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

AUGUSTIN GODINOT, Ecole Normale Supérieure Paris-Saclay, France
FABIEN TARISSAN, Université Paris-Saclay, CNRS, ISP, Ecole Normale Supérieure Paris-Saclay, France

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. However, we still know little about how classic recommendation paradigms affect users' behavior in terms of diversity. 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*. We study the impact the parameters of the model have on different aspects of the diversity of recommendations, as well as the diversity of users' attention before and after recommendations.

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 and that diversity also increases with the number of items recommended. 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. 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.

ACM Reference Format:
Augustin Godinot and Fabien Tarissan. 2021. Measuring the effect of collaborative filtering on the diversity of users' attention. In *RecSys '21: The ACM Conference Series on Recommender Systems, Sept. 27–Oct. 01, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/00.0000/0000000.0000000

1 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 [7] to news exposure on social media platforms [15], purchasing recommendations on online stores [22], and generating playlists in streaming services [8]. 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

Manuscript submitted to ACM

systems have been criticized for their role in the appearance of echo chambers and the spread of fake news [9]. In this context, *diversity* has been proposed as a way of mitigating the side effects resulting from the specialization of recommender systems [10]. 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* [13], in relation to diversity. By conducting random walks on heterogeneous information networks that describe the relation between users, products, and categories, we were able to quantify the diversity of users' attention [23] 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 [2] 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 [19], 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.

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 working on recommender systems could examine the ethical effects of 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 Section 2 the existing literature on recommender systems and their relation to diversity measures , in Section 3 we provide details about the recommendation system we studied in this paper and the method we used to measure its effect on diversity. In Section 4, 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 Section 5 we conclude the paper and propose possible avenues for future work.

2 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. These range from ecology [18] to economics [6] and information theory [24], 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 [28], 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 [23] 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 Section 3.1 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 [27], in particular in the context of the diversity of music recommendations [25, 33], 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' [4] and to take the average over all pairs (i, i'). Other metrics can be used, however, such as the inverse Pearson correlation [31] or the hamming distance [16].

As an alternative, when matrix factorization is not used, the authors of [32] 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 [1]. Another diversity measure, used in particular when metadata (such as tags or categories) are available, is the topic coverage [17]: 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* [26].

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 [30], while [20] 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 framework introduced in [23], 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.

3 DIVERSITY IN HETEROGENEOUS INFORMATION NETWORKS 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?

The work set out in [23] 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 (Section 3.1), 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* [13] (Setion 3.2).

3.1 A formalism to analyse diversity

Tripartite graph. 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: $w_U^I : E_U^I \to \mathbb{R}_+$ (the number of times a user has listened to a song) and $w_I^C : E_I^C \to \mathbb{R}_+$ (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 Figure 1(a).

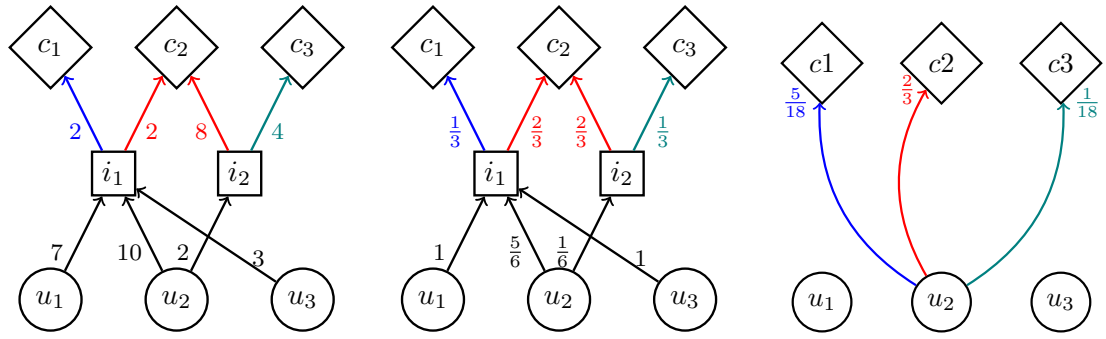
Random walk and bipartite projection. For every node v, we define the probability to reach a neighbor $z \in N(v)$ as $p_{v\to z} = \frac{w(v,z)}{d_w(v)}$. Then, for each bottom node $u \in U$ and top node $t \in C$, we define the probability $p_{u\to t}$ to reach t from u through I as:

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

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 Figure 1(c) for an example centered on user u_2).

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

[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) Tripartite graph (original weight) (b) Tripartite graph (normalized weights) (c) Bipartite projection (for u_2)

Fig. 1. Example of a tripartite graph and the random walk bipartite projection.

listening habits (weighted by the play count). This is similar to the work conducted in [30]. 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 [23] 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) $D_\alpha(p)$ [11, 14]. 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:

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

Interestingly, depending on the values of α, this diversity index expresses well-known diversity measures. $\alpha = 0$ is the richness diversity (which captures the *variety*), $\alpha \to 1$ is the exponential of the Shannon entropy, $\alpha = 2$ is the Herfindal diversity and $\alpha = \infty$ is the Berger diversity (which captures the *balance*). In the rest of the paper, we will study the α-diversity for $\alpha = 0, 2$ and ∞.

3.2 Collaborative filtering

The model. In addition to the set U and I, let $\mathcal{R} = \{r_{ui} \mid u \text{ rated } i\}$ 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 [13]. 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 $x = (x_{u_1} \dots x_{u_n})$ (resp. $y = (y_{i_1} \dots y_{i_m})$) 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 Equation 5 is optimized with a *Regularized Alternative Least Squares* method as described in [13]. However, each least square problem (which is equivalent to solving a linear system) is solved via *Conjugate Gradient Descent* [29].

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 $\mathcal{R}_{\text{test}}$ containing listening selections for users used to train the model, therefore all having a user latent factor. This eliminates 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 $L_u = (i_1, \dots, i_{|L|})$, ordered by their score $\widehat{p}_{ui}$, we define the *Discounted Cumulative Gain* (DCG) in Equation 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).

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

4 ANALYSING THE EFFECT OF COLLABORATIVE FILTERING IN TERMS OF DIVERSITY

This section presents the results we obtained by using the approach presented in Section 3 in the context of a dataset recording user activity on an online platform featuring musical content. We first present the general setting (Section 4.1), before investigating different questions: how diversified are the recommendations and what is the impact of the parameters (Section 4.2)? What is the effect of the recommendations on users' diversity (Section 4.3)? How to explain the differences in the way recommendations affects users' diversity (Section 4.4)?

	number of users	number of songs	number of tags	number of user–item links
full dataset	100, 000	169, 730	1, 000	3, 760, 629
training set	100, 000	167, 671	1, 000	3, 574, 692
test set	10, 000	54, 157	1, 000	185, 937

Table 1. Statistics of the dataset, the training set and the test set

4.1 Experimental setting

Dataset. The dataset we used comes from the *Million Song Dataset* project [2]. To create the first two layers of the triparite graph and the user–item links, we used the *Echo Nest user tast 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. The relevant statistics of the dataset are summarized in Table 1 (with the size of the training set and test sets).

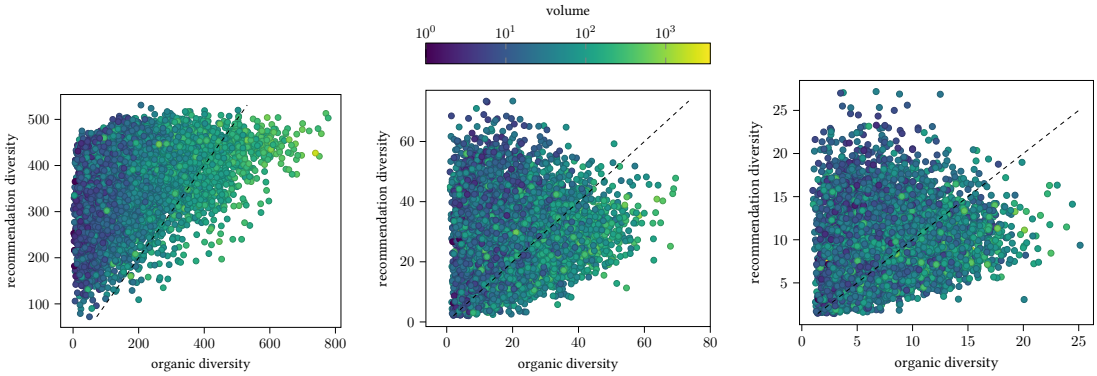
Implementation. We ran all of the experiments using the Python programming language on a 40 cores `Intel Xeon` server, equipped with 256 GB of RAM. We used the package `lenskit` [5] to instantiate, train and evaluate the recommender system. The code is available on GitHub[5] along with instructions to install and run it. Following the test procedure described in Section 3.2, the hyper-parameters were chosen to empirically maximize the *Normalized Disconted Cumulative Gain*, which resulted in 3000 latent factors, $\mu = 40$ and $\lambda = 10^6$. Unless stated otherwise, those are the values used in the rest of the paper.

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.

4.2 Analysis of the recommendations

Diversity of the recommendations. First, we investigate the properties of the recommendations in terms of diversity. Figure 2 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

[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]blinded, see code attached to the submission.
[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.



(a) $\alpha = 0$. (b) $\alpha = 2$. (c) $\alpha = +\infty$.

Fig. 2. Diversity of the recommendations ($k = 50$) with respect to the organic diversity of the users

is no strong relation between organic diversity and recommendation diversity since low organic diversity does not necessarily lead to low recommendation diversity. Some users 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 the recommendations tend to be more diverse than the organic ones[7]. This is particularly true for the variety (or coverage) captured by $\alpha = 0$, but it can also be observed (although to a lesser extent) for $\alpha = 2$ and $\alpha = \infty$.

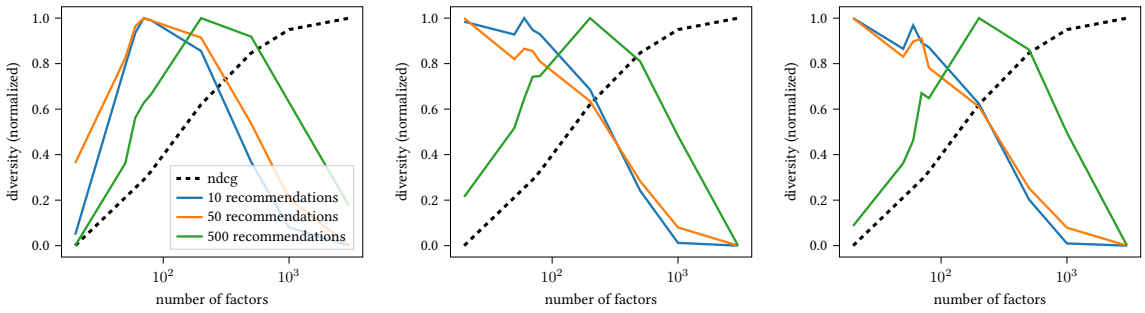
However, one can clearly see that this relation depends on the value of the organic diversity. As this value increases (for all values of α), it becomes more difficult for the recommendations to reach the same level of diversity.

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 [12]. To investigate this relation, in Figure 3 we show how the number of latent factors impacts the diversity 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 high number of recommendations ($k = 500$) or for $\alpha = 0$. For the latter case, there does seem to be a trade-off between performance and diversity. This supports the observations made in [12] 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.

4.3 Effect of the recommendations on users' diversity

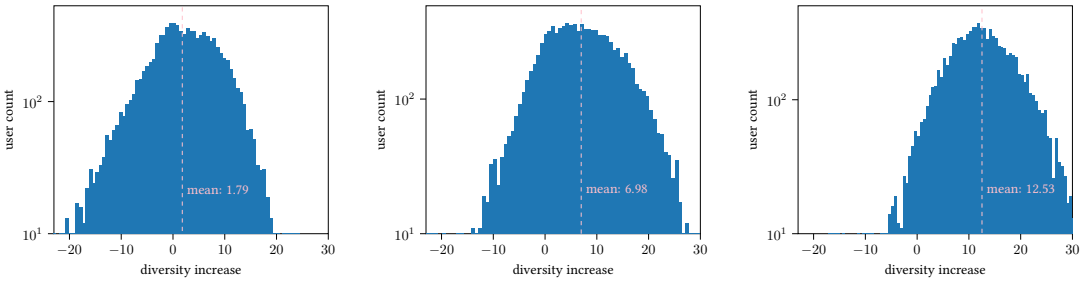
Figure 2 presented in Section 4.2 reveals that a list of recommendations can be diversified, even for users whose past musical records are narrow in scope. However, this plot does not provide any information on the impact of the recommendations on users' musical habits. How diverse are the user's listening habits *after being exposed* to the

[7]This observation stands as soon as a minimum number of items are recommended.
[8]The values have been normalized in order to ease the comparison.



(a) 0-diversity. (b) 2-diversity. (c) ∞-diversity.

Fig. 3. Diversity[8] and performance of the recommendations



(a) $k = 10$ recommendations. (b) $k = 50$ recommendations. (c) $k = 500$ recommendations.

Fig. 4. Distribution of the diversity increase ($\alpha = 2$) for different numbers of recommendations.

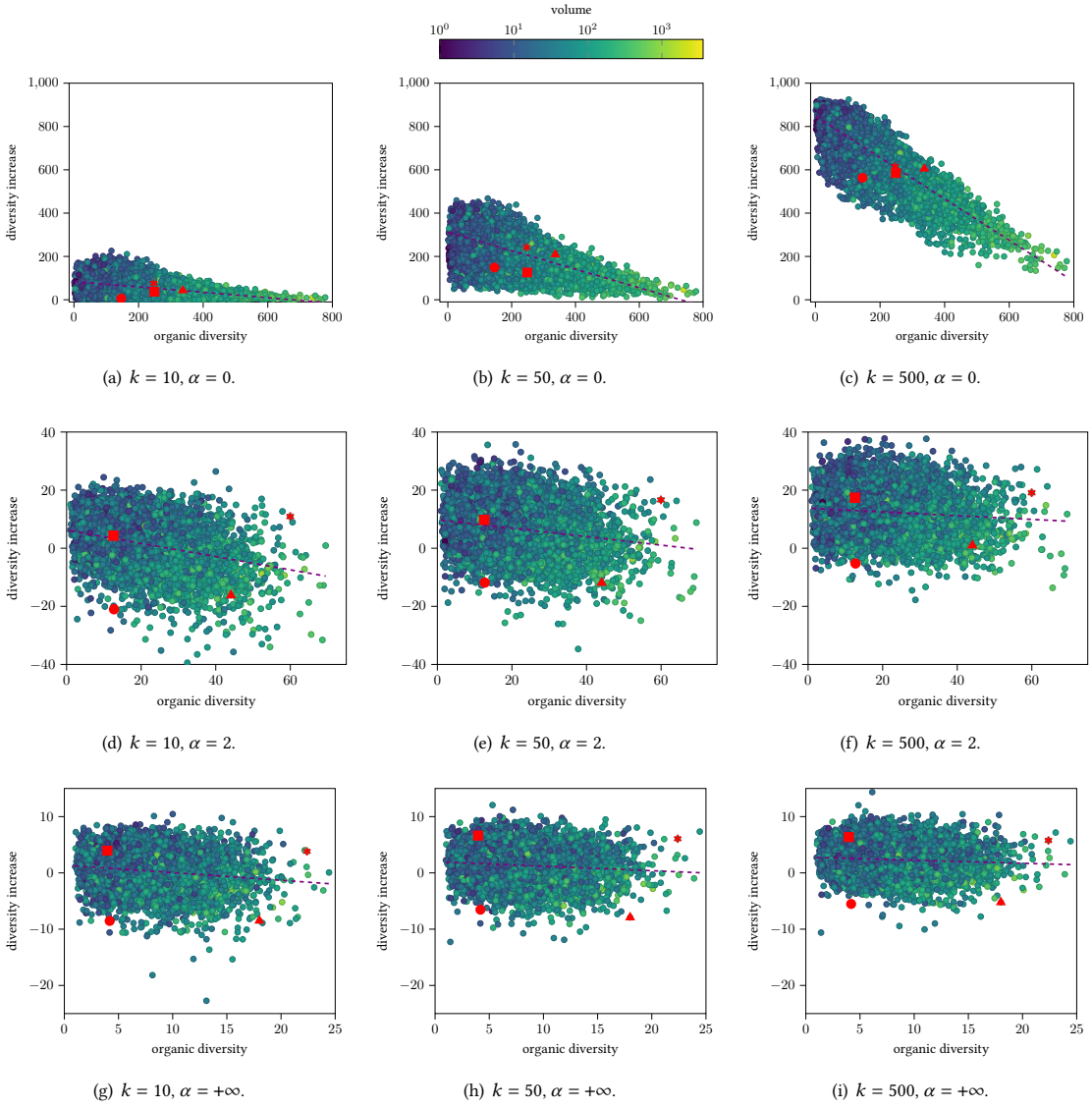
recommendations? To investigate this question, for each user we computed the *diversity increase*, which is defined as

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

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 [21] that account for the saturation effects in the diversity of users' attention.

Figure 4 presents the distribution of the diversity increase (with $\alpha = 2$) for different numbers of recommended items ($k = 10, 50, 500$). Clearly, 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 61% of users. However, the log-scale of the



(a) $k = 10, \alpha = 0$. (b) $k = 50, \alpha = 0$. (c) $k = 500, \alpha = 0$.

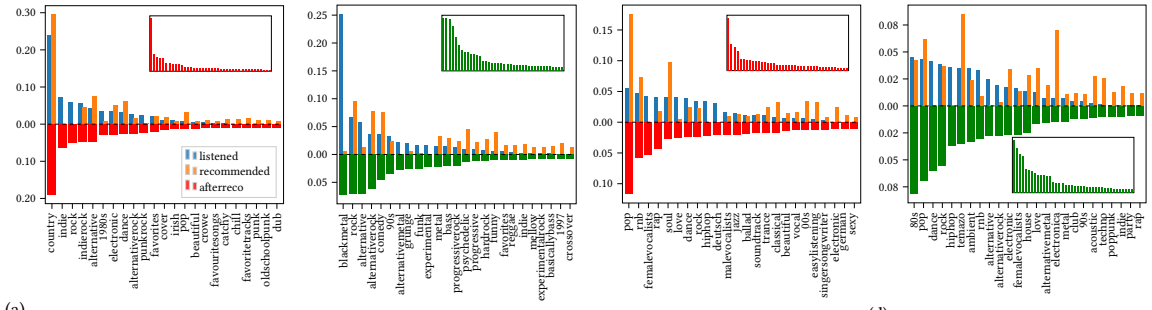
(d) $k = 10, \alpha = 2$. (e) $k = 50, \alpha = 2$. (f) $k = 500, \alpha = 2$.

(g) $k = 10, \alpha = +\infty$. (h) $k = 50, \alpha = +\infty$. (i) $k = 500, \alpha = +\infty$.

Fig. 5. 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 Section 4.4.

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 at least one order of magnitude lower than the number of items recommended.

To achieve a more detailed and comprehensive picture of the effect recommendations have on users' diversity, in Figure 5 we have plotted the diversity increase for different numbers of recommendations ($k = 10, 50, 500$) and different



(a) User E6CD (circle) Low organic diversity Decrease

(b) User C0D9 (square) Low organic diversity Increase

(c) User 1653 (triangle) High organic divsersity Decrease

(d) User 7671 (star) High organic diversity Increase

Fig. 6. Effect of recommendations on different users.

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 (dark blue on the plots) or whose musical listening selections are very narrow in scope (left-hand side of 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.

Although they are not conclusive, these results indicate that collaborative filtering impacts the diversity of musical listening habits in a specific way. To study this relation in greater depth, in the next Section we will investigate exactly how the recommendations affect users' diversity.

4.4 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, and how, as a result, recommendations have different effects on users' attention. In Figure 6, we present four users whose past musical records are either narrow (left and middle left) or diverse (middle right and right) and for whom recommendations ($k = 50$) generate either a decrease (left and middle right) or an increase (middle left and right) in diversity ($\alpha = 2$). For each 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 [9]. Finally, users presented in this picture are also visible in Figure 5 with their respective red marker.

This figure reveals some interesting aspects in terms of how recommendations impact users. First, in all four cases, one can observe that a significant fraction of the musical categories associated with the recommendations are completely new to the users, thus increasing variety. 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 leads to 263.51 new categories. This more than triples 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 E6CD (left), for instance, the most recommended category, *country*, 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 country music.

By contrast, the musical exposure of user C0D9 (middle left) is diversified as a result of the recommendations, although he/she also has one particular musical focus: *black-metal*. One reason for this is that black metal songs barely appear in the list of recommendations. Instead, in addition to new ones, the recommendations expose other, less dominant categories in the user's past musical record (such as *rock*, *alternative* and *alternative-rock*). 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. The study of users 1653 and 7671 (middle right and right) show this is not the case. Both are highly active users (in the top 3% of the most active users on the dataset) with a very high organic diversity (top 1%). Yet, they both experience a completely different impact of the recommendations exposure: while the main category of user 1653 (*pop*) features strongly in the recommendations, thus unbalancing the range of musical categories to which he/she is exposed, the musical record of user 7671, on the contrary, is broadened by the recommendations due to the more nuanced musical exposure.

5 CONCLUSION & 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 (Figure 2) and globally improve the diversity of most users (Figure 4). However, there are some limits to this capacity to diversify users' musical habits. First, it usually undermines the performance of the models (Figure 3). 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 (Figure 5). 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 (Figure 6). 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.

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

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 working on recommender systems could examine the ethical effects of algorithmic recommendations. This approach calls for future investigations, two of which are discussed below.

Individual vs. *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 [3].

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 '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 $User \rightarrow Item \rightarrow User \rightarrow Item \rightarrow Category$. 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 [28], in a unified framework.

REFERENCES

[1] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, and Sepideh Mahabadi. 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, NY, USA, 1–12. https://doi.org/10.1145/2488388.2488390

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[3] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/3287560.3287601

[4] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to Recommend Accurate and Diverse Items. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 183–192. https://doi.org/10.1145/3038912.3052585

[5] Michael D. Ekstrand. 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, NY, USA, 2999–3006. https://doi.org/10.1145/3340531.3412778

[6] Corrado Gini. 1921. Measurement of Inequality of Incomes. *The Economic Journal* 31, 121 (1921), 124–126. https://doi.org/10.2307/2223319

[7] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* 35, 12 (Dec. 1992), 61–70. https://doi.org/10.1145/138859.138867

[8] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and Sequential User Embeddings for Large-Scale Music Recommendation. In *Fourteenth ACM Conference on Recommender Systems*. ACM, Virtual Event Brazil, 53–62. https://doi.org/10.1145/3383313.3412248

[9] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (Sept. 2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

[10] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure Diversity as a Design Principle for Recommender Systems. *Information, Communication & Society* 21, 2 (Feb. 2018), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

[11] M. O. Hill. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* 54, 2 (1973), 427–432. https://doi.org/10.2307/1934352

[12] David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. 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, NY, USA, 75–76. https://doi.org/10.1145/3391403.3399532

[13] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, Pisa, Italy, 263–272. https://doi.org/10.1109/ICDM.2008.22

[14] Lou Jost. 2006. Entropy and Diversity. *Oikos* 113, 2 (2006), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

[15] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News Recommender Systems – Survey and Roads Ahead. *Information Processing & Management* 54, 6 (Nov. 2018), 1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

[16] John Paul Kelly and Derek Bridge. 2006. Enhancing the Diversity of Conversational Collaborative Recommendations: A Comparison. *Artificial Intelligence Review* 25, 1-2 (2006), 79–95. https://doi.org/10.1007/s10462-007-9023-8

[17] Chang Li, Haoyun Feng, and Maarten de Rijke. 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, NY, USA, 33–42. https://doi.org/10.1145/3383313.3412245

[18] Kevin Shear McCann. 2000. The Diversity–Stability Debate. *Nature* 405, 6783 (May 2000), 228–233. https://doi.org/10.1038/35012234

[19] E. Pariser. 2011. *The Filter Bubble: What The Internet Is Hiding From You.* Penguin Books Limited.

[20] Bibek Paudel, Thilo Haas, and Abraham Bernstein. 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, NY, USA, 215–223. https://doi.org/10.1145/3109859.3109916

[21] Rémy Poulain and Fabien Tarissan. 2020. Investigating the Lack of Diversity in User Behavior: The Case of Musical Content on Online Platforms. *Information Processing & Management* 57, 2 (March 2020), 102169. https://doi.org/10.1016/j.ipm.2019.102169

[22] Bruno Pradel, Savaneary Sean, Julien Delporte, Sébastien Guérif, Céline Rouveirol, Nicolas Usunier, Françoise Fogelman-Soulié, and Frédéric Dufau-Joel. 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* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 377–385. https://doi.org/10.1145/2020408.2020470

[23] Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring Diversity in Heterogeneous Information Networks. *Theoretical Computer Science* 859 (March 2021), 80–115. https://doi.org/10.1016/j.tcs.2021.01.013

[24] Alfréd Rényi 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, 547–561.

[25] Markus Schedl, Peter Knees, and Fabien Gouyon. 2017. New paths in music recommender systems research. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 392–393.

[26] Lei Shi. 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, NY, USA, 57–64. https://doi.org/10.1145/2507157.2507165

[27] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. 2019. How Good Your Recommender System Is? A Survey on Evaluations in Recommendation. *International Journal of Machine Learning and Cybernetics* 10, 5 (May 2019), 813–831. https://doi.org/10.1007/s13042-017-0762-9

[28] Andy Stirling. 2007. A General Framework for Analysing Diversity in Science, Technology and Society. *Journal of The Royal Society Interface* 4, 15 (Aug. 2007), 707–719. https://doi.org/10.1098/rsif.2007.0213

[29] Gábor Takács, István Pilászy, and Domonkos Tikk. 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, Illinois, USA, 297. https://doi.org/10.1145/2043932.2043987

[30] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 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, NY, USA, 209–216. https://doi.org/10.1145/2645710.2645743

[31] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 109–116.

[32] Isaac Waller and Ashton Anderson. 2019. Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1954–1964. https://doi.org/10.1145/3308558.3313729

[33] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 13–22.