experimental setting), before investigating different questions: how diversified are the recommendations and what is the impact of the parameters (Sect. Analysis of the recommendations)? What is the effect of the recommendations on users' diversity (Sect. Effect of the recommendations on users' diversity)? How to explain the differences in the way recommendations affects users' diversity (Sect. Examining the effect of representative recommendations)?

Experimental setting

Dataset

The dataset we used comes from the *Million Song Dataset* project (Bertin-Mahieux et al. 2011). To create the first two layers of the triparite graph and the user–item links, we used the *Echo Nest user taste profile* dataset of the project. It features triplets of (user, item, play count) that describe how many times a user has listened to a given song. To add the third layer of the tripartite graph and create the item–category links, we used the *last.fm* dataset of the project. It contains (item, tag, strength) triplets that describe the tags associated to each song with their strength. Furthermore, in order to obtain a coherent tripartite graph, we performed the following operations: we selected only the 1, 000 most popular tags[3], deleted songs with ambiguous identifiers[4], and deleted songs with no tag and users with no songs. Finally, in order to reduce the training time of our models, we randomly sampled 100, 000 users and their recorded items.

Implementation

We ran all of the experiments using the Python programming language on a 40 cores `Intel Xeon` server, equipped with 256 GB of RAM. We used the package `lenskit` (Ekstrand 2020) to instantiate, train and evaluate the recommender system. The code is available on GitHub[5] along with instructions to install and run it. Following

[3] This step was necessary to avoid misleading interpretations due to the inconsistency of the use of the tags in the dataset.

[4] See http://millionsongdataset.com/blog/12-2-12-fixing-matching-errors/.

[5] https://github.com/grodino/recodiv

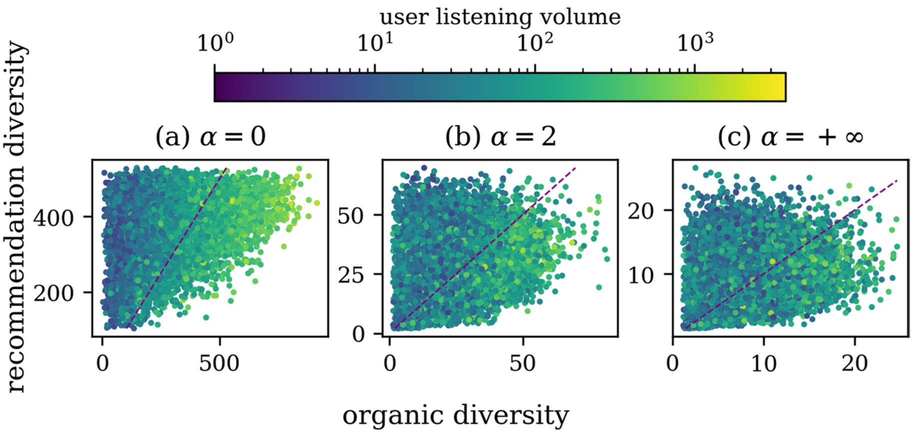


Fig. 3 Diversity of the recommendations ($k = 10$) with respect to the organic diversity of the users

the test procedure described in Sect. Collaborative filtering, the dataset was split in 5 user folds, using the leave-one-out strategy to create the train/test datasets. The hyperparameters were then chosen to empirically maximize the average *Normalized Disconted Cumulative Gain*, which resulted in 512 latent factors, $\mu = 40$ and $\lambda = 5 * 10^{-3}$. Unless stated otherwise, those are the values used in the rest of the paper. When studying particular users the first fold was used.

Tripartite graph of the recommendations

To evaluate the diversity of the recommendations, we define a second tripartite graph $\mathbb{T}_r$. In this graph, the links between songs and categories are the same as in the dataset but the links between the users and the songs now represent the recommendations (instead of the musical record of the user). In addition, to account for the impact of the position $rank(i, u)$ of an item i in the list L_u of recommendations generated to user u, the weight of the corresponding u–i link in the tripartite graph is set to $|L_u| - rank(i, u)$. The *diversity of the recommendations* L_u exposed to a user u is therefore the α-diversity of u in $\mathbb{T}_r$. Finally, to differentiate between the two tripartite graphs, we will refer to the *organic diversity* of a user as his/her diversity *before* being exposed to the recommendations.

Analysis of the recommendations

Diversity of the recommendations

First, we investigate the properties of the recommendations in terms of diversity. Figure 3 presents the recommendation diversity (for $k = 50$ recommendations and $\alpha = 0, 2$ and ∞) with respect to the organic diversity of each user. The *users' volume* (the sum of the play counts of all items the user has listened to) is represented by a color in a log scale[6].

This figure provides several pieces of information. First, we can observe that there is no strong relation between organic diversity and recommendation diversity since low organic diversity does not necessarily lead to low recommendation diversity. Some users

[6] Because the model is only trained on a subset of the dataset, the organic diversity and the volume are computed on the train dataset.

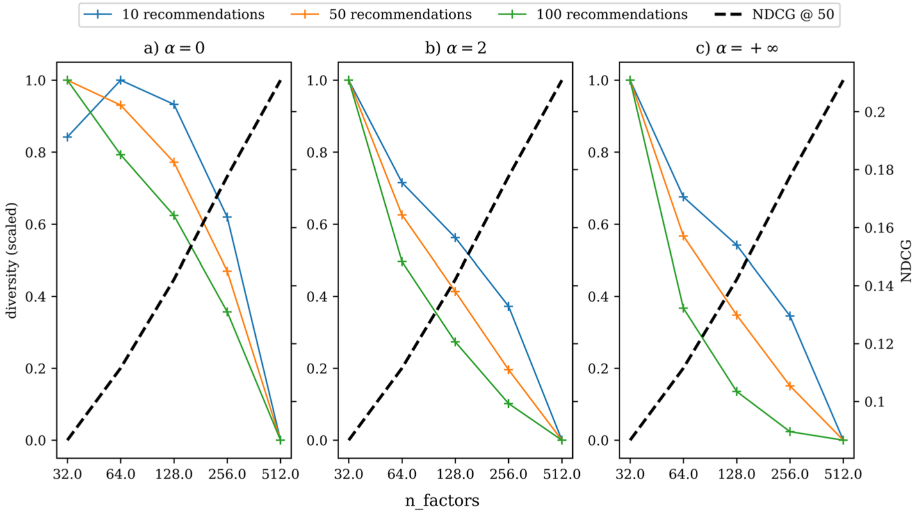


Fig. 4 Diversity (normalized) and performance of the recommendations

with a low organic diversity are exposed to diversified recommendations,while others have a narrower exposure. In addition, and whatever the order of the diversity, we can observe that, as soon as a minimum number of items are recommended, the recommendations tend to be more diverse than the organic ones. This is particularly true for the variety captured by $\alpha = 0$, but it can also be observed (although to a lesser extent) for $\alpha = 2$ and $\alpha = \infty$.

However, one can clearly see that this relation depends on the value of the organic diversity. As this value increases (for all values of α), it becomes more difficult for the recommendations to reach the same level of diversity.

Impact of the parameters of the model

As regards the impact of the parameters of the model, it has been suggested that there is a trade-off between the usual notion of performance (engagement or accuracy) and the notion of diversity (Holtz et al. 2020). To investigate this relation, in Fig. 4 we show how the number of latent factors impacts the diversity[7] of the recommendations exposed to the users, along with the model performance (measured by the NDCG, in black dashed lines).

As expected, it can be observed that when the number of latent factors increases, so does the performance of the model. With more latent factors, the model is efficient in recommending items related to the user's past musical record. However, this efficiency has a clear effect on diversity, which decreases when the number of latent factor increases, except for a particularly low number of recommendations ($k = 10$) and for $\alpha = 0$. In this situation, there does seem to be a trade-off between performance and diversity. This supports the observations made in Holtz et al. (2020) and clarifies the effect: the suggested trade-off can be observed for the *variety* but vanishes as soon as the *balance* exposure is taken into account in the diversity measure.

[7] The values have been normalized in order to ease the comparison.

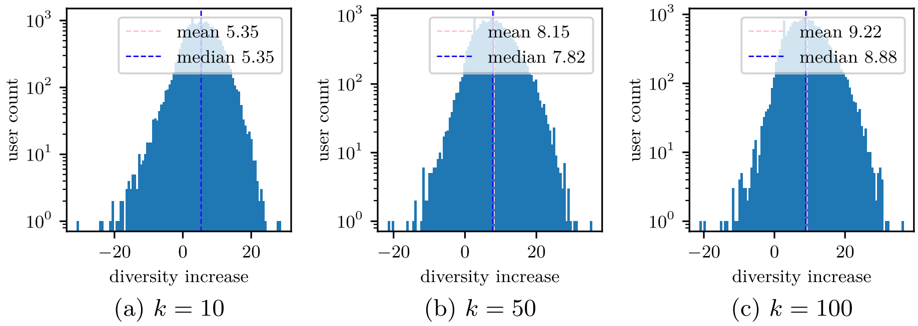


Fig. 5 Distribution of the diversity increase ($\alpha = 2$) for different numbers of recommendations

Effect of the recommendations on users' diversity

Figure 3 presented in Sect. Analysis of the recommendations reveals that a list of recommendations can be diversified, even for users whose past musical records are narrow in scope. However, this plot does not provide any information on the real impact of the recommendations on users' musical habits. How diverse are the user's listening habits *after being exposed* to the recommendations? To investigate this question, for each user we computed the *diversity increase*, which is defined as

$$\Delta_\alpha = D_\alpha(\mathbb{T}_{t,r}) - D_\alpha(\mathbb{T}_t) \tag{7}$$

where $\mathbb{T}_t$ is the tripartite graph associated to the training dataset and $\mathbb{T}_{t+r}$ the tripartite graph in which we added the recommendations. Thus a positive value of Δ_α indicates that the recommendations have improved the musical habits of a user in terms of diversity, while a negative value indicates the opposite effect.

It is worth noting here that, in order for this diversity measure to make sense, one has to carefully adapt the weights in $\mathbb{T}_{t+r}$ and in particular to derive a relevant notion of play count for the recommendations. We chose to use a linear relation between the weight and the rank. Assuming that a user would listen to as many recommendations as the number u_v of items he/she has listened to and that the last item recommended is not listened to ($w(u, i_{n_r}) = 0$), we used the following weight function:

$$w(u,i) = \frac{2u_v}{k(k-1)}(k - \text{rank}(i,u))$$

This choice clearly hides a simple user model and other choices could be investigated. In particular, one could use the models proposed in Poulain and Tarissan (2020) that account for the saturation effects in the diversity of users' attention.

Figure 5 presents the distribution of the diversity increase (with $\alpha = 2$) for different numbers of recommended items ($k = 10, 50, 100$). Surprisingly, contrary to what Fig. 3 suggested, the musical listening habits of most users are diversified by the recommendations. Even with only ten items recommended, there is a positive increase for more than 89% of users. However, the log-scale of the y-axis could be misleading and one could object that the increase, although mostly positive, is relatively small. This does in fact prove to be the case and one can observe that the mean increase is always lower than 10, even for hundreds of recommendations.

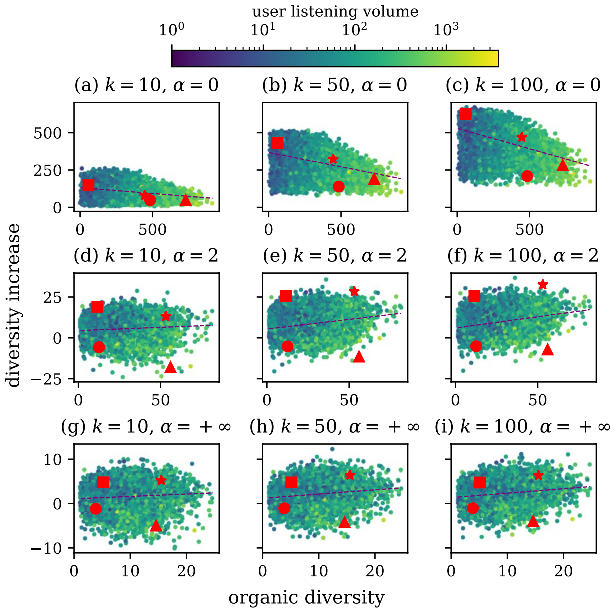


Fig. 6 Diversity increase with respect to the organic diversity for different numbers of recommendations (k) and diversity orders (α). Dots marked with a red symbol refer to the users studied in Sect. Examining the effect of representative recommendations

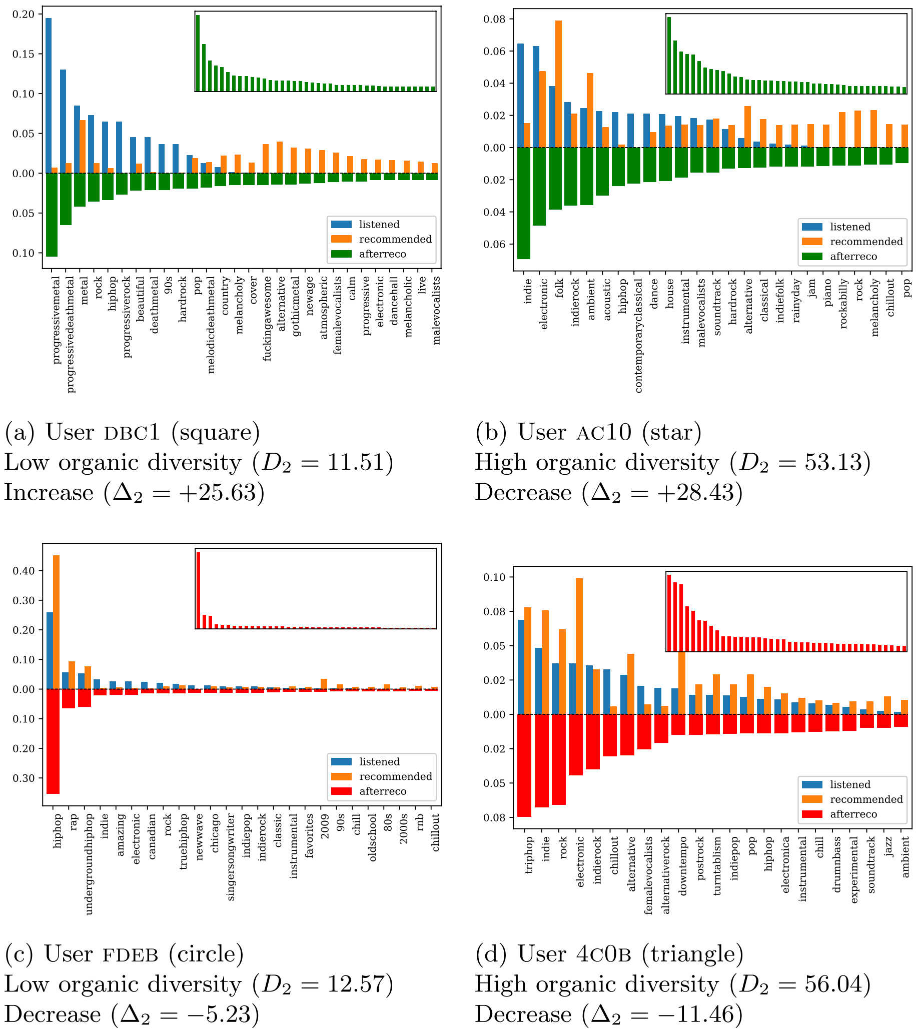
To achieve a more detailed and comprehensive picture of the effect recommendations have on users' diversity, in Fig. 6 we have plotted the diversity increase for different numbers of recommendations ($k = 10, 50, 100$) and different orders of diversity ($\alpha = 0, 2, \infty$) for all users. The effect of the number of recommendations on the diversity of users' attention is clear: diversity increases with the number of items recommended (from left to right) for all diversity orders. However, one can also observe that the capacity of the recommendations to improve users' diversity is closely related to the users' profiles. Such an increase can mainly be seen in users that listen to a small number of songs (darker dots on the plots). For very active users or those with broad musical listening habits, the recommendations barely have any positive increase.

The plots also reveal that the recommendations do not have the same effect when diversity is considered in terms of variety or balance. In relation to variety ($\alpha = 0$, top row), we can see that the recommendations improve diversity even with very few recommendations ($k = 10$). This is not the case for other diversity orders. This suggests that the main effect of the recommendations is to introduce new categories into users' musical habits. As soon as balance is taken into account in the measure (middle and bottom row), the recommendations are less effective in increasing diversity.

These results indicate that collaborative filtering impacts the diversity of musical listening habits in a specific way. To study this relation in greater depth, we will now investigate exactly how the recommendations affect users' diversity.

Examining the effect of representative recommendations

We will conclude this section by presenting the cases of some specific users. This will shed light on the relation between recommended items and users' past musical records,



(a) User DBC1 (square)
Low organic diversity ($D_2 = 11.51$)
Increase ($\Delta_2 = +25.63$)

(b) User AC10 (star)
High organic diversity ($D_2 = 53.13$)
Decrease ($\Delta_2 = +28.43$)

(c) User FDEB (circle)
Low organic diversity ($D_2 = 12.57$)
Decrease ($\Delta_2 = -5.23$)

(d) User 4C0B (triangle)
High organic diversity ($D_2 = 56.04$)
Decrease ($\Delta_2 = -11.46$)

Fig. 7 Effect of recommendations ($k = 50$) on different users

and how, as a result, recommendations have different effects on users' attention. In Fig. 7, we present four users whose past musical records are either narrow (left column) or diverse (right column) and for whom recommendations ($k = 50$) generate either a decrease (bottom row) or an increase (top row) in diversity ($\alpha = 2$). For each user case, the top part of the plot displays the proportion of the categories of the past musical record (left blue line) and of the recommendations (right orange line) independently, while the bottom part shows the distribution of the categories as a result of the recommendations, that is *after the user is exposed* to the recommendations. It is worth noting that, to ease the comparison between the different cases, we only display the most important categories[8]. Finally, users presented in this picture are also visible in Fig. 6 with their respective red marker.

This figure reveals some interesting aspects in terms of how recommendations impact users. First, in all four cases, a significant fraction of the musical categories associated

[8] The reader might refer to the inset for a larger part of the final distribution.

with the recommendations are completely new to the users, thus increasing variety. This is obvious in Fig. 7a–c, even if only the main categories are displayed. In Fig. 7d, the new categories belong to less dominant categories, thus not visible in the distribution. To support this observation, we computed the proportion of new categories when a positive increase is observed in the whole dataset: on average, a positive diversity increase (for $\alpha = 2$ and $k = 50$) leads to 322 new categories. This more than quadruples the number of categories of the past musical record of an average user.

Second, we can observe that users with a similar profile in terms of organic diversity might be differently affected by recommendations. The real effect of the recommendations relies strongly on how they fit in with the users' musical habits. For user FDEB (bottom left), for instance, the most recommended category, *hiphop*, corresponds to the main category in his/her past musical record. This completely unbalances the musical landscape of this user, whose musical record was already largely restricted to hiphop music.

By contrast, the musical exposure of user DBC1 (top left) is diversified as a result of the recommendations, although he/she also has one particular musical focus: *progressive metal* and *progressive death metal*. One reason for this is that those categories barely appear in the list of recommendations (only metal, related to the former, is highly represented). Instead, the recommendations expose other new or less dominant categories in the user's past musical record (such as *alternative* or *gothic metal*). This provides a far more balanced range of musical categories, leading to a higher diversity.

One might wonder whether this effect could be due to the fact that those users have a low organic diversity to begin with. The study of users 4C0B and AC10 (right column) show this is not the case. Both are highly active users (respectively in the top 7% and top 25% of the most active users on the dataset) with a very high organic diversity (top 2%). Yet, they both experience a completely different impact of the recommendations exposure : while the four main categories of user 4C0B (*triphop*, *indie*, *rock* and *electronic*) are also the four main categories of the recommendations, thus unbalancing the range of musical categories to which he/she is exposed, the musical record of user AC10, on the contrary, is broadened by the recommendations due to the more nuanced musical exposure. In particular, on can notice that *indie* (his/her most listened categories) is one of the least recommended.

Conclusion and perspectives

In this paper, we have investigated the impact recommender systems have on the diversity of users' attention. Specifically, we examined a recently proposed framework that exploits the relations between users, items and categories to measure the diversity of a user's attention. By applying a collaborative filtering approach to a dataset that records users' past musical records, we were able to undertake a detailed analysis of how recommendations affect diversity in this context.

The results presented in this paper all show that recommendations tend to have a relatively high degree of diversity (Fig. 3), and globally improve the diversity of most users (Fig. 5). However, there are some limits to this capacity to diversify users' musical habits. First, it usually undermines the performance of the models (Fig. 4). Second, this improvement does not benefit all users. Rather, it is limited to users with a low level of activity or whose musical records are narrow in spectrum (Fig. 6). Finally, when recommendations do successfully improve diversity, this is mainly due to the discovery of new categories close to the musical records of users, which enhances *variety*. It usually fails, however, to provide a *balance* exposure to the different musical categories (Fig. 7).

These results are in line with recent papers which identified key complex effect of algorithmic recommendations in the context of musical platforms, although using different approaches. In particular, it is worth noting that (Anderson et al. 2020) also specifically identified that algorithmic recommendations are more effective for users with a lower organic diversity, while (Villermet et al. 2021) showed that if recommendations tend generally to drive new exploration of least popular content (hence increasing variety), such effect primarily depends on the users mode of consumption. These results, along with the ones presented in this paper, all show that when it comes to diversity, the extent to which recommendations fit in with a user's musical habits actually proves more important than the diversity of the recommendations itself.

We believe that the method proposed in this paper, as well as the practical investigation conducted on a collaborative filtering approach, applied to a real musical dataset, shed new light on how researchers could leverage network representations to examine the ethical effects of algorithmic recommendations. This approach calls for future investigations, two of which are discussed below.

Individual versus collective diversity

By focusing on *individual diversity* (that is, the diversity computed from one node of the network), we were able to highlight the impact recommendations have for each individual user. We then used the average computed across all the individual diversities to measure the impact the recommendation algorithm has from a global perspective. However, this approach fails to uncover certain situations that could be intuitively described as not diversified. For example, if a very diverse subset of items were systematically recommended to all users, the average individual diversity would be measured as high, although one could argue that this is extremely undiversified from a global perspective since a unique subset is exposed to all users. Fortunately, the framework we used in this paper makes it possible to detect this situation by measuring *collective diversity* in addition to individual diversity. Analyzing the effect of recommendations through these collective lenses would undoubtedly provide some meaningful insights. This approach would be helpful, for instance, for measuring the effect of polarization in news recommendations (Celis et al. 2019).

Integrating dissimilarity

In contrast with most of the literature on recommender systems, we did not explicitly explore the dissimilarity facet of diversity. Instead, we focused on the variety (also referred to as the â€˜coverage') and the balance of the exposure. However, most standard dissimilarity metrics are based on the scalar product between the vectors of items, extracted from users' musical records. Therefore, these metrics result in a combination of user dissimilarity and item dissimilarity. Translated in the context of the diversity measure we used, dissimilarity is close to what would be measured with the meta-path *User→Item→User→Item→Category*. Adding such a dimension to the approach proposed in this paper would pave the way for analyzing the three facets of the diversity, as highlighted by Stirgling (Stirling 2007), in a unified framework.

Author contributions

AG and FT contributed equally to the work presented in this papier. AG contributed in particular to data curation and preparation, design and implementation the recommander system, training of the model, simulations, analysis, visualisation and edition of the paper. FT contributed in particular to the methodology, conception and design of the work, analysis and interpretation of results, supervision, validation, drafting the first version of the paper and edition. All authors read and approved the final manuscript.

Funding

This work has been funded in part by the Agency of French National Research (ANR) Grant ANR-20-CE38-0011.

Availability of data and materials

The dataset used in this study is composed of two dataset collected independently but freely available on the website of the Million Song Dataset project: http://millionsongdataset.com/ In particular, we used: - the user-taste profile dataset: http://millionsongdataset.com/tasteprofile/ - the Last.fm dataset: http://millionsongdataset.com/lastfm/ These two datasets were merged as described in Sect. Analysing the effect of collaborative filtering in terms of diversity. An anonymized GitHub archive is also provided. It contains the programs that can be used in order to reproduce the results presented in the present paper.

Declarations

Competing interests

The Authors declare to have no competing interest with the content of the study.

Received: 5 September 2022 Accepted: 22 December 2022
Published online: 25 January 2023

References

Abbar S, Amer-Yahia S, Indyk P, Mahabadi S (2013) Real-time recommendation of diverse related articles. In: Proceedings of the 22nd international conference on world wide web. WWW '13. Association for Computing Machinery, New York, pp 1–12. https://doi.org/10.1145/2488388.2488390

Anderson A, Maystre L, Anderson I, Mehrotra R, Lalmas M (2020) Algorithmic effects on the diversity of consumption on spotify. Association for Computing Machinery, New York, pp 2155–2165. https://doi.org/10.1145/3366423.3380281

Berger WH, Parker FL (1970) Diversity of planktonic foraminifera in deep-sea sediments. Science 168(3937):1345–1347

Bertin-Mahieux T, Ellis DPW, Whitman B, Lamere P (2011) The million song dataset. In: Proceedings of the 12th international conference on music information retrieval (ISMIR 2011)

Celis LE, Kapoor S, Salehi F, Vishnoi N (2019) Controlling polarization in personalization: an algorithmic framework. In: Proceedings of the conference on fairness, accountability, and transparency. FAT* '19. Association for Computing Machinery, New York, pp 160–169. https://doi.org/10.1145/3287560.3287601

Cheng P, Wang S, Ma J, Sun J, Xiong H (2017) Learning to Recommend Accurate and Diverse Items. In: Proceedings of the 26th international conference on world wide web. WWW '17, pp. 183–192. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE . https://doi.org/10.1145/3038912.3052585

Ekstrand MD (2020) LensKit for Python: next-generation software for recommender systems experiments. In: Proceedings of the 29th ACM international conference on information & knowledge management. CIKM '20. Association for Computing Machinery, New York, pp 2999–3006. https://doi.org/10.1145/3340531.3412778

Gini C (1921) Measurement of inequality of incomes. Econ J 31(121):124–126. https://doi.org/10.2307/2223319

Goldberg D, Nichols D, Oki BM, Terry D (1992) Using collaborative filtering to weave an information tapestry. Commun ACM 35(12):61–70. https://doi.org/10.1145/138859.138867

Gotelli NJ, Colwell RK (2011) Estimating species richness. Biol Divers Front Meas Assess 12:39–54

Hansen C, Hansen C, Maystre L, Mehrotra R, Brost B, Tomasi F, Lalmas M (2020) Contextual and sequential user embeddings for large-scale music recommendation. In: Fourteenth ACM conference on recommender systems. ACM, Virtual Event Brazil, pp 53–62. https://doi.org/10.1145/3383313.3412248

Helberger N (2019) On the democratic role of news recommenders. Digit J 7(8):993–1012. https://doi.org/10.1080/21670811.2019.1623700

Helberger N, Karppinen K, D'Acunto L (2018) Exposure diversity as a design principle for recommender systems. Inf Commun Soc 21(2):191–207. https://doi.org/10.1080/1369118X.2016.1271900

Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54(2):427–432. https://doi.org/10.2307/1934352

Holtz D, Carterette B, Chandar P, Nazari Z, Cramer H, Aral S (2020) The engagement-diversity connection: evidence from a field experiment on spotify. In: Proceedings of the 21st ACM conference on economics and computation. EC '20. Association for Computing Machinery, New York, pp 75–76. https://doi.org/10.1145/3391403.3399532

Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: 2008 eighth IEEE international conference on data mining. IEEE, Pisa, pp 263–272. https://doi.org/10.1109/ICDM.2008.22

Jost L (2006) Entropy and diversity. Oikos 113(2):363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Karimi M, Jannach D, Jugovac M (2018) News recommender systems—survey and roads ahead. Inf Process Manag 54(6):1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

Kelly JP, Bridge D (2006) Enhancing the diversity of conversational collaborative recommendations: a comparison. Artif Intell Rev 25(1–2):79–95. https://doi.org/10.1007/s10462-007-9023-8

Li C, Feng H, de Rijke M (2020) Cascading hybrid bandits: online learning to rank for relevance and diversity. In: Fourteenth ACM conference on recommender systems. RecSys '20. Association for Computing Machinery, New York, pp 33–42. https://doi.org/10.1145/3383313.3412245

MacArthur RH (1965) Patterns of species diversity. Biol Rev 40(4):510–533

McCann KS (2000) The diversity–stability debate. Nature 405(6783):228–233. https://doi.org/10.1038/35012234

Pariser E (2011) The filter bubble: what the internet is hiding from you. Penguin Books Limited, London

Paudel B, Haas T, Bernstein A (2017) Fewer flops at the top: accuracy, diversity, and regularization in two-class collaborative filtering. In: Proceedings of the eleventh ACM conference on recommender systems. RecSys '17. Association for Computing Machinery, New York, pp 215–223. https://doi.org/10.1145/3109859.3109916

Poulain R, Tarissan F (2020) Investigating the lack of diversity in user behavior: the case of musical content on online platforms. Inf Process Manag 57(2):102169. https://doi.org/10.1016/j.ipm.2019.102169

Pradel B, Sean S, Delporte J, Guérif S, Rouveirol C, Usunier N, Fogelman-Soulié F, Dufau-Joel F (2011) A case study in a recommender system based on purchase data. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '11, pp. 377–385. Association for Computing Machinery, New York. https://doi.org/10.1145/2020408.2020470

Ramaciotti Morales P, Lamarche-Perrin R, Fournier-S'niehotta R, Poulain R, Tabourier L, Tarissan F (2021) Measuring diversity in heterogeneous information networks. Theor Comput Sci 859:80–115. https://doi.org/10.1016/j.tcs.2021.01.013

Rényi A et al (1961) On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics. The Regents of the University of California, pp 547–561

Rhoades SA (1993) The herfindahl-hirschman index. Fed Res Bull 79:188

Sakai T, Zeng Z (2019) Which diversity evaluation measures are "good"? In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery, New York, pp 595–604. https://doi.org/10.1145/3331184.3331215

Schedl M, Knees P, Gouyon F (2017) New paths in music recommender systems research. In: Proceedings of the eleventh ACM conference on recommender systems. ACM, pp 392–393

Shannon CE, Weaver W (1963) The mathematical theory of communication. 1949. University of Illinois Press, Urbana

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Shi L (2013) Trading-off among accuracy, similarity, diversity, and long-tail: a graph-based recommendation approach. In: Proceedings of the 7th ACM conference on recommender systems. RecSys '13. Association for Computing Machinery, New York, pp 57–64. https://doi.org/10.1145/2507157.2507165

Silveira T, Zhang M, Lin X, Liu Y, Ma S (2019) How good your recommender system is? A survey on evaluations in recommendation. Int J Mach Learn Cybern 10(5):813–831. https://doi.org/10.1007/s13042-017-0762-9

Stirling A (2007) A general framework for analysing diversity in science, technology and society. J R Soc Interface 4(15):707–719. https://doi.org/10.1098/rsif.2007.0213

Takács G, Pilászy I, Tikk D (2011) Applications of the conjugate gradient method for implicit feedback collaborative filtering. In: Proceedings of the Fifth ACM conference on recommender systems—RecSys '11. ACM Press, Chicago, p 297. https://doi.org/10.1145/2043932.2043987

Vargas S, Baltrunas L, Karatzoglou A, Castells P (2014) Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: Proceedings of the 8th ACM conference on recommender systems. RecSys '14. Association for Computing Machinery, New York, pp 209–216. https://doi.org/10.1145/2645710.2645743

Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the fifth ACM conference on recommender systems. ACM, pp 109–116

Villermet Q, Poiroux J, Moussallam M, Louail T, Roth C (2021) Follow the guides: disentangling human and algorithmic curation in online music consumption. Association for Computing Machinery, New York, pp 380–389. https://doi.org/10.1145/3460231.3474269

Waller I, Anderson A (2019) Generalists and specialists: using community embeddings to quantify activity diversity in online platforms. In: The world wide web conference. WWW '19, pp 1954–1964. Association for Computing Machinery, New York. https://doi.org/10.1145/3308558.3313729

Zhang YC, Séaghdha D.Ó, Quercia D, Jambor T (2012) Auralist: introducing serendipity into music recommendation. In: Proceedings of the Fifth ACM international conference on web search and data mining. ACM, pp 13–22

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.