Investigating the lack of diversity in user behavior: The case of musical content on online platforms

Rémy Poulain, Fabien Tarissan

Investigating the Lack of Diversity in User Behavior: the Case of Musical Content on Online Platforms

Rémy Poulain[a,*], Fabien Tarissan[b]

[a] *Sorbonne Université, CNRS, LIP6, F-75005, Paris, France*
[b] *Université Paris-Saclay, CNRS, ISP, ENS Paris-Saclay, Cachan, France*

Abstract

Whether to deal with issues related to information ranking (*e.g.* search engines) or content recommendation (on social networks, for instance), algorithms are at the core of processes that select which information is made visible. Such algorithmic choices have a strong impact on users' activity *de facto*, and therefore on their access to information. This raises the question of how to measure the quality of the choices algorithms make and their impact on users. As a first step in that direction, this paper presents a framework with which to analyze the diversity of information accessed by users in the context of musical content.

The approach adopted centers on the representation of user activity through a tripartite graph that maps users to products and products to categories. In turn, conducting random walks in this structure makes it possible to analyze how categories catch users' attention and how this attention is distributed. Building upon this distribution, we propose a new index referred to as the (calibrated) herfindahl diversity, which is aimed at quantifying the extent to which this distribution is diverse and representative of existing categories.

To the best of our knowledge, this paper is the first to connect the output of random walks on graphs with diversity indexes. We demonstrate the benefit of such an approach by applying our index to two datasets that record user activity on online platforms involving musical content. The results are threefold. First, we show that our index can discriminate between different user behaviors. Second, we shed some light on a saturation phenomenon in the diversity of users' attention. Finally, we show that the lack of diversity observed in the datasets derives from exogenous factors related to the heterogeneous popularity of music styles, as opposed to internal factors such as recurrent user behaviors.

Keywords: Network analysis, Diversity, Tripartite graph, Online platform, Music, Random model

*corresponding author.
Email addresses: remy.poulain@lip6.fr (Rémy Poulain), fabien.tarissan@ens-paris-saclay.fr (Fabien Tarissan)

1. Introduction

Online networks and digital platforms have become increasingly essential in our everyday lives. Not only do they shape our interactions in the real world, but they also constrain our actions in virtual spaces enabled by the internet and the web. At a time when data are systematically processed by online platforms, the traces left by users contain key information revealing their behavior and tastes, which lead most digital platforms to propose algorithmic recommendations to their users.

At the same time, the existence of algorithmic recommendations feeds into prophecies about the expected positive outcomes of their use [1] as well as fears about the evolution of an algorithmically controlled society [2]. Can algorithms really be objective and neutral [3]? What impact do they have on the users exposed to their output [4]? Do they shape and narrow our vision of what information is available online [5]? Do they generate any kind of discriminations when applied to large-scale infrastructures [6, 7]?

These issues are all at the core of intense debates and relate to many different contexts ranging from the exploitation of private data [8] to the impact of search engines on elections [9, 10], the dissemination of fake news [11], and filter bubble phenomena on social media [12]. Consequently, this increasing use of digital platforms and their recommendation systems has led the scientific community to focus on the impact of algorithmic decisions on users' behavior [2, 13, 14, 15, 16].

However, while the need to address these issues is commonly agreed upon [17, 18, 19, 20], there is no consensus within the scientific community around the question of *how* to measure the impact of recommendation systems. While it is clear that an efficient recommendation system requires serendipity and diversity [21], how might such properties be quantified?

The importance of this question is reflected in the emergence of intense scientific activity dedicated to analyzing the *diversity* of information proposed to users [15, 22], particularly in the context of music recommendation systems [23, 24, 25, 26, 27, 28, 29] which is the context of our validation settings. Indeed, whether to provide purchasing recommendations (*i.e.*, the suggestion to purchase an item on *Amazon*) or information recommendations (*i.e.*, the suggestion to read a post in *Newsfeed* or listen to a song on *Spotify*), algorithms strongly affect what is made visible to users. One might wonder, in turn, whether the choices made by platforms to make certain information visible are representative of the diversity of existing information.

This paper is intended to generate significant progress in that regard by providing a framework with which to formally analyze online users' behavior. By quantifying the extent to which the information accessed by a user is *diverse*, we aim to provide a way of measuring how narrowed and/or biased their behavior is. In turn, this diversity provides key information for studying the impact of algorithmic recommendations and comparing the effect of different algorithmic choices.

Research objective and findings. The objective of this paper is to propose a general and formal framework to analyze the diversity of user behavior.

To this end, we propose to exploit the network structure generated by user activity in order to reveal the diversity of information accessed by such users. More precisely, we will represent user activity as a tripartite graph that maps users to products and products to categories. In turn, conducting random walks in this structure will make it possible to analyze how categories catch users' attention and how this attention is distributed. Building upon this distribution, we propose a new index referred to as the *(calibrated) herfindahl diversity* aimed at quantifying this diversity.

To determine the relevance of the proposed approach, we apply it to two datasets related to online (music) platforms and analyze the results. The study reveals, in particular, that:

- The index makes it possible to **differentiate** between different behaviors and **identify the diversity** of users that listen to contrasting music styles (see Section 5 for details);
- The index **highlights a saturation process** in the diversity of users' attention (see Section 6 for details);
- The lack of diversity as captured by the index **derives from exogenous factors** related to the heterogeneous popularity of music styles, as opposed to internal factors such as recurrent user behavior (see Section 8 for details).

Outline of the paper. The paper is organized as follows. After reviewing the standard techniques used in literature to study diversity (Section 2) , we describe the formalism used to represent a user's activity and present our score measuring the diversity of such activity (Section 3). Then, we present the information contained in the two datasets considered in this study in order to validate the approach (Section 4) and show how our index can be used to analyze the diversity both of music audiences (Section 5) and of users' attention (Section 6). To consolidate our results, we investigate two questions raised by our approach in greater depth: how might the different facets of diversity be captured (Section 7) and how might random models be exploited in order to explain the analyses (Section 8). Finally, we conclude the paper and pave the way for further work (Section 9)[1].

[1] Compared to the preliminary results presented in [30], this paper investigates the relevance of our approach on a second dataset involving the music industry (see in particular Sections 4.2, and 5.3 as well as part of the analyses made in Section 6) and explores two aspects that were left as a perspective (Sections 7 and 8). The last section in particular elaborates on the reasons explaining the lack of diversity observed in users' attention.

2. Related work

Many papers have focused on diversity in various complex systems in order to highlight its importance for ensuring the viability of systems in the long term, with topics ranging from ecology [31] to politics [32], science [33, 34] and social media [35, 36], to cite just a few.

Likewise, a large number of propositions have been put forward to formally capture the notion of diversity by defining various metrics [33, 37]. Among these, one might cite, in particular, the Shannon [38] and Renyi [39] entropy, the Gini coefficient [40], the Hirschman-Herfindahl index [41, 42], and the Berger-Parker index [43]. The profusion of indexes reveals the difficulty of reaching a consensus on the correct way to capture this notion. This paper does not seek to offer such a consensus but rather to propose a new way to build on these different indexes in order to capture different facets of diversity (see Section 7, in particular).

The specific context of the case study examined in this paper – namely, musical content – is no exception to the rule [44, 45, 46, 47, 48, 49]. However, studies have long focused mainly on the economic aspect, seeking to identify the key actors who monopolize the markets (thereby limiting its diversity). It is only recently, due to the extensive use of recommendation systems by digital platforms, that scientific papers related to diversity in online music platforms have emerged [23, 24, 25, 26, 27, 28, 29]. For the most part, such publications examine the diversity of the items recommended to a user by analyzing the *distance* between all pairwise items recommended. As highlighted in various papers (such as [23]), no consensus has emerged on how distance should be computed. However, such studies generally use a particular function applied to feature vectors that define the items, such as cosine similarity [50], the inverse Pearson correlation [51], and hamming distance [52].

In that regard, and even if there is a direct relation between this line of research and our study, we adopt a different vantage point in this paper in that we focus on *user behavior* [53] rather than on the distance between the items. As such, what matters in our measure of diversity relates closely to *how a user accesses* the set of items he/she purchases (here, the songs listened to). One implication of this shift in the standard approach lies in the crucial role of the paths (depicted in a tripartite graph here: see Section 3.1) that relate a user to the features of the songs. Such a dynamic cannot be examined by reducing the analyses to the computation of distance measures. This is why we propose, instead, to rely on random walks that make it possible to capture the relation between the users and the features of the songs they listen to with great precision.

In that sense, and to the best of our knowledge, this study is the first to propose connecting diversity indexes to the output of a random walk on (tripartite) graphs in order to analyze the diversity of user behavior. In light of this, we would position our study in line with the long trend of studies in the scientific community that propose formal, sound ways to quantify diversity in complex systems.

Another way to consider the importance of the approach proposed and tested

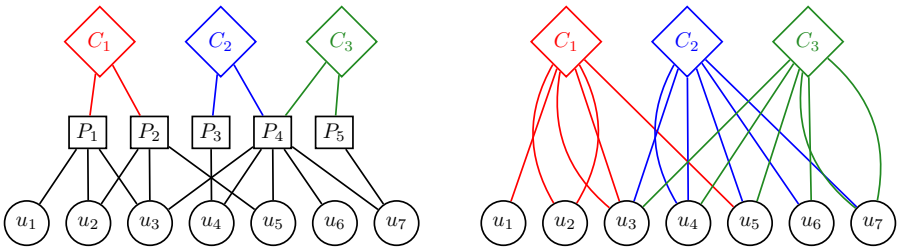


Figure 1: Example of a tripartite structure (left) and its bipartite projection (right).

in this paper is to recognize that all standard analyses using distance measures are particular cases of the formal approach proposed here, using graphs. If a tripartite graph is projected into a bipartite representation (see Section 3.1 and Figure 1 (right), in particular) relating the bottom layer (users) to the top layer (features), a random walk becomes insignificant and does not convey any additional information compared to standard indexes.

We argue, however, that not only do we lose a great deal of structural information in such a projection (as already demonstrated [54]) but we also lose the ability to precisely scrutinize the dynamics (the paths) that link a user to the features of what he/she listens to.

3. A formalism to analyse diversity

In this section, we provide all necessary definitions for conducting a formal analysis of user activity. We begin by defining tripartite graphs (Section 3.1), before then proposing the herfindahl diversity index (Section 3.2).

3.1. Tripartite graph

A bipartite graph is a graph with two disjoint sets of nodes and such that links relate a node in one set to a node in the other set. Formally, it is defined by a triplet $\mathbb{B} = (\top, \bot, E)$ where $\bot$ is the set of *bottom* nodes (*e.g.* users), $\top$ is set of *top* nodes (*e.g.* songs), and $E \subseteq \top \times \bot$ the set of links (*e.g.* that relate the users to the songs they have listen to). For each node $u \in \top$, one defines the set of its *neighbors* $N(u) = \{v \in \bot \mid (u, v) \in E\}$ and a similar definition is derived for $v \in \bot$. We refer to the size of the set of neighbors as the *degree* of the node: $d(u) = |N(u)|$.

Besides, both in the context of purchasing recommendation or news recommendation, the $\top$ nodes can be mapped to their *categories*. Products can for instance be related to their type (books, vehicles, tools, ...) and news can be related to their thematic (international, sport, fashion, ...). This leads to a second bipartite graph mapping products to categories.

Thus, in order to analyze the complete structure of users' activity, we propose to describe it as a *tripartite graph* $\mathbb{T} = (\top, \vdash, \bot, E^{\top}_{\mathbb{T}}, E^{\bot}_{\mathbb{T}})$ where $\top$ stands for

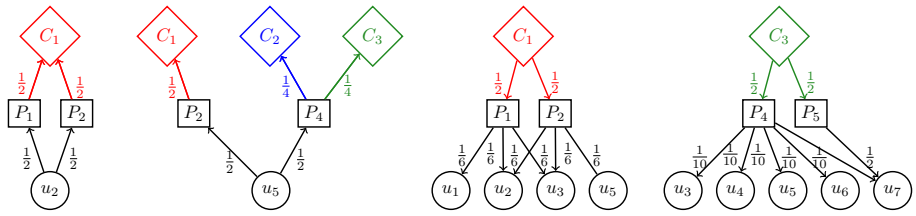


Figure 2: Random walks from different nodes of the tripartite graph of Figure 1: from u_2 (left), u_5 (middle left), C_1 (middle right) and C_3 (right).

the categories, $\vdash$ for the products, $\perp$ for the users, $E_{\mathbb{T}}^{\top} \subseteq \top \times \vdash$ for the relations between categories and products and $E_{\mathbb{T}}^{\perp} \subseteq \vdash \times \perp$ for the relations between products and users (see Figure 1 left for an example).

In addition, information related to the weight of bottom links (number of times a product is bought, an article read, a song listened to, etc.) can be taken into account by defining a weight function $w_{E_{\mathbb{T}}^{\perp}} : E_{\mathbb{T}}^{\perp} \mapsto \mathbb{R}$. In that case, in addition to the degree of a node $v \in \perp$, we also use the *weighted degree* of $v \in \perp$, defined by:

$$d_w(v) = \sum_{u \in N(v)} w(u, v)$$

From this tripartite graph, it is then possible to study how the categories are related to users' activity by analyzing the bipartite projection of $\mathbb{T}$. Formally, this projection is defined by the bipartite graph $Pr(\mathbb{T}) = (\top, \perp, E_{Pr(\mathbb{T})})$ where $E_{Pr(\mathbb{T})} = \{(u, v) \in \top \times \perp \mid \exists z \in \vdash \ s.t.\ (u, z) \in E_{\mathbb{T}}^{\top} \, and \, (z, v) \in E_{\mathbb{T}}^{\perp}\}$. Figure 1 right is an example of the result of such a projection.

In case the tripartite graph is weighted, the projection derives a weight function $w_{E_{Pr(\mathbb{T})}} : E_{Pr(\mathbb{T})} \mapsto \mathbb{R}$ defined formally by:

$$w_{E_{Pr(\mathbb{T})}}(u, v) = \sum_{z \in N(u) \cap N(v)} w_{E_{\mathbb{T}}^{\perp}}(z, v)$$

3.2. Diversity score

Once they have been depicted in the form of a tripartite graph, we intend to analyze how the induced relations between bottom and top nodes (here, between users and categories) are distributed. To do so, we rely on random walks on the structure. Starting with a user u, we compute the distribution of the probabilities to reach the different categories, through the products linked to u.

The aim is then to distinguish between a perfect situation in which all relations are uniformly distributed across all categories (highest diversity) and the worst situation in which few categories capture all the links (lowest diversity). To this end, we make use of the *Herfindahl–Hirschman index* [41], which is widely used in economics to study market concentration and, in particular, to identify monopoly situations.

We have adapted the original definition to our context. Formally, for each node $u \in \bot$ and for each node $v \in \vdash$, we can define the probability to reach v from u as $p_{u\to v} = \frac{w(u,v)}{d_w(u)}$. Similarly for every $v \in \vdash$ and $z \in \top$, one can define $p_{v\to z} = \frac{w(v,z)}{d_w(v)}$. By extention, one can then derive the probability to reach z in $\top$ from u in $\bot$ as:

$$p_{u\to z} = \sum_{v\in\vdash} p_{u\to v} p_{v\to z}$$

It is worth noting here that if $v \notin N(u)$, then $p_{u\to v} = 0$. This means the last definition can be usefully replaced by:

$$p_{u\to z} = \sum_{v\in N(u)} p_{u\to v} p_{v\to z}$$

For each node $u \in \bot$ one can now compute such probability $p_{u\to z}$ for every node $z \in \top$. This allows us to define the output of a random walk issued from u as the distribution of probabilities for every $\top$ node. Such a distribution then makes it possible to apply any diversity index, such as the Herfindahl–Hirschman index.

Formally, let RandWalk($\mathbb{T}$, u) denote a random walk issued from $u \in \bot$ in $\mathbb{T}$ and let P be a distribution of probabilities generated by the random walk, *i.e.* $P = \text{RandWalk}(\mathbb{T}, u) = (p_{u\to z})_{z\in\top}$. Then, the *herfindahl diversity* of node u in $\mathbb{T}$ is defined by:

$$\texttt{hd}(\mathbb{T}, u) = \left(\sum_{z\in\top} p_{u\to z}^2\right)^{-1}$$

A high value of `hd` thus indicates that the categories of the products related to a particular user are almost uniformly distributed, while a low value indicates a concentration of its products towards a small number of categories.

It is worth noting that, for each user u, the herfindahl diversity is formally bounded by the number of categories associated with the random walk, that is, the degree of u in the bipartite projection. This upper bound is reached when the distribution is uniform. Thus, the total number of $\top$ nodes is an upper bound for any herfindahl diversity.

Returning to the example of Figure 1, one can see that this coefficient makes it possible to differentiate between users u_2 and u_5. Although both are related to two products, their situations differ entirely. While u_2 only accesses products attached to category C_1, thus giving it the lowest diversity value $\texttt{hd}(u_2) = 1$, u_5 is, on the contrary, related to all three possible categories through a relatively balanced distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. Its herfindahl diversity is, then, $\texttt{hd}(u_5) = \frac{8}{3}$, which is clearly higher than $\texttt{hd}(u_2)$ and close, in particular, to the highest value in this example (3, the total number of categories). This is depicted in Figure 2 (left and middle left).

This approach focuses on a random walk starting with a user (*i.e.*, a $\bot$ node), which allows us to study the diversity of users' *attention*. Similarly, the herfindahl diversity can be computed based on a random walk starting from

200 a $\top$ node, which then makes it possible to study the diversity of the *audience* captured by a particular category.

Applying this to the example of Figure 1, one might note that category C_1 exhibits a higher diversity ($\texttt{hd}(C_1) = \frac{18}{5}$) than category C_3 ($\texttt{hd}(C_3) = \frac{10}{4}$). This is because the distribution of probability is closer to a uniform distribution for C_1 than it is for C_3: See Figure 2 (middle right and right) for a visual comparison.

To conclude this section dedicated to the formalism, let us remak that if one wants to be completely formal, one could define a general herfindahl diversity by adding the notion of *path* in the tripartite structure. This would lead to formally take into account the sequence of layers that a random walk can use to derive the distribution of probabilities. In the two examples taken above, we considered the paths $\mathbb{P} = [\bot, \vdash, \top]$ (diversity of users attention) and $\mathbb{P} = [\top, \vdash, \bot]$ (diversity of categories audience) and defined the herfindahl diversity as $\texttt{hd}$ ($\mathbb{T}$, u) instead of $\texttt{hd}(\mathbb{T}, \mathbb{P}, u)$.

Although such a definition makes it possible to conduct a more general analysis of the diveristy in a tripartite structure, we limit the paths, definitions and analyses to the two cases exemplified in this study – from users towards categories and from categories towards users – and omit the extensive notation.

4. Dataset

This section provides some details on the two datasets used in this study: namely, the *Million Song Dataset* (Section 4.1) and the *Amazon Dataset* (Section 4.2). For both cases, we describe the information contained in the metadata of the records and which pre-process operations we performed. We then provide some descriptive statistics (Section 4.3) relating to the tripartite graphs.

4.1. Million Song Dataset

The first dataset we used stems from the *Million Song Dataset* (*MSD*) project [55], which provides a collection of audio features and metadata related to user activity on online music platforms. Provided by *The Echo Nest* (now owned by *Spotify*), this project offers access to a *user taste profile* dataset[2]. It contains triplets (user, song, play count) that describe how many times a user has listened to a given song. The dataset contains approximately 48 million triplets, involving 1 million users and $300,000$ songs. This constitutes the bottom and middle layers of our tripartite structure.

To add the third layer (the categories), we also made use of the *last.fm* dataset[3] from which we extracted the tags associated with a song. For each

[2]available at `https://labrosa.ee.columbia.edu/millionsong/tasteprofile`
[3]available at `https://labrosa.ee.columbia.edu/millionsong/lastfm`

song, the dataset provides a list of tags intended to describe the music categories to which the song belongs. This comprises around $500,000$ songs and approximately the same amount of tags.

Since the two datasets were recorded separately, we performed the following operation to obtain a coherent tripartite graph. First, we mapped each song to its unique MSD identifier[4]. Then, we only kept information for songs that were present in both two datasets. Furthermore, since the uses of the tags were extremely contrasting[5], we focused on the most popular tags and retained only the $1,000$ most frequent tags. Finally, we removed all the songs with no tags and, consequently, all users with no songs.

This resulted in a tripartite graph comprising $1,019,190$ users $\perp$ nodes), $234,379$ songs ($\vdash$ nodes), and $1,000$ tags ($\top$ nodes).

4.2. Amazon Dataset

The second dataset used in this study is a record of all the reviews made between May 1996 and July 2014 on the *Amazon* online marketplace [56, 57]. To remain with the context of the music industry, we focused exclusively on items related to the categories *CDs & Vinyl* and *Digital Music*.

Similarly to the *MSD* case, we used two independent datasets in order to generate a tripartite graph. First, we used the *ratings dataset*[6] which contains a list of tupples (user, item, rating, timestamp) relating the users ($\perp$ nodes) to the products ($\vdash$) they are reviewing. This dataset provided information on approximately $500,000$ distinct users and $450,000$ songs.

Second, we used the *metadata dataset*[6], which provides extensive details on each product proposed on *Amazon*, including the price, other products also bought, other products also viewed, etc. From this dataset, we extracted the information that connects the products ($\vdash$) with their different categories, expressed as lists of tags. For instance, item I, related to the category "[*CD& Vinyl, Children's Music, Stories*]", generates two links: one between I and the tag *Children's Music*, and the other between item I and the tag *Stories*[7].

Finally, as with the previous case, when merging the information contained in the two datasets, we only kept the most popular tags and removed songs with no tags and users with no reviews.

This resulted in a tripartite graph comprising $465,248$ users ($\perp$ nodes), $445,514$ songs ($\vdash$ nodes), and 250 tags ($\top$ nodes). It is worth noting that compared to *MSD*, the number of songs is similar but the number of tags and users is significantly lower.

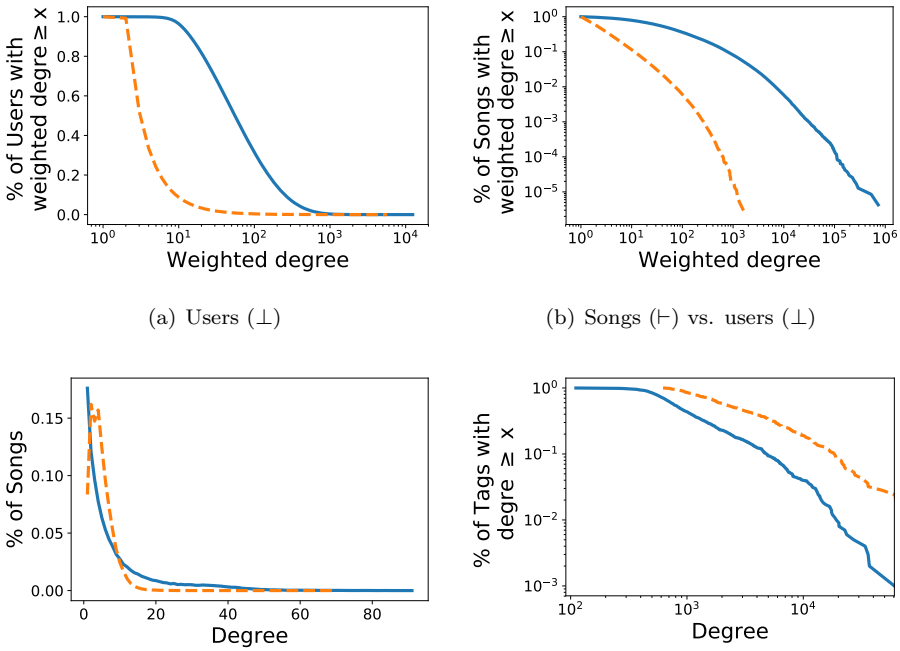
---

[4] we removed tracks known to be matched to wrong songs, see `https://labrosa.ee.columbia.edu/millionsong/blog/12-2-12-fixing-matching-errors`.

[5] some tags, such as *"rock"* or *"metal"*, clearly describe musical content but others, such as *"webfound" "polyglotism"*, are more problematic.

[6] see `http://jmcauley.ucsd.edu/data/amazon/`.

[7] We systematically omit tags *CD& Vinyl* and *Digital Music* since all selected categories involve at least one of the two tags.



(a) Users ($\perp$)

(b) Songs ($\vdash$) vs. users ($\perp$)

(c) Songs ($\vdash$) vs. tags ($\top$)

(d) Tags ($\top$)

Figure 3: Degree distributions in the tripartite graphs representing the Million Song Dataset (plain blue lines) and Amazon Dataset (dashed orange lines).

4.3. Characteristics

In addition to the information concerning size described above, Figure 3 shows the distribution of the links in the two tripartite graphs extracted from the datasets.

Figure 3(a) presents the inverse cumulative distribution of the weighted degree for the users for both cases (plain blue lines for *MSD*, dashed orange lines for *Amazon*). Although the complete range of values is high (the x-axis is in log scale), as is the maximal value ($12,387$ for *MSD*, $5,706$ for *Amazon*), the order of magnitude remains narrow for all users. A user recorded by *MSD* (only) listens to a song 105 times and listens to 37 different songs on average while they (only) write an average of 5 reviews on *Amazon*. These values are representative of a random user for each dataset and show that user behavior is quite homogeneous.

In contrast, Figure 3(b) shows that the popularity of the songs is very heterogeneous. It presents the distribution of the degrees for the songs towards the users (cumulative weighted degree distribution, in log-log scale) for the two datasets. Some songs are very popular (listened to more than $100,000$ times in *MSD* and reviewed more than $1,000$ times on *Amazon*) while the vast majority

of the songs have a small number of play counts (91% of the songs are listened to less than $1,000$ times on *MSD*, 64% less than 100 times) or of reviews (99% of the songs are reviewed by fewer than 100 users) and a small audience (78% of the songs are listened to by fewer than 100 different users and 88% are reviewed by fewer than 10 users).

Regarding the links towards the tags (Figure 3(c)), a large number of songs have a very small degree. In particular, 72% of the songs have fewer than 10 tags on *MSD* (93% on *Amazon*) from the $1,000$ that are possible (250 in Amazon). This is to be expected as the tags are intended to describe the content and feeling related to a song, which naturally results in a small number of possible combinations.

Finally, Figure 3(d) presents the inverse cumulative distribution of the tags towards the songs. The plot (in log-log scale) clearly exhibits a heterogeneous distribution of the use of the tags. Similarly to the play count of the song, some tags are extremely popular while the majority of the tags are used a small number of times by users.

All in all, the two datasets exhibit properties that are usually observed in similar systems. In particular, the popularity of the songs is highly heterogeneous and the behavior of a random user is regular.

Before analyzing the diversity, let us recall that the bipartite projection of the structure being studied allows us to study how users are related to tags. In the context of these datasets, we will use the term *volume* to refer to the weighted degree of the links in the projection. Thus, the *volume of tag* t is defined as the sum of the number of play counts (or reviews) for all songs with tag t. Similarly, the *volume of user* u is the sum of the number of tags for all songs listened to (or reviewed) by u, multiplied by their play count. When there is no ambiguity, we use the term 'volume' only.

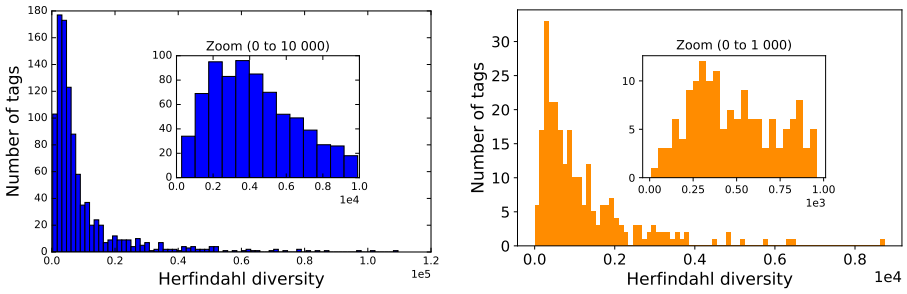
Let us turn to the results. We will begin by showing how the proposed herfindahl diversity can be used to study the diversity of the *audience* captured by a category (Section 5) before studying the diversity of users' *attention* (Section 6).

5. Diversity of the tag audience

Figure 4 shows the distribution of the diversity for all the tags. This distribution is very heterogeneous, with an average value of $9,699$ in *MSD* (median of $5,111$) and $1,197$ in *Amazon* (median of 806). However, it exhibits some tags with a particularly high diversity (higher than $100,000$ in *MSD* and close to $10,000$ in *Amazon*). This shows that tags may have very different audiences, ranging from a very broad audience to a very narrow one.

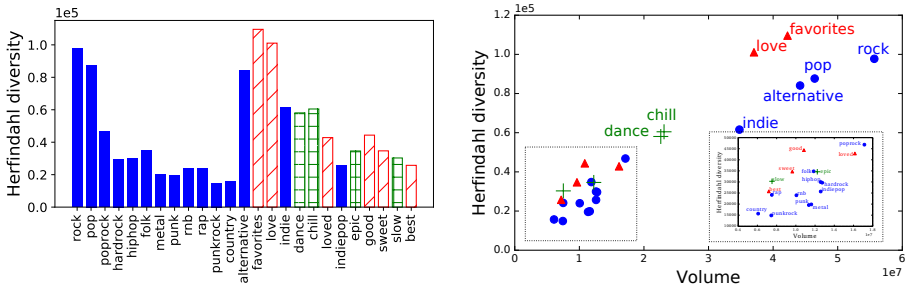
Concerning the latter (see insets in Figure 4), one can see, however, that tags with a diversity score lower than $10,000$ in *MSD* (which represents 75% of the tags) present a more homogeneous diversity value[8], centered around $5,000$.

[8]Note that as we zoom in, the size of the bins in the inset is not the same as the one in



(a) Million Song Dataset (b) Amazon Dataset

Figure 4: Distribution of the diversity of the tag audience for *MSD* (left) and *Amazon* (right).



(a) Diversity of tags audience (b) Herfindahl *vs.* volume

Figure 5: Analysis of the diversity of a selection of 25 tags from the *Million Song Dataset*

A similar observation can be made regarding the *Amazon Dataset* for tags with a diversity score lower than 1,000 (59% of the tags).

This diversity suggests that a more in-depth investigation should be carried out into how the score behaves according to the tags and, in particular, whether it can differentiate between different tag audiences. In order to investigate this question, and because the manual analysis of several thousand tags is not feasible here, we will now focus on a selection of 25 tags for each dataset.

5.1. Focusing on 25 tags for the Million Song Dataset

Even among the 1,000 most popular tags on the *Million Song Dataset*, different sorts of tags appear. This is because different users have very different ways of using tags to describe the content of a song. Some tags, for instance,

the larger scale.

are clearly used to describe the music style related to a song (such as 'rock', 'metal' or 'country'). We will refer to those tags as *style tags* and will represent them on Figures using plain blue markers.

In contrast, others tags relate less to the content of a song than to its context, for instance, the emotion experienced when listening ('awesome', 'best'), the period at which the song was created ('1986', '70s') or even why or when it is listened to ('to sleep', 'shower'). We will refer to those tags as *generic tags* from now on and will depict them using red triangles (or red dash lines).

Sometimes, a tag can also be a mix of the two aforementioned categories, mainly because of its polysemy. For instance, 'chill' clearly refers to an emotion but is also increasingly used to refer to a style of music. We will refer to those tags as *mix tags* hereafter and will use green cross markers to represent them.

This distinction is interesting since one can expect the herfindahl diversity to be able to differentiate between style and generic tags; as the former naturally leads to a narrower audience than the latter, its diversity is likely to be lower.

In order to investigate this question, we manually selected 25 of the 50 most popular tags: 15 are style tags, 6 are generic tags, and 4 are mix tags. The list is provided in Figure 5 with the herfindahl diversity of each tag.

Surprisingly, on Figure 5(a) one can see that highly diverse tags appear in the two extreme categories: 'rock', 'pop' and 'alternative' for style tags; 'favorites' and 'love' for generic ones. Similarly, tags with a low diversity can be observed in each category: 'country' and 'metal' for style tags; 'best' for generic ones; and 'slow' for mix ones.

This observation calls into question what is actually captured by the herfindahl diversity. It is interesting that tags like 'favorites' and 'best', for instance, have such a different diversity although they are likely to be synonymous in this context.

Considering those tags in particular, the diversity score of a tag proved to be strongly correlated to the volume of its audience (*i.e.*, to the number of times that songs in its category have been listened to, see Section 4.3). The larger its audience, the higher its diversity. This can be clearly observed in Figure 5(b). This raises the question of compensating for the effect of the volume so that the score perfectly captures the diversity instead of its volume. This will be the subject of the following section.

5.2. Calibrated herfindahl diversity

In network science, one common way to take into account the effect of a property on a score is to compare that score to its expected value for random networks with similar properties. This is usually carried out using models that generate random networks respecting the properties of interest.

In our context, we used a variant of the *configuration model* [58] which randomly assigns edges to match a given degree sequence without adding any other expected property. We used it to shuffle the bottom part of the tripartite structure in order to randomly reassign the links between the users and the songs.

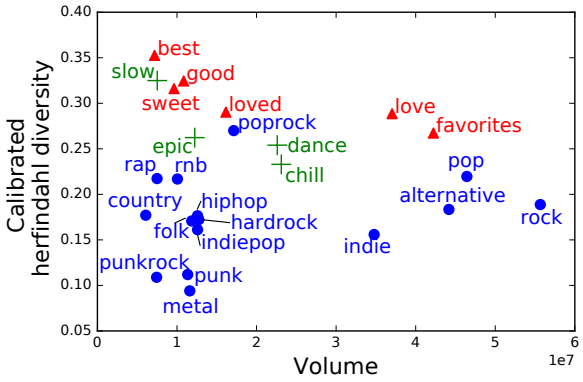


Figure 6: Calibrated diversity of tag audience according to its volume (*MSD*)

More precisely, we generated tripartite graphs with the same number of nodes and links, but such that the links of $E^{\perp}_{\mathbb{T}}$ (with their weight) are distributed uniformly at random among $\perp$ and $\vdash$ nodes according to their observed degree. The other links (*i.e.*, between $\vdash$ and $\top$) remain unchanged.

This means that, compared to the original, the songs on the tripartite graphs have the exact same tags and are listened to the same number of times but by random users. Similarly, every user listens to the same number of songs but those songs are selected uniformly at random.

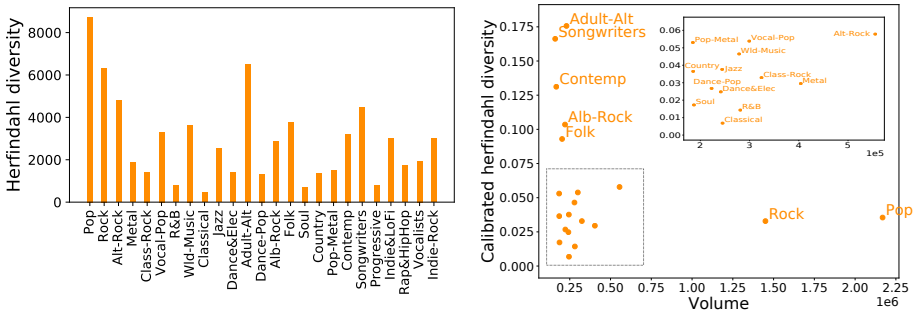
As such, we were able to generate several random tripartite graphs and compute the average herfindahl diversity for every tag. This average value was then used to divide the herfindahl diversity, giving the *calibrated herfindahl diversity*. Formally, let $\mathrm{Rand}(\mathbb{T})$ be the random tripartite graph generated from $\mathbb{T}$ by the model; the calibrated herfindahl diversity of u in $\mathbb{T}$ is defined by:

$$\texttt{chd}(\mathbb{T}, u) = \frac{\texttt{hd}(\mathbb{T}, u)}{\texttt{hd}(\mathrm{Rand}(\mathbb{T}), u)}$$

It is worth noting here that the volume of a tag is the same in both the original and the random graphs. This allows for a fair comparison between the two values, thus legitimating the definition of the calibrated herfindahl diversity.

The result is shown in Figure 6. As expected, the comparison with the random model compensates for the impact of the volume on the way diversity is computed. No particular correlation can be observed between volume and the calibrated herfindahl diversity. This does not rule out tags with a high volume from still having a relatively high diversity, since they have a better chance of reaching a broader audience. In particular, one can see that diversity increases slightly with volume.

Moreover, the calibrated diversity seems to restore the balance between tags that are very close on the semantic level yet very different in terms of use. The tags 'love' and 'loved', for instance, exhibit a similar diversity (around



(a) Diversity of tags audience

(b) Calibrated herfindahl *vs.* volume

Figure 7: Analysis of the diversity of a selection of 25 tags from the *Amazon Dataset*.

0.29) though they have a very different volume ('love' is almost 4 times more used than 'loved'). Similar observations can be drawn for the tags 'indie' and 'indiepop'.

In addition, the plot shows that the calibrated herfindahl diversity can differentiate between generic tags (upper part of the plot) and style tags (mostly in the bottom part): generic and mix tags all have a high diversity (between 0.23 and 0.35) while style tags tend to have a lower value (from 0.09 to 0.22).

The only exception is 'poprock', which has a relatively high diversity for a style tag (0.27) despite its low volume (17M total play counts). Although it is difficult to draw a conclusion from this value alone, one can note that, taken independently, the tags 'rock' and 'pop' also have a high diversity. This might explain why 'poprock' builds on their success and reaches a broader and more diverse audience.

Finally, the fact that the calibrated diversity is now independent of the volume allows for comparisons of types of music with similar volumes. For instance, one can see that, though they all have a similar audience size (around 10M total play counts), 'rap' and 'rnb' reach a far more diverse audience than 'metal' and 'punk', which seem to form rather dense and coherent sub-groups of listeners.

5.3. Focusing on 25 tags from the Amazon Dataset

Compared to the *Million Song Dataset*, the tags used in *Amazon* are selected by the platform itself, rather than the users. This results in a far more standardized meaning for the tags to describe the musical content. Such a description is formally provided by a hierarchical tree of tags which summarizes the content of the songs at different levels. Tags such as 'love' or 'favorites' are thus prohibited.

This does not mean that all the tags in *Amazon* have the same level of precision. Because a tag is described as a path in a tree (see Section 4.2) which we broke down into as many tags as there are 'tag-nodes' in that path, the

position of a tag has a direct impact on the set of users to which it connects. Intuitively, the deeper the tag in the tree, the narrower the audience is likely to be.

Although it would be worth investigating this remark more deeply, we only focused here on the 25 tags most used by *Amazon* to describe the songs proposed on the platform. The list is provided in Figure 7, with the herfindahl diversity of each tag (left).

However, the pure herfindahl score (as defined in Section 3.2) is affected by the volume of the tags and normalization is then required. By applying the same model to derive the calibrated herfindahl score (Section 5.2), we managed to dispense with the correlation in order to achieve a more comprehensive picture of the diversity of the tags. Figure 7(b) shows the results of the normalization. One can clearly see that, similarly to what we observed for the *MSD* case, although 'Pop' and 'Rock' have a particularly high volume, the diversity of their audiences is lower than for other tags, such as 'Adult-alternative' or 'Singer-Songwriters', to cite the highest. The latter, in particular, could obviously stand for a *generic tag* (in the sense used in Section 5.1) since it is more likely to depict a property of the singer (who also composes the song) than of the musical content of the song. As such, they are likely to be used to tag songs belonging to very different types of music, hence the high diversity of their audiences.

Again, although it is difficult to draw general conclusions as to the high diversity of some tags in both datasets[9], the two results (Figures 5(b) and 7(b)) support our claim that the calibrated herfindahl score (referred to hereinafter simply as the *herfindahl diversity*) provides useful information on the true diversity of the tags used in the context of musical content. In particular, it is independent of the volume of the tags.

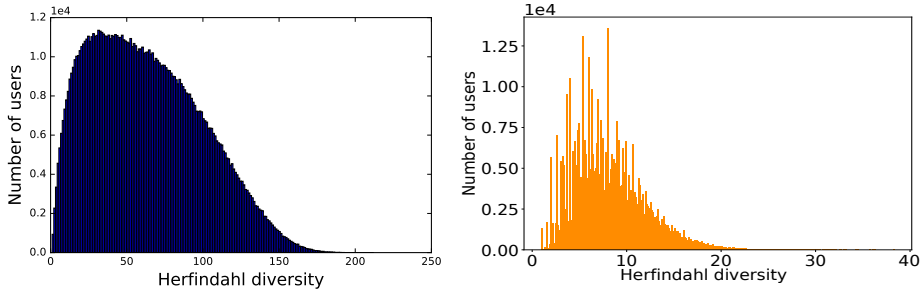
6. Diversity of users' attention

We will now turn to the analysis of the diversity of users' attention. Formally, the approach is similar to the one used in the previous section, except that the random walks start from $\perp$ nodes instead of $\top$ ones.

Figure 8 shows the distribution of the herfindahl diversity of all users' attention in *MSD* and *Amazon*. The plots, in the lin-lin scale, clearly demonstrate a homogeneous distribution, well-centered on average values (the mean and median are, respectively, 63 and 59 for *MSD*, and 8 and 7 for *Amazon*) even if some users with a particularly high diversity of attention can be observed. This is in line with the distribution of users' volume (Section 4.3) and manual investigations revealed that most of the highly diverse users correspond to the outliers observed in Figure 3(a).

In addition to the distribution of diversity, the previous section focused more precisely on 25 tags with a view to ascertaining whether the index introduced could differentiate between different behaviors. Unfortunately, due to

[9]To justify this diversity, more in-depth observations on the users would be required.



(a) Million Song Dataset (b) Amazon Dataset

Figure 8: Distribution of the diversity of users' attention.

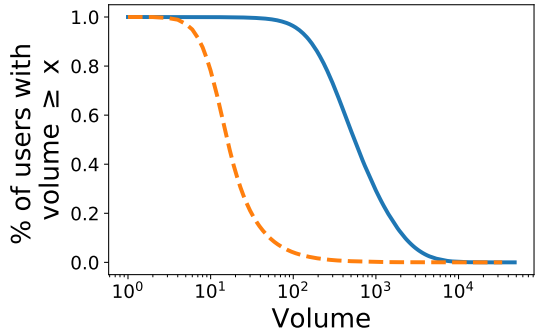


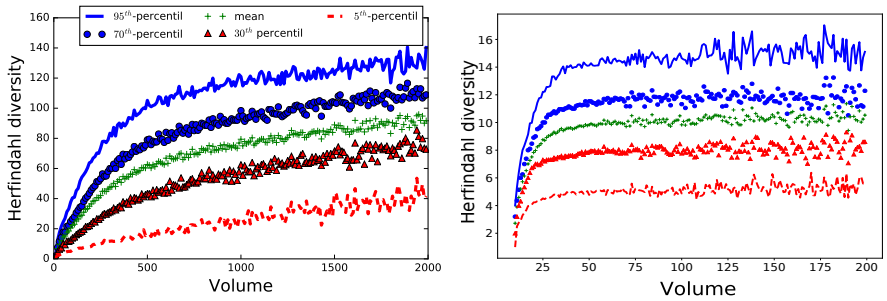
Figure 9: Distribution of users' volume for *MSD* (plain blue line) and *Amazon* (dashed orange line).

anonymization processes, we have no information on the users1 [10]. Therefore, we cannot relate the herfindahl diversity to external explanations, similarly to what we did with the tags at the semantic level.

However, it is possible to examine how the diversity of a user's attention relates to its volume. Indeed, while the diversity can be expected to be correlated to the volume, it is not clear how the correlation functions.

Figure 9 provides some elements to help investigate this question. In particular, the plain blue line displays the distribution of the user volume for the *Million Song Dataset*, *i.e.*, the number of tags for all the songs listened to by the user multiplied by their play counts. Although the x-axis is in log-scale, one

[10] a user is simply identified by a hash value.



(a) *Million Song Dataset* (b) *Amazon Dataset*

Figure 10: Evolution of the diversity of users' attention as a function of its volume.

can see that the order of magnitude is homogeneous in the network. The vast majority of the users (87%) have a volume between 10 and 2,000.

A similar observation can be made for the *Amazon Dataset* (dashed orange line). Although the volume may be higher than 10,000, most of the users (96%) do not write more than 100 reviews.

This provides enough elements to study how the diversity of a user's attention evolves as its volume grows. Figure 10 depicts this evolution for volumes less than or equal to 2,000 for *MSD* and 200 for *Amazon*[11]. For each volume, the plot shows the mean herfindahl diversity observed, along with the 5^{th}, 30^{th}, 70^{th} and the 95^{th} percentiles.

In the two cases, the influence of volume on diversity can be seen to operate in two phases. Below 500 for *MSD* (see Figure 10(a)) and 40 for *Amazon* (see Figure 10(b)), the progression is sharp, particularly for the upper part of the population (mean and above), while the progression is slower for higher volumes.

This indicates some sort of saturation in the diversity of users' attention: as the volume of listening or reviews reaches a certain threshold (500 and 40 in our observations, respectively), an average user starts to listen repeatedly to similar musical content or write reviews related to similar musical content (approximated by the tags). Consequently, the diversity tends to increase more slowly.

The redundancy observed in users' listening explains why the average diversity (63 for *MSD*, 8 for *Amazon*) is far from the maximal theoretical values (1,000 and 250, respectively). Although the volume tends to widen a user's musical perspective, its taste for a limited amount of different contents limits their musical diversity.

The analysis proposed thus far (Sections 5 and 6) has relied on choices

[11] for higher values, we had too few users to have confidence on the observations

made to capture the notion of diversity. Several elements could have been approached differently and exploring their impact allows us to investigate two new questions: how can we account for different notions of diversity (Section 7) and can we exploit random models in order to provide objective explanations for the analyses (Section 8)?

7. Exploring other diversity indexes

Given a distribution of probabilities $P = (p_i)_i$ issued from a random walk, in Section 3.2 we proposed using an adaptation of the herfindahl index to quantify the extent to which the distribution is close to a uniform distribution, hence the maximal diversity. Other choices could have been made by using other indexes such as the ones defined below:

Richness [59]	$\texttt{div}_0(P)$	$= \sum_i \mathbb{1}_{p_i>0}$
Shannon [38]	$\texttt{div}_1(P)$	$= 2^{-\sum_i p_i log(p_i)}$
Herfindahl [42]	$\texttt{div}_2(P)$	$= (\sum_i p_i^2)^{-1}$
Berger-Parker [43]	$\texttt{div}_\infty(P)$	$= (max_i(p_i))^{-1}$

Interestingly, these four indicators can be unified using the following definition of a diversity function `div`:

$$\texttt{div}_\alpha(P) = (\sum_i p_i^\alpha)^{\frac{1}{(1-\alpha)}}$$

where α is now a parameter of the *level* of diversity one wishes to capture. In this paper, we have used this very diversity function, with $\alpha = 2$.

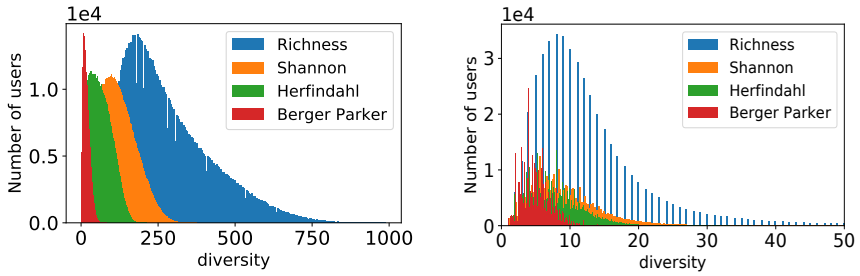
To facilitate the explanation that follows, we will focus on the interpretation of those indicators when applied to random walks issued from users ($\perp$ nodes), *i.e.*, when analyzing the diversity of users' attention (see Section 6).

In this context, when $\alpha = 0$ (Richness), one simply counts the number of categories that a random walk reaches, without considering its distribution. On the contrary, when $\alpha = \infty$ (Berger-Parker), one only considers the value of the category with the highest probability of being reached, without considering the other categories.

Intuitively, these two indicators constitute the two dimensions one wishes to capture in a diversity index: the number of categories reached and their distribution. In that regard, when $\alpha = 1$ (Shannon[12], widely used in information theory) or $\alpha = 2$ (Herfindahl, mostly used in economics and social sciences), a more nuanced point of view can be provided by accounting for those two dimensions.

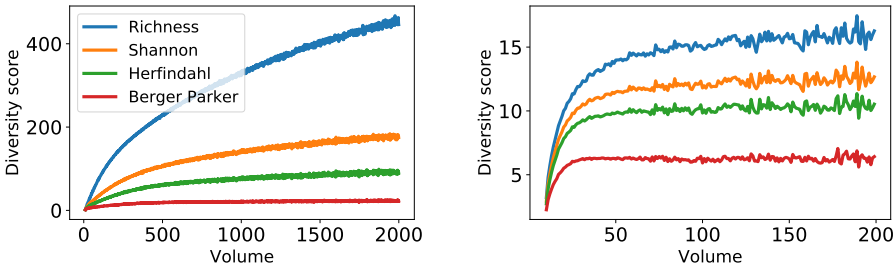
It is worth noting that, like the herfindahl diversity used in this article, those indicators are all bounded by the number of categories reached by the random walks (which determines the Richness value) and that, for $\alpha \geq 1$, the

[12] more formally, for $\alpha \to 1$, $\texttt{div}_\alpha(P)$ tends to Shannon.



(a) *Million Song Dataset* (b) *Amazon Dataset*

Figure 11: Distribution of the four diversity indexes for users' attention



(a) *Million Song Dataset* (b) *Amazon Dataset*

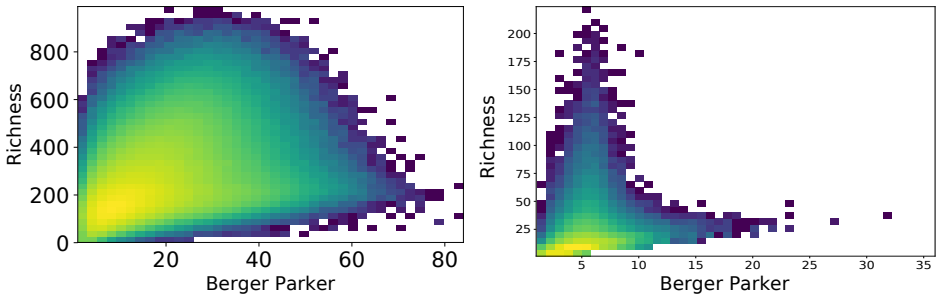
Figure 12: Evolution of the mean of the four diversity indexes as a function of the volume

maximum value is reached when the distribution is uniform. This provides a unified framework within which to explore the different facets of diversity.

This is what we investigate in Figure 11, which displays the distribution of the four diversity scores for the two datasets. One can see that the higher the order α of diversity, the narrower the distribution of the diversity towards low values. This shows that, as the order of diversity increases, it becomes more difficult to exhibit a high diversity score.

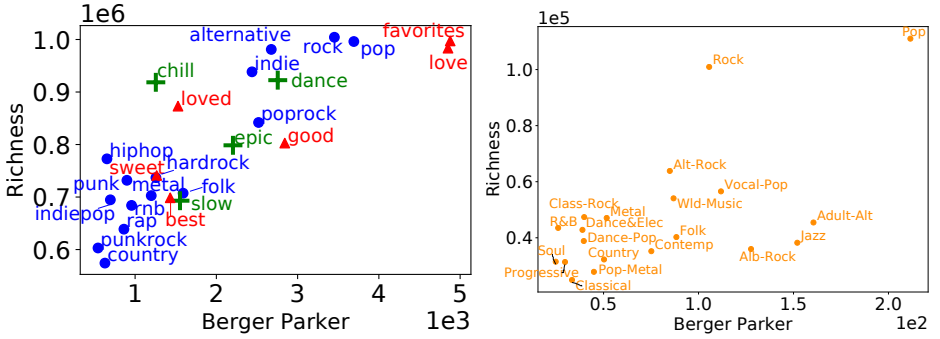
This is also confirmed when analyzing the evolution of the diversity of users' attention according to volume (Figure 12). The same saturation process can be seen as in Section 6, except that it converges towards different values of diversity. In particular, the threshold observed (around 500 for *MSD*, and 40 for *Amazon*) is not affected by the order of the index, which confirms the analysis of the saturation process.

Observed at the macro-level, this tendency does not necessarily mean that the scores capture the same behavior at the individual level. A given node can be highly diverse on a given order of diversity but relatively low on another. To demonstrate this, we compared the Richness and the Berger-Parker diversity of



(a) *Million Song Dataset* (b) *Amazon Dataset*

Figure 13: Richness diversity compared to Berger-Parker diversity (users' attention)



(a) *Million Song Dataset* (b) *Amazon Dataset*

Figure 14: Richness diversity compared to Berger-Parker diversity (tag audience)

each user (Figure 13) and each tag selected (Figure 14).

If there is a natural trend that shows a correlation between the two scores – particularly strong for the tags in *MSD* (Figure 14(a)) and *Amazon* (Figure 14(b)) as well, to a lesser extent, as for the users in *MSD* (Figure 13(a)) –, one can detect some users and tags with very contrasting diversity behavior.

For instance, 'hiphop' and 'punk' have a high Richness value on *MSD* (Figure 14(a)) but a low Berger-Parker value. This indicates that, although both reach a high number of users (hence a high Richness), only a relatively small core of users regularly listen to songs from these categories (hence a low Berger-Parker diversity).

Conversely, tags 'Jazz' and 'Adult-Alt' reach a relatively small number of users in *Amazon* (Figure 14(b)), but they write more regularly about songs of these categories. The distribution is thus closer to a uniform distribution, hence with a relatively high Berger-Parker value.

These examples show that analyzing diversity at different orders provides a more comprehensive picture of diversity in a dataset.

8. Impact of random models.

To try to compensate for the impact of volume on the herfindahl diversity (see Section 5.2), a second choice was made. Following a common method, we compared the score of the diversity to what it would have been in a similar random structure. In this approach, everything relies on the notion of *similar random structure*. We chose to shuffle the links between the bottom and middle layers of the tripartite graphs (*i.e.*, between the users and the songs) while keeping the rest unchanged.

This choice was motivated by the fact that the volumes of the tags remain unchanged in the process, thus legitimating the comparison of the values. However, several other ways to disturb the structure could have been proposed.

In order to investigate the different choices and their relation to the diversity of the tag audience and users' attention (Section 5 and 6), we have defined and compared several random models below. For the sake of simplicity, we chose to focus exclusively on the herfindahl diversity studied in the first part of the paper (that is, the diversity for $\alpha = 2$, see Section 7) and to exclude other indexes.

Our interest in using random models in this paper derives from the possibility of assigning users to categories. If we forget, for a moment, the tripartite structure used so far to represent the information contained in the datasets, the problem can be formulated as follows. Let ν be the number of objects (users), τ the number of sets (tags), v_u the value of object $u < \nu$ (number of times a song is listened to by u), and a_t the probability of placing an object in set t (probability of reaching tag $t < \tau$). Then the expected value E for the herfindahl index is given by:

$$E = \sum_t \alpha_t (\sum_u {v_u}^2 + \alpha_t (1 - \sum_u v_u^2))$$

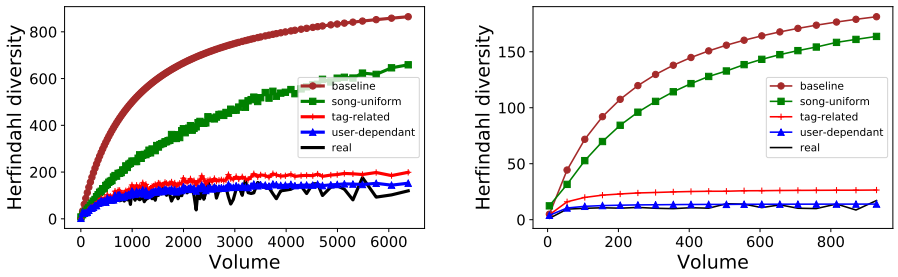
Given this generic formula, different models lead to a different value for E[13]. In particular, there are several ways to take into account the specific tripartite structure related to the diversity we wish to capture. Considering different realistic constraints, one can assume that:

baseline: a user u selects the tags with equal probability. In that case, $\alpha_t = \frac{1}{\tau}$ and $v_u = \frac{1}{V_u}$ (where V_u is the volume of user u)

Note that this baseline model does not account for the middle layer of the tripartite graphs. The distribution of the songs in the structure has no impact and users are directly related to the tags, uniformly at random.

song-uniform: a user u selects a song s uniformly at random. Then, u splits its value v_u into $\frac{1}{d_\top(s)}$ equal objects placed in the $d_\top(s)$ tags attached to s (where $d_\top(s)$ is the degree of s towards the tags).

[13] to be completely formal, the expected herfindahl *diversity* is the inverse of E (see Section 7).



(a) *Million Song Dataset* (b) *Amazon Dataset*

Figure 15: Evolution of the expected herfindahl diversity as a function of the volume

This is the model used in the first part of the paper, which led to the *calibrated herfindahl diversity*.

tag-related: building on the former model, one can add that the probability α_t to place an object in the set t depends on the number of songs tagged as t. Formally, $\alpha_t = \frac{d_\perp(t)}{\sum_{t'} d_\perp(t')}$ where $d_\perp(t')$ is the degree of a tag t'.

This is one way to take into account the popularity of the music styles in the model since α_t is now proportional to the popularity (degree) of t.

user-dependant: Finally, relying on the fact that a user that listens to a song has a natural tendency to listen to songs of the same music style, one can limit the selection of the songs to dependent variables.

The evolution of the expected diversity according to the users' volume is displayed in Figure 15 for the four models presented above: baseline (brown discs); song-uniform (green squares); tag-related (red crosses); and user-dependent (blue triangles). In addition, we display the evolution of the (not calibrated) herfindahl diversity (black line).

The first thing to observe is that all the models show the saturation process observed on the real evolution, even if the baseline and the song-uniform models exhibit a smoother inflection. The second observation is that, as is to be expected, by adding realistic constraints to the model, the curves get closer to the real plot, showing that these properties have a direct impact on diversity.

This also explains why diversity stagnates as volume increases. The empirical limits observed in the data (around 200 for *MSD*, 40 for *Amazon*) are indeed far from the maximal diversity that could be achieved (1,000 and 250, respectively). In that regard, if the diversity of the baseline model tends to be close to the theoretical maximum, one can observe, on the other hand, that the tag-related and user-dependent models successfully reproduce the empirical limit depicted by the real curve. This shows that the properties taken into account in those models mechanically limit the diversity of users' attention.

In relation to the tag-related model, the straightforward interpretation is that the heterogeneous popularity of the tags (see Figure 3(d)) sufficiently explains why the diversity of attention is limited. Even if the tendency of users to listen to similar musical content (property added in the user-dependent model) supports this observation, by becoming even closer to the real curve, the gap between the two models is small and one might conjecture that the main reason for the limitation of diversity is an *exogenous pressure* related to music popularity, rather than internal factors modeled by user behavior.

The latter result is interesting in itself, as it relates to a more general debate on whether limitations of diversity (the echo chamber phenomena on social platforms, for instance) are due to user behavior (the user tends to be more likely to focus repeatedly on similar content) or by external factors (such as social pressures, algorithmic recommendations, and so on).

While this study does not offer a conclusion to the discussion, we believe that both the methodology adopted (the use of the calibrated herfindahl diversity computed on a tripartite graph) and the results achieved (the comparisons of the different models) provide strong evidence of external reasons for the empirical observation that diversity is limited.

9. Conclusion & perspectives

In this paper, we have investigated the question of quantifying diversity in users' activity. We represented this activity in the form of a tripartite graph relating users to products and products to categories. We then performed random walks from user nodes towards tag nodes and, conversely, from tag nodes towards user nodes. These random walks allowed us to study how users' attention and the tag audience are distributed.

Building on these distributions, we defined herfindahl diversity as a way of quantifying how close the distributions of probabilities obtained from the random walks are to a uniform distribution, which is representative of a perfect diversity.

We applied this approach to two datasets that record user activity on online platforms involving musical content. The results were threefold. First, by analyzing the tag audience, we showed how to compensate for the effect of volume on diversity and defined the *calibrated herfindahl diversity*. The latter score proved to be a good indicator for differentiating between style tags (low diversity) and generic tags (high diversity), independently of their volume (Section 5).

Second, by focusing on users' attention, we studied the relation between the volume of user activity and its diversity. This revealed a saturation phenomenon: while the growth of diversity is initially high as volume increases, this progression slows down after a certain threshold, indicating that when the volume is too high, users tend to repeatedly connect to similar musical content (Section 6).

Finally, by investigating the relation between different models and the expected herfindahl diversity of the structures generated, we provided evidence that explains the reason for the limitation observed in the diversity of users. Ex-

ogenous factors related to the heterogeneous popularity of music styles generate constraints that seem to limit the expression of diversity by users (Section 8).

Along with these results, we also demonstrated that our approach is sufficiently generic to capture different mathematical notions of diversity. In particular, we illustrated that although the values of diversity at different orders are globally correlated, our general framework provides a more nuanced picture of the different facets of diversity at the individual level (Section 7).

This work has the potential to be expanded on in several directions. The first development would consist in applying the method to other domains in order to study whether the current observations also hold in contexts that do not involve musical content. Another interesting avenue would be to continue the study proposed in Section 8 by seeking similar results with other diversity indexes, such as those proposed in Section 7. This would refine our understanding of which facets of diversity are impacted by the properties captured by the different models.

An even more promising development would be to use this approach to analyze the behavior not of users but rather of the recommendation systems used on online platforms. In all the analyses conducted in the present study, a link between a user and a song was triggered by the user's activity on the platform. However, depending on the data available, such a link could depict a song *recommended* by an algorithm. This would in turn allow us to study the diversity of the recommendation itself.

Combining the two analyses – the diversity of users' activity and that of the algorithmic recommendations – should shed light on the intricate relation between users and recommendation algorithms that shape activity on online platforms [53]. We believe that our framework would provide an ideal tool for such analyses and would open up many promising avenues of future research.

Finally, along with the straightforward developments presented above, the results set out in this paper have the potential to provide more important outcomes for society itself. Returning to the initial objectives underpinning our research question (see Section 1), our results demonstrate the relevance of the approach for incorporating the notion of diversity into recommendation systems.

One evident potential use would be to help analyse the *effects* of online algorithms, which is an interesting and useful avenue of investigation for the research community. As advocated in a recent article published in *Nature*, we need the scientific community to contribute to the detection of biases in algorithms [60]. Such a task would clearly entail designing, testing and demonstrating the relevance of fairness metrics (such as the one captured by our framework) that can monitor their effects [61]. Diversity measures could even be incorporated *into* recommendation systems, which would then directly integrate diversity as an optimization criterion of the underlying algorithm [29]. This is expected to result in the mechanical proposition of a more diverse set of items, leading to an improved user experience.

There is no doubt that designing effective ways to audit (or improve) recom-

mendation algorithms would empower users when it comes to their dependency on online social platforms. By providing a more comprehensive picture of the properties related to the items recommended to them, one can in turn expect users to become more aware of the implications driven by their use of online services.

Acknowledgment

This work is funded in part by the European Commission H2020 FET-PROACT 2016-2017 program under grant 732942 (ODYCCEUS), the ANR (French National Agency of Research) under grant ANR-15-CE38-0001 (AlgoDiv) and the CNRS (French National Center for Scientific Research) under grant PICS RÉCITAL (n° 245 709).

[1] V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, Boston, 2013.

[2] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, Cambridge, MA, USA, 2015.

[3] T. Gillespie, The relevance of algorithms, in: P. B. Tarleton Gillespie, K. Foot (Eds.), Media Technologies, MIT Press, 2014, pp. 167–194.

[4] S. Barocas, A. D. Selbst, Big Data's Disparate Impact, California Law Review 104 (3) (2016) 671–732.

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

[6] L. Sweeney, Discrimination in online ad delivery, Queue 11 (3) (2013) 10:10–10:29. doi:10.1145/2460276.2460278.

[7] A. Datta, J. Makagon, D. Mulligan, M. Tschantz, Discrimination in online personalization: A multidisciplinary inquiry, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ACM, 2018.

[8] A. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. Lagendijk, Q. Tang, Privacy in recommender systems, in: N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, X.-S. Hua (Eds.), Social Media Retrieval, Computer Communications and Networks, Springer Verlag, 2013, pp. 263–281.

[9] R. Epstein, R. E. Robertson, The search engine manipulation effect (seme) and its possible impact on the outcomes of elections, Proceedings of the National Academy of Sciences 112 (33) (2015) E4512–E4521. doi:10.1073/pnas.1419828112.

[10] R. E. Robertson, D. Lazer, C. Wilson, Auditing the personalization and composition of politically-related search engine results pages, in: Proceedings of the 2018 World Wide Web Conference, ACM, 2018, pp. 955–965. doi:10.1145/3178876.3186143.

[11] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management (2019) 102025doi:https://doi.org/10.1016/j.ipm.2019.03.004.

[12] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (6239) (2015) 1130–1132. doi:10.1126/science.aaa1160.

[13] D. Beer, The social power of algorithms, Information, Communication & Society 20 (1) (2017) 1–13. doi:10.1080/1369118X.2016.1216147.

[14] D. Wolfram, Search characteristics in different types of web-based ir environments: Are they the same?, Information processing & management 44 (3) (2008) 1279–1292.

[15] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, D. Kluver, Exploring author gender in book rating and recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, 2018, pp. 242–250. doi:10.1145/3240323.3240373.

[16] E. Bozdag, Bias in algorithmic filtering and personalization, Ethics and Information Technology 15 (3) (2013) 209–227. doi:10.1007/s10676-013-9321-6.

[17] G. Adomavicius, Y. Kwon, Improving aggregate recommendation diversity using ranking-based techniques, IEEE Transactions on Knowledge and Data Engineering 24 (5) (2012) 896–911. doi:10.1109/TKDE.2011.15.

[18] P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, N. J. Stroud, Bursting your (filter) bubble: Strategies for promoting diverse exposure, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion, ACM, 2013, pp. 95–100. doi:10.1145/2441955.2441981.

[19] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on World Wide Web, ACM, 2005, pp. 22–32.

[20] B. Smyth, P. McClave, Similarity vs. diversity, in: Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, ICCBR '01, Springer-Verlag, 2001, pp. 347–361.

[21] A.-A. Stoica, C. Riederer, A. Chaintreau, Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity, in: Proceedings of the 2018 World Wide Web Conference, ACM, 2018, pp. 923–932. doi:10.1145/3178876.3186140.

[22] C. Sha, X. Wu, J. Niu, A framework for recommending relevant and diverse items., in: IJCAI, 2016, pp. 3868–3874.

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

[24] M. Schedl, P. Knees, F. Gouyon, New paths in music recommender systems research, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 2017, pp. 392–393.

[25] M. Kaminskas, F. Ricci, Contextual music information retrieval and recommendation: State of the art and challenges, Computer Science Review 6 (2-3) (2012) 89–119.

[26] T. Cebrián, M. Planagumà, P. Villegas, X. Amatriain, Music recommendations with temporal context awareness, in: Proceedings of the fourth ACM conference on Recommender systems, ACM, 2010, pp. 349–352.

[27] J. H. Lee, H. Cho, Y.-S. Kim, Users' music information needs and behaviors: Design implications for music information retrieval systems, Journal of the association for information science and technology 67 (6) (2016) 1301–1330.

[28] M. Slaney, W. White, Measuring playlist diversity for recommendation systems, in: Proceedings of the 1st ACM workshop on Audio and music computing multimedia, ACM, 2006, pp. 77–82.

[29] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, T. Jambor, Auralist: introducing serendipity into music recommendation, in: Proceedings of the fifth ACM international conference on Web search and data mining, ACM, 2012, pp. 13–22.

[30] R. Poulain, F. Tarissan, Quantifying the diversity in users activity: an example study on online music platforms, in: Proceedings of the Fifth International Conference on Social Networks Analysis, Management and Security, IEEE, 2018, pp. 3–10.

[31] K. S. McCann, The diversity–stability debate, Nature 405 (2000) 228–233.

[32] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (6239) (2015) 1130–1132. doi:10.1126/science.aaa1160.

[33] A. Stirling, A general framework for analysing diversity in science, technology and society, Journal of the Royal Society, Interface 4 (15) (2007) 707–719. doi:10.1098/rsif.2007.0213.

[34] L. Zhang, R. Rousseau, W. Glänzel, Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account, Journal of the Association for Information Science and Technology 67 (2016) 1257–1265. doi:10.1002/asi.23487.

[35] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 519–528. doi:10.1145/2187836.2187907.

[36] Q. Li, Y. Liu, Exploring the diversity of retweeting behavior patterns in chinese microblogging platform, Information Processing & Management 53 (4) (2017) 945–962.

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

[38] C. E. Shannon, A mathematical theory of communication, Bell system technical journal 27 (3) (1948) 379–423.

[39] A. Rényi, 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, University of California Press, Berkeley, CA, 1961, pp. 547–561.

[40] C. Gini, Measurement of inequality of incomes, The Economic Journal 31 (121) (1921) 124–126.

[41] A. Hirschman, The paternity of an index, The American economic review 54 (5) (1964) 761–762.

[42] S. A. Rhoades, The herfindahl-hirschman index, Federal Reserve Bulletin, 79 (1993) 188.

[43] W. H. Berger, F. L. Parker, Diversity of planktonic foraminifera in deep-sea sediments, Science 168 (3937) (1970) 1345–1347.

[44] P. J. Alexander, Entropy and popular culture: Product diversity in the popular music recording industry, American Sociological Review 61 (1) (1996) 171–174.

[45] P. D. Lopes, Innovation and diversity in the popular music industry, 1969 to 1990, American Sociological Review 57 (1) (1992) 56–71.

[46] M. Bacache-Beauvallet, M. Bourreau, F. Moreau, Piracy and creation: The case of the music industry, European Journal of Law and Economics 39 (2) (2015) 245–262.

[47] B. Ferwerda, A. Vall, M. Tkalcic, M. Schedl, Exploring music diversity needs across countries, in: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, ACM, 2016, pp. 287–288. doi:10.1145/2930238.2930262.

[48] M. Slaney, W. White, Measuring playlist diversity for recommendation systems, in: Proceedings of the 1st ACM workshop on Audio and music computing multimedia, ACM, 2006, pp. 77–82. doi:10.1145/1178723.1178735.

[49] Z. Zhou, K. Xu, J. Zhao, Homophily of music listening in online social networks of china, Social Networks 55 (2018) 160 – 169. doi:https://doi.org/10.1016/j.socnet.2018.07.001.

[50] M. T. Ribeiro, A. Lacerda, A. Veloso, N. Ziviani, Pareto-efficient hybridization for multi-objective recommender systems, in: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, ACM, 2012, pp. 19–26. doi:10.1145/2365952.2365962.

[51] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, ACM, 2011, pp. 109–116. doi:10.1145/2043932.2043955.

[52] J. P. Kelly, D. Bridge, Enhancing the diversity of conversational collaborative recommendations: A comparison, Artificial Intelligence Review 25 (1-2) (2006) 79–95. doi:10.1007/s10462-007-9023-8.

[53] S. Yang, Analysis of user behavior, in: B. Fang, Y. Jia (Eds.), Groups and Interaction, De Gruyter, 2019, Ch. 1, pp. 1–62. doi:10.1515/9783110599411.

[54] M. G. Everett, S. P. Borgatti, The dual-projection approach for two-mode networks, Social Networks 35 (2) (2013) 204–210. doi:http://dx.doi.org/10.1016/j.socnet.2012.05.004.
URL http://www.sciencedirect.com/science/article/pii/S0378873312000354

[55] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, P. Lamere, The million song dataset, in: Proceedings of the 12th International Conference on Music Information Retrieval, University of Miami, 2011, pp. 591–596.

[56] R. He, J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: Proceedings of the 25th International Conference on World Wide Web, ACM, 2016, pp. 507–517. doi:10.1145/2872427.2883037.

[57] J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2015, pp. 43–52. doi:10.1145/2766462.2767755.

[58] M. E. Newman, The structure and function of complex networks, SIAM review 45 (2) (2003) 167–256.

[59] R. H. MacArthur, Patterns of species diversity, Biological reviews 40 (4) (1965) 510–533.

[60] R. Courtland, Bias detectives: the researchers striving to make algorithms fair, Nature 558 (7710) (2018) 357–360.

[61] B. Salimi, L. Rodriguez, B. Howe, D. Suciu, Interventional fairness: Causal database repair for algorithmic fairness, in: Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, ACM, 2019, pp. 793–810. doi:10.1145/3299869.3319901.

Appendix A. Additional results

This section provides additional results deriving from the fact that random models are at the core of our method to define the calibrated version of the herfindahl diversity. As explained in Section 5.2, the herfidahl diversity measured on a real network needs to be compared with the *expected* herfindahl diversity that would have been observed on a random network with similar properties.

We chose to use the configuration model [58] that generates random networks while preserving the degree distributions. The analyses presented in Sections 5 and 6 were all based on this random model. However, Section 8 showed that other choices could have been appropriate and one might question what the results would have been for other models. This is the subject of this appendix.

It should be noted that, as highlighted in Section 8, the model referred to as the *song-uniform* model corresponds to the calibrated herfindahl diversity analyzed in the paper. For this reason, in this section we only present the results for the three other models: namely, the *baseline*, *tag-related*, and *user-dependant* models.

Figures A.16 and A.17 show the calibrated herfindahl diversity of the tag audience when using the *baseline* (top row), *tag-related* (middle row) and *user-dependent* models (bottom row) on the two datasets. Those figures correspond to Figures 6 and 7(b), respectively, both of which use the *song-uniform* model on the same datasets.

The first observation to be drawn from these figures is that, as expected, the *baseline* model (top row of the figures) fails to compensate for the effect of volume on the diversity score. In that regard, the *song-uniform* model employed in this paper is more effective.

A more interesting point to consider is what can be learnt from the models that factor in more realistic properties. While both the *tag-related* (middle row) and the *user-dependent* (bottom row) models do not differ greatly from the *song-uniform* model in the sense that they all compensate for the effect of volume, subtle differences can nonetheless be noted.

For instance, if one looks at the top inset for *MSD* (middle column of Figure A.16), it can be seen that the relation between the tags 'poprock', 'dance'

and 'chill' varies as the models change. While the diversity of 'poprock' is clearly higher with the *song-uniform* model (Figure 6), its value decreases with the *tag-related* model, where its diversity is in between that of 'dance' and 'chill'. The diversity of 'poprock' is even lower in the *user-dependant* model, where both 'chill' and 'dance' are considered as more diverse tags.

A similar observation can be made for the tags 'Contemp', 'Wild-Music' and 'Alt-Rock' for the *Amazon* dataset, for instance, when comparing the middle column of Figure A.17 and Figure 7(b).

Although the general conclusions drawn in this paper remain largely consistent with these observations, the last remarks call for a deeper investigation of the impact of the model, which would constitute an interesting avenue of analysis for future studies.

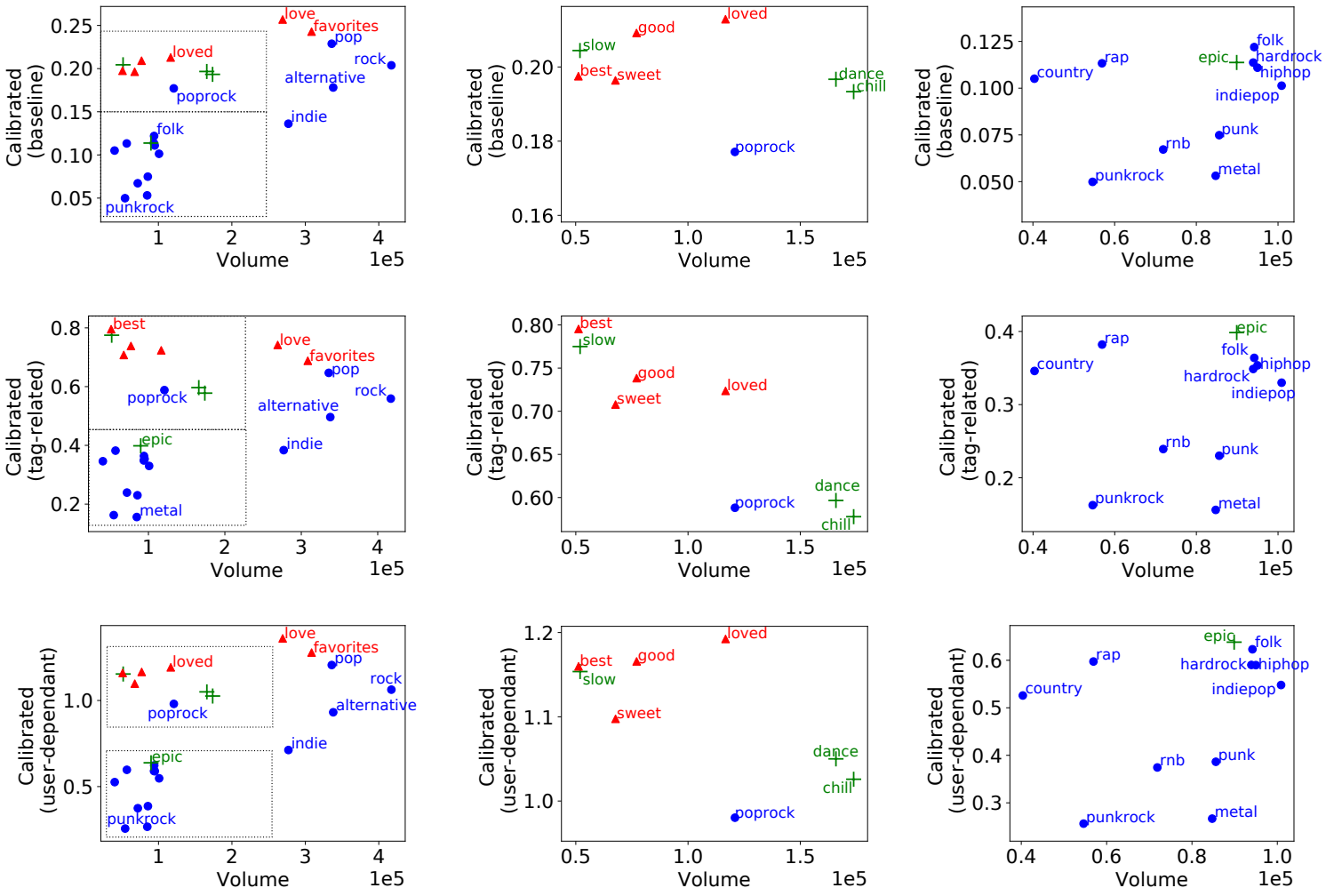


Figure A.16: Calibrated herfindahl diversity of the tag audience according to its volume for the *Million Song Dataset* using the *baseline* model (top row), the *tag-related* model (middle) and the *user-dependat* model (bottom). For each model, we show the result for all tags (left column), the top inset (middle) and the bottom inset (right).

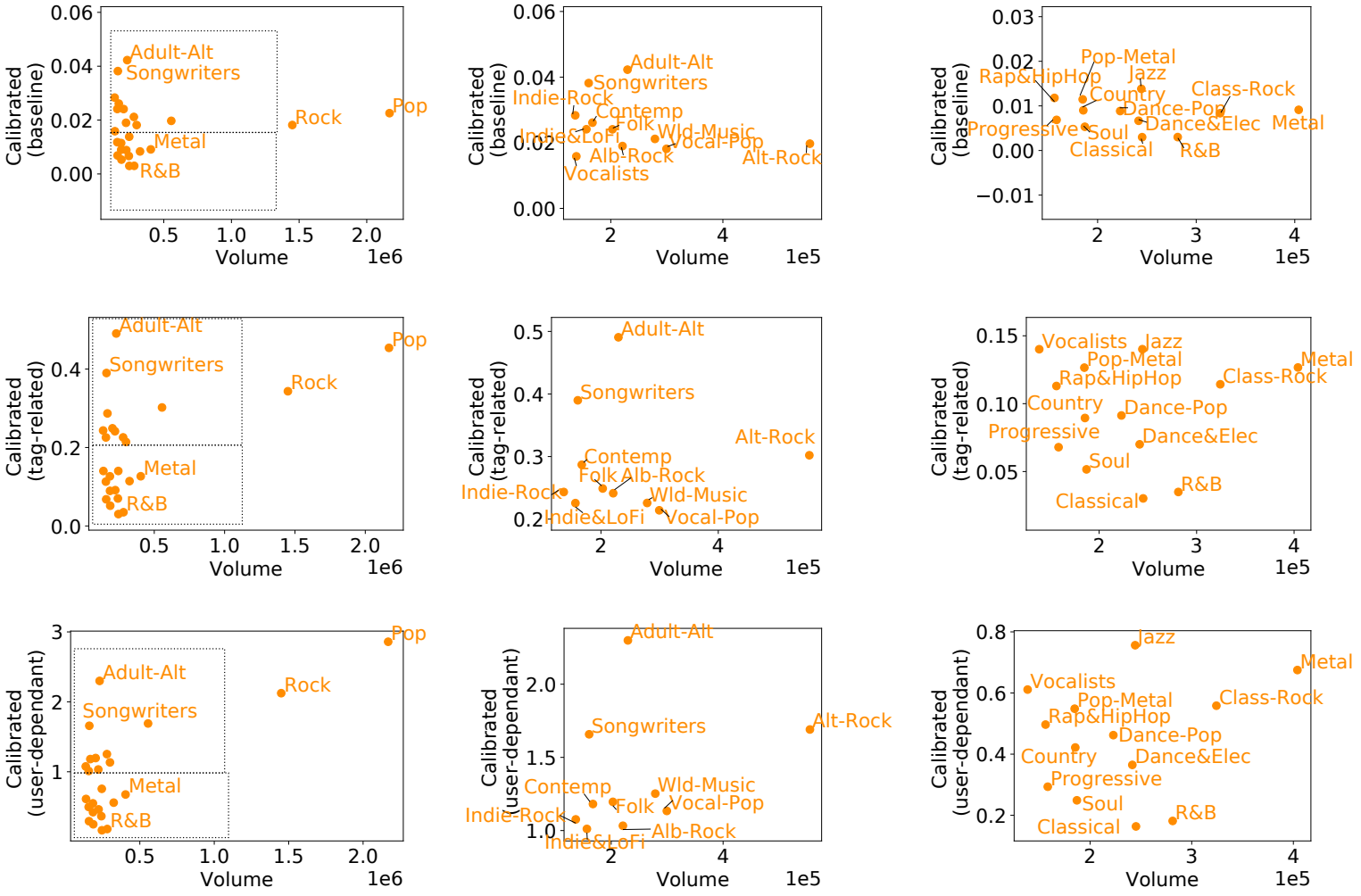


Figure A.17: Calibrated herfindahl diversity of the tag audience according to their volume for the *Amazon Dataste* using the *baseline* model (top row), the *tag-related* model (middle) and the *user-dependat* model (bottom). For each model, we show the result for all tags (left column), the top inset (middle) and the bottom inset (right).