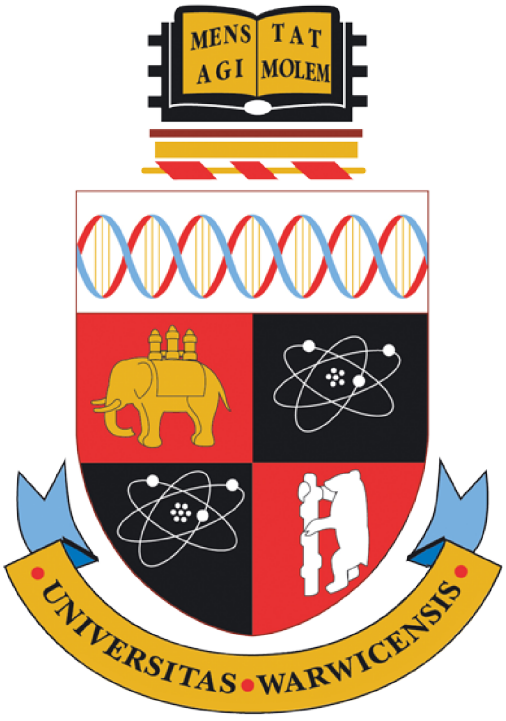
Measuring the diversity of set of items consumed by users with applications to streamed music

Gabrielė Stravinskaitė

1620789

Supervised by Julia Brettschneider (University of Warwick)
and Lucas Maystre (Spotify)



Department of Statistics
University of Warwick
United Kingdom
6 September 2020

Acknowledgements

I thank Julia Brettschneider for her superb assistance during the whole project, especially with helping to translate the concepts into more rigorous mathematical language. I am grateful to Lucas Maystre for help with the embedding, generating ideas about the evaluation strategy for the predominately qualitative problem as well as for helping to keep my work relevant for the applications. Finally, I thank Eavan Thomas Pattie for assisting me in the idea development, discussions about the mathematical concepts and notation as well as for correcting my English mistakes.

Contents

Measuring the diversity of set of items consumed by users with applications to streamed music

Gabrielė Stravinskaitė

September 6, 2020

Abstract

Algorithmic recommendation systems are an important part of many online platforms. However, research indicates that the commonly used recommender systems only work for those users who have a rather non-diverse taste. Furthermore, algorithmic recommendations lead a user towards even less diverse taste. This is an undesirable consequence because the diversity of a user's taste is positively associated with the longevity of that user in a platform. Thus, there is a need to detect diverse users and understand them better. This project focuses on measures of diversity which could help to classify users according to their taste. Online music platforms were chosen as a domain of application. Statistical diversity measures such as Shannon entropy, Kullback-Leibler divergence, Gini-Simpson index and Hill number alongside the geometrical diversity measures such as the Generalist-Specialist and TS-SS scores are evaluated. Statistical diversity measures could be preferred when the interpretation of the score is important whilst the geometrical scores could be favoured for being able to determine how similar different songs in a playlist are. Finally, using a naive recommender system prototype, it was concluded that both geometrical diversity measures outperform the statistical scores in explaining which users are unlikely to benefit from the simple recommendation algorithm.

1 Introduction

In many online platforms, users engage with a huge number of different items, be it movies, music, posts, etc. The set of items they encounter as a result of their own exploration or algorithmic recommendations shape their experience and the longevity of a user in a specific platform. One important aspect of the user's behaviour is how *diverse* is the set of items she is using. Previous research showed that the users with a more diverse taste tend to have higher retention of the platform [1, 19] which is a desirable quality for any online platform provider. Therefore, there is a need to measure the level of diversity of a user as well means to promote it. But *what is* a diverse taste? How one would describe the diversity by considering the actual set of items consumed by the user? It might be relatively simple to define an extremely diverse and extremely non-diverse taste. We could, for example, assume that at one extreme a user engages to all the content in the platform at an exactly equal number of times whilst on the other extreme, a user only engages with only

one item a multiple number of times. However, even this is up to debate. The real question arises when we want to define the diversity of the users in-between these two extremes which are likely to be all or almost all of the users. Is the user who listens to Lady GaGa, Beyoncé and Madonna at equal frequencies is more or less diverse than a user who listens to the same three artists plus Slipknot, yet listens to Lady GaGa substantially more than to the other artists? This and similar questions are central to this project: what does the diverse taste mean and how can we measure it? This question will be attempted to answer with the application to the streamed music. Namely, the *EchoNest* (now part of *Spotify*) *Million Song Dataset* is used. As a result, several diversity measures are described and the notion of the diversity they carry along are evaluated.

The structure of the work is as follows. In section 2, previous work on the diversity measurement and the application of them in the music context are discussed. Section 3 proceeds to the theoretical evaluation of the statistical diversity measure, namely the Shannon entropy, Kullback-Leibler divergence, Gini-Simpson index and the Hill number. These measures are described when the diversity is measured on the set of songs as well as on the set of larger hierarchical categories such as artists, genres, etc. Section 4 then introduces the idea behind the geometrical diversity measures alongside the embedding space needed to implement them. Generalist-Specialist score introduced in the previous literature [1, 19], as well as a novel TS-SS score, [8] are evaluated from the theoretical perspective. Sections 5 and 6 prepare for the practical evaluation by describing the *Million Song Dataset*, the subsample of data being used as well as the practical evaluation strategy. Sections 7 to 10 discuss the results of the practical evaluation of all the diversity measures introduced. Section 11 attempts to show the diversity scores in action by simulating a naive recommender system prototype. Section 12 concludes the results.

2 Background

There is a rich literature about the diversity measures and their application in the information theory. These measures, stemming from the Shannon entropy and evolving to more sophisticated measures of information and its compression are widely applied in the fields such as thermodynamics, cryptography and ecology to name a few. The application of the information theory to the biological diversity is perhaps mathematically the closest well-researched field to that of the diversity of the music playlists, thus a lot of findings can be relatively easily adapted. Among the popular biodiversity measures are Shannon entropy, Gini-Simpson index, Rényi entropy and Hill number which can be all defined using different parameterisations of the Sharma-Mittal index [4]. These measures define the diversity in terms of the "surprise" associated with a set of items (Shannon entropy), average "surprise" (Rényi entropy), probability of two randomly drawn with replacement items being of a different type (Gini-Simpson index) and a number of equally abundant items needed to result to the same species pattern as empirically observed (Hill number). The different definitions of the diversity appeared to sometimes give different stories depending on the measure used even using the same data [10]. This stresses the importance of having a well-defined and

concrete definition of diversity before opting for a certain measure.

Regarding the online-platforms, the diversity of the users in *GitHub*, *Reddit* as well as *Spotify* was estimated using the Shannon entropy, Gini-Simpson index and the modification of the cosine-similarity introduced as the Generalist-Specialist score [1, 19]. Shannon entropy and Gini-Simpson index were argued to be inferior measurements owing to their inability to distinguish how similar are the song listened by the same user. Instead, the geometrical Generalist-Specialist score defined on the embedded song vectors was claimed to be a better option owing to its inherent geometrical notion of the similarity as a physical distance. Nonetheless, the correlation between the Generalist-Specialist score and the Shannon entropy appeared to be -0.77 and the correlation between the Generalist-Specialist and the Gini-Simpson index reached -0.60 suggesting that the measures would have likely resulted in the qualitatively similar conclusions about the behaviour of the users conditional on their diversity.

Whilst giving the correct theoretical critique of the Shannon entropy and the Gini-Simpson index versus the geometrical similarity measures, the authors of the papers evaluating the diversity in the online platforms did not discuss the individual features of the similarity measures. Thus, the contribution of this work is to re-evaluate the already mentioned measures alongside some previously not discussed measures in terms of the notion of the diversity they bear whilst focusing on the streamed music playlists.

To facilitate the further analysis we borrow the notion of the *generalist* and *specialist* user from the previous work [1]. We thus define a *specialist* user to be the one who, according to any definition of the diversity, has a very non-diverse taste in music and thus listens to a relatively narrow spectrum of songs. The *generalist*, on the other hand, is a user who listens to a diverse (again, according to any definition of diversity) set of songs.

3 Statistical similarity measures: theory

We first begin by defining a finite discrete probability space in which the diversity measures will be defined. Let $\mathcal{X} = \{x_1, x_2, ...x_N\}$ represent the discrete collection of songs and $S_i \subseteq \mathcal{X}$ - a collection of songs user i is listening to. Moreover, let p_{ij} define the probability of x_j being listened to by user i. One way of defining p_{ij} would be to assign a relative frequency of song j being played by the user i (for $i = 1, ..., n$), i.e

$$p_{ij} = \frac{f_{ij}}{\sum_{j \in S_i} f_{ij}} = \frac{f_{ij}}{F_i}$$

where f_{ji} is the number of times song j was listened to by user i. We will adopt this measure of p_{ij} for our purposes.

3.1 Shannon entropy

Shannon entropy is a very fundamental and thus popular metric in the information theory. It arises from the idea that the *value* of information depends on how *surprising* it is. If an event is very

likely, there is little surprise when observing it and thus it gives us little value in understanding the underlying distribution better. Yet, if a very rare event is observed, it is more surprising and thus gives us much more information about the distribution of the interest. We want the measure of the expected information surprise to have the following properties:

- to be entirely defined by the probability distribution of the data source.
- give maximum value for the most unpredictable - uniform - distribution and yield zero for the cases with 100% probability of an event occurring, i.e. $I(1) = 0$ where $I(\cdot)$ stands for the expected information surprise measure.

  As a result, the information surprise measure should be a function $I : [0, 1] \mapsto [0, \infty)$
- to be larger for the less likely outcomes, i.e. $I(m) > I(n)$ if $m < n$ where m, n stand for some probabilities.
- to be a continuous function in probability, i.e. $I(m)$ is continuous function of m.
- to allow for the independent events to be additive. This is because we want the two independent events to communicate the same information jointly as well as separately, i.e. $I(mn) = I(m) + I(n)$.

Claude Shannon [16] derived a measure of the information surprise that meets all of those criteria by replacing the expectation of a logarithm function with an average of a discrete space as follows:

$$\mathbb{E}[I(\Gamma)] = \mathbb{E}\left[log_b\left(\frac{1}{P(\gamma_i)}\right)\right] = -\sum_{i=1}^{N} P(\gamma_i) log_b(P(\gamma_i))$$

where $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_N\}$ is some set of events, $P(\gamma_i)$ - probability of an event γ_i happening and b - logarithm base usually set to be 2, 10 or e. When $b = 2$ the entropy can be also perceived as the expected number of bits you need to describe the distribution of the events. Due to this interpretation, we will from now use the logarithm base of 2.

We can adapt the Shannon entropy to suit the music diversity application using the notation introduced at the beginning of section 3 and define it as follows:

$$D^{Sh}(S_i) = -\sum_{j \in S_i} p_{ij} log_2(p_{ij})$$

Having the user's i playlist as the distribution we calculate Shannon entropy which yields us the following interpretation. The more song-loyal the user is, or in other words, the more often she tends to listen to the same songs for an extended period, the less surprising the user's next song choice will be. Hence, we need less information to explain the user's behaviour and thus the entropy will be low. On the other hand, if the user listens to all the songs in her playlist at equal frequencies, her playlist distribution will resemble that of the uniform distribution and hence would

result in high entropy. Thus low Shannon entropy score can be interpreted as a sign of a person being more specialist and high entropy, similarly is a sign of a person being more generalist.

Shannon entropy offers another interpretation of the *relative* surprise when comparing the observed user with an extremely generalist user who listens to all the songs uniformly and random. This idea will be discussed further in the following subsection about the Kullback-Leibler divergence.

Shannon Entropy is a desirable diversity metric in a music setting due to the two main reasons. Firstly, it takes the relative frequency of a song being listened into the account directly via p_{ij}. Secondly, due to the definition of entropy, it also takes the number of different songs in the playlist into account. This is best seen with the equiprobable relative frequencies of the songs. Consider a person who listens to 4 different songs at the same frequency ($p_{ij} = 0.25,\ \forall j$) and thus obtains Shannon entropy of $D^{Sh}(S_i) = -log_2 0.25 = 2$. If the same user now starts listening to another new song and splits her time so that she again listens to all 5 songs at equal frequencies ($p_{ij} = 0.2,\ \forall j$), the new entropy becomes $D^{Sh}(S_i) = -log_2 0.2 \approx 2.3219$. This can be a desirable property if one would expect a user's taste in music to be more diverse if she listens to more songs as opposed to a user who listens to fewer songs. Note, this clear relationship breaks down once we move away from the uniform distribution case since an additional song might alter the underlying distribution substantially and thus re-weight the frequencies considerably. Yet, *in general*, when comparing two similar users with two similar playlist distributions, the user who is listening to more songs is expected to have a higher entropy provided none of the users are an extreme specialists. This is, however, not rigorous and will not always hold. Nonetheless, for some applications having a larger number of songs in the playlist might not correspond to a more *diverse* taste. Hence, this must be considered when opting for the Shannon entropy in any application.

To understand Shannon's entropy better, let's consider five different fictional users. Now, define P_i as the collection of the relative song frequencies p_{ij} of a user i, i.e $P_i = \{p_{i1}, p_{i2}, ..., p_{iN}\}$. Let the relative frequency distributions $P_i,\ i = 1, 2, 3, 4, 5$ to be as follows

$$
\begin{aligned}
P_1 &= \{0, 0.1, 0.5, 0.1, 0.3\} \\
P_2 &= \{0.6, 0.1, 0.1, 0.1, 0.1\} \\
P_3 &= \{0.1, 0.2, 0.4, 0.2, 0.1\} \\
P_4 &= \{0.2, 0.2, 0.2, 0.2, 0.2\} \\
P_5 &= \{0.7, 0.1, 0.1, 0.1, 0\}
\end{aligned}
$$

We label the user 5 as an extreme specialist, user 4 - an extreme generalist and users 1, 2 and 3 as in-between cases with user 2 and 1 having a tendency towards being a specialist and user 3 - towards being a generalist. The resulting entropies of each of these user playlists are:

$$D^{Sh}(S_1) \approx 1.68547$$
$$D^{Sh}(S_2) \approx 1.77095$$
$$D^{Sh}(S_3) \approx 2.12192$$
$$D^{Sh}(S_4) \approx 2.32192$$
$$D^{Sh}(S_5) \approx 1.35677$$

Thus, Shannon entropy managed to correctly differentiate between all five of these users with user 4 having the largest entropy and user 5 - the lowest. Users 1, 2 and 3 have accordingly increasing entropy values together with an increasing degree of generalist nature. In this sense, Shannon entropy suggests that user 1 is more specialist than the user 2. The "correctness" of this concussion depends solely on the application and it reflects the idea that the Shannon entropy tends to consider a user with a larger number of songs in the playlist as having a more diverse taste.

However, the issue with Shannon entropy is that it does not take the similarity between different songs played by the same user into the account. Hence, two users with an identical number of songs and listening frequencies would yield the same entropy regardless of how similar the songs each of them listen to are. To illustrate this, consider two users A and B. Suppose that A listens to 4 different songs from the same artist with frequencies p_{Aj}. User B also listens to 4 songs at the frequencies p_{Bj} such that the elements of P_B are identical to the elements of P_A ($P_B = P_A$). However, each of the songs listened by the user B come from different genres like Metal, R&B, Pop and Rap. We would consider user B to have a more diverse taste than user A, yet Shannon's entropy does not take that into account and thus would result in identical entropies $D^{Sh}(S_A) = D^{Sh}(S_B)$.

3.2 Kullback-Leibler divergence - relative entropy

Kullback-Leibler divergence, also called relative entropy, is an entropy-style measure which aims to compare the similarity of two distributions. In the discrete case, it can be expressed as

$$\sum_{i=1}^{N} P(\gamma_i) log_b\Big(\frac{P(\gamma_i)}{Q(\gamma_i)}\Big)$$

Commonly, $P(\gamma_i)$ is defined as an observed distribution and $Q(\gamma_i)$ as the theoretical one. Regarding the application in the music setting and using the notation introduced at the beginning of this chapter, Kullback-Leibler divergence can be expressed as

$$D^{KL}(S_i, Q_i) = \sum_{j \in S_i} p_{ij} log_2\left(\frac{p_{ij}}{q_{ij}}\right)$$

where q_{ij} stands for the probability that some user, chosen as the comparison to the user i,

listens to the song j.

Kullback-Leibler divergence $D^{KL}(S_i)$ is very similar to the Shannon entropy $D^{Sh}(S_i)$ and thus offers a similar interpretation with a slight twist. Instead of representing an expected surprise associated with seeing a song x_j being played, Kullback-Leibler divergence represents the *relative* surprise associated with the song x_j being played. In other words, relative entropy shows the expected surprise you experience after observing the user's i playlist distribution P_i when you expected it to be Q_i.

In general, once we observe the value of $D^{KL}(S_i) = 0$, it would signal that both distributions are identical. Consequently, the larger in absolute terms the value of $D^{KL}(S_i)$ is, the stronger is the evidence that the distributions differ. The interpretation of the relative entropy can be made more precise once the theoretical comparison distribution Q_i is defined. For example, we could define a synthetic listener who listens to all the songs in the data adjusted by their popularity, i.e. the more people have played the song at least once, the more popular the song is. In other words, a single user who has listened to a song x_j at least once increases the frequency of the song x_j being listened by a synthetic user by exactly one. Formally, we would define the elements of the playlist distribution Q_j of such a user as:

$$q_j^I = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j \in S_i\}}{\sum_{j=1}^{N}\sum_{i=1}^{n} \mathbb{1}\{x_j \in S_i\}}$$

Note, with this definition, we would have the same synthetic user for any user i, thus we can drop the subscript i and only define the playlist of the synthetic distribution as Q^I. Such definition of Q^I would label a song as popular if it is being listened by a large number of users.

Alternatively, one could simply define Q as the average frequency of a song being played across the users as:

$$q_j^{II} = \frac{\sum_{i=1}^{n} p_{ij}}{n}$$

where n is the number of users. In this case, the popularity is defined as the overall average frequency of a song being listened to. It is, however, possible that Q^{II} will be distorted by specific artists or genres which overall, are not popular among the wide public, yet have extremely loyal and active fandoms. Thus, q_j^{II} might appear to be less robust than q_j^I. It is, however, an empirical question whose conclusions will depend on the question of an application.

Regardless of the choice of q_j^I or q_j^{II}, the playlist distribution of the synthetic listener would resemble that of an extreme generalist with a very small probability of any particular song being listened (given a large enough sample size). Thus, larger absolute values of $D^{KL}(S_i)$ would indicate that a person resembles the specialist type more.

One interesting property of $D^{KL}(S_i)$ under the definition of Q as of the average user (under any definition of average) is that we control for the song popularity. Hence, we would expect an increase in absolute terms in the relative entropy when a user listens to a lot of relatively non-popular songs and a decrease in the relative entropy in the absolute terms if a user listens to a relatively popular

song more often than an average user does. Thus, in absolute terms larger value of $D^{KL}(S_i)$ can also indicate that a user has a less *mainstream* taste in music.

To illustrate the relative entropy let's consider the same example users we consider with Shannon entropy. Now, we define a synthetic user Q^I as:

$$Q^I = \{0.306, 0.194, 0.139, 0.194, 0.167\}$$

Note, here we treat the five example users with the playlist distributions P_1, P_2, P_3, P_4, P_5 as a subsample of the whole dataset with five different songs. P_5 and P_2 distributions can be defined as that of the specialists with a *mainstream* music taste, P_1 - a *non-mainstream* specialist, P_3 - a mild *non-mainstream* generalist and P_4 - an overall extreme generalist. Consequently, Kullback-Leibler divergence yields the following values for all of these distributions:

$$\begin{aligned} D^{KL}(S_1, Q^I) &\approx 0.98574 \\ D^{KL}(S_2, Q^I) &\approx 0.27015 \\ D^{KL}(S_3, Q^I) &\approx 0.39220 \\ D^{KL}(S_4, Q^I) &\approx 0.05188 \\ D^{KL}(S_5, Q^I) &\approx 0.59695 \end{aligned}$$

User 1 has the highest relative entropy value indicating that it has the most different distribution from the synthetic user. Meanwhile, user 5 has the second highest entropy score. This aligns with our reasoning since the user 5 is a mainstream specialist whilst user 1 is a non-mainstream specialist. Note, the non-mainstream nature of the user 1 outweighs the fact that without taking music popularity into the account, user 5 is a stronger generalist having most of her P_5 mass on x_1. Similarly, user number 2 and 3 have a very similar relative entropy score with $D^{KL}(S_2, Q^I) < D^{KL}(S_3, Q^I)$. Here, the more non-mainstream nature of the user 3 outweighs the specialist nature of the user 2 and hence, user 2 is deduced to be more similar to a synthetic user who could be labelled to be a mainstream generalist. This conclusion, however, is up to debate and presents one potential drawback of the Kullback-Leibler divergence with popularity measure. In general, $D^{KL}(S_i, Q)$ with Q as an "average" user will struggle to differentiate a mild mainstream specialist from the mild non-mainstream generalist. Lastly, user number 4 has the lowest relative entropy score labelling it as the most similar user to the mainstream generalist Q^I.

As we can see from this example, by looking at the $D^{KL}(S_i)$ score, it might be difficult to decide whether a user is more specialist or mainstream and which force is stronger in determining a specific Kullback-Leibler divergence score. This serves as a potential drawback. Yet, for some applications,

it could be overcome if we simply define $D^{KL}(S_i)$ as a similarity to a mainstream generalist.

One advantage of the relative entropy is that this measure is very flexible. By adjusting the definition of the theoretical distribution Q_i we can create diversity measures with different reference categories as well as take into account data particularities.

Lemma 1. *Shannon entropy, up to a constant, can be seen as a special case of Kullback-Leibler divergence with* $Q = Unif(0, N)$.

Proof. *Since* $Q = Unif(0, N)$*, all elements of* Q *can be defined as* $q_j = \frac{1}{N}$, $\forall j$

$$\begin{aligned} D^{KL}(S_i, Q) &= \sum_{j \in S_i} p_{ij} log_2\Big(\frac{p_{ij}}{q_j}\Big) \\ &= \sum_{j \in S_i} p_{ij}(log_2(p_{ij}) - log_2(q_j)) \\ &= \sum_{j \in S_i} p_{ij}(log_2(p_{ij}) - C) \\ &\propto -\sum_{j \in S_i} p_{ij} log_2(p_{ij}) \\ &= D^{Sh}(S_i) \end{aligned}$$

where C *is an arbitrary constant.*

From Lemma 1 it follows that Shannon entropy implicitly defines a reference theoretical user who listens to all the songs at the same frequency. This is because the synthetic user described would have the same q_j for all the songs and would only act as a constant in the Kullback-Leibler divergence. Hence, in this case, Shannon entropy would be proportional to Kullback-Leibler divergence and would give another interpretation of the relative expected surprise arising with a playlist distribution P_i when the uniform listening pattern was expected to be seen. As before, a higher absolute value of the relative entropy indicates more notable deviance from the extreme generalist case. The issue is, however, that the synthetic user who listens to all the songs uniformly at the same frequency is a very unrealistic, theoretical concept. Thus, it might not be useful for particular applications.

Finally, as with the Shannon entropy, Kullback-Leibler divergence does not take into the account the similarity of different songs and treats each individual song as completely unrelated to all the other songs even though their style, artist, genre might be very similar.

3.3 Gini-Simpson index

Gini-Simpson index represents a very simple idea similar to the one exploited by Shannon entropy. At the crux of the Gini-Simpson index is the idea that an arbitrary set is diverse if there is a low probability that when sampling two items with replacement from it, the probability of drawing two identical items is low. In music application, the Gini-Simpson index would denote the probability

that when sampling two songs from the same playlist with replacement, these songs are different. Mathematically it can be represented as:

$$D^G(S_i) = \sum_{j \in S_i} p_{ij}(1 - p_{ij}) = 1 - \sum_{j \in S_i} p_{ij}^2$$

This measure would give a flavour of how passionate a person is about some particular song and would directly reflect the frequency to which she is listening to it. Hence, a high Gini index is a sign of a generalist behaviour and low - that of a specialist. Similarly to Shannon entropy, Gini index is also affected by the number of items in the playlist, thus the more songs a user is listening to, the higher her Gini-Simpson index is *expected* to be. As before, it is not rigorous and will not always hold.

Interpretation, however, is different from that we have seen with entropy. Now, the Gini index has a meaning as a probability that two songs chosen from the playlist are different. This is a very simple and straightforwards interpretation and thus can be seen as an advantage of a Gini-Simpson index.

Consider the same example users which were discussed in the previous subsections. Their Gini-Simpson indices are:

$$\begin{aligned} D^G(S_1) &= 0.64 \\ D^G(S_2) &= 0.6 \\ D^G(S_3) &= 0.74 \\ D^G(S_4) &= 0.80 \\ D^G(S_5) &= 0.48 \end{aligned}$$

As before, Gini index manages to correctly differentiate different types of users with user 4 being the most diverse and user 5 - the least diverse. All other users fall in between these values accordingly to their degree of specialist nature. Nonetheless, the ordering of these users in the specialist-generalist spectrum is slightly different from that encountered with the Shannon entropy. Now, user 2 is perceived as more specialist than user 1. The distribution P_1 is slightly more evenly distributed than that of P_2 thus the conclusion reached by the Gini index is not surprising. Yet, this already shows how different measures can lead to different interpretations of the results with Gini index placing higher importance of evenness of the distribution that the Shannon entropy. Lastly, from the value of $D^G(S_4)$ we can see clearly the interpretation of the Gini score since the probability of drawing two different items with replacement from the uniform distribution is $1 - \frac{1}{N}$.

Another Gini-Simpson's index advantage over the Shannon entropy is that the Gini index does not require logarithmic transformation, thus Gini-Simpson improves the computational complexity by a logarithmic factor as compared to the Shannon entropy. Besides that, the Gini index does not

offer a significant improvement over Shannon entropy because it fails at the key drawback of the entropy too. As before, the Gini index does not take into account the relative similarity of songs and treats them as completely different entities.

3.4 Hill number

Hill number, which arose in the study of ecology [9], can be defined as the number of equally abundant (i.e. frequent) classes needed to yield the average proportional abundance obtained in the dataset (which do not have to be of an equal abundance). It is defined as the reciprocal of a weighted generalised mean of the abundance as

$$\left(\sum_{i=1}^{N} \gamma_i^m\right)^{\frac{1}{1-m}}$$

where $\gamma_i \in \Gamma$ is the relative abundance of a class i and m acts as a sensitivity adjusted term for abundant vs rare class. The choice of $m > 1$ exaggerates the weight of the abundant classes and $m < 1$ - of the rare classes. This means that with a higher value of m, distributions that are close to the uniform will get higher Hill number. Meanwhile, distributions with a substantial part of a mass on a few points will get lower Hill number with an increase in m.

Lemma 2. *For $m = 2$ case we implicitly define Hill number to measure the expected number of classes of equal abundance as based on the Gini-Simpsons coefficient with uniformly distributed classes [4].*

Proof. *Given the uniformly equally frequent classes ($\gamma_i = \frac{1}{N}$), the Gini-Simpsons index is defined as $\sum_{i=1}^{N} \frac{1}{N}\left(1 - \frac{1}{N}\right) = 1 - \frac{1}{N}$ and we thus solve for N as*

$$\begin{aligned} D^G(\Gamma) &= 1 - \frac{1}{N} \\ N &= \frac{1}{1 - D^G(\Gamma_i)} \\ N &= \frac{1}{\sum_{i=1}^{N} \gamma_i^2} \\ N &= \Big(\sum_{i=1}^{N} \gamma_i^2\Big)^{-1} \end{aligned}$$

With application to music, Hill number would measure how many different songs a person would have to listen to at equal frequencies to yield the average proportional pattern of the playlist observed in the dataset. It can be defined as follows:

$$D^H(S_i, m) = \left(\sum_{j \in S_i} p_{ij}^m\right)^{\frac{1}{1-m}}$$

Naturally, a larger value of $D^H(S_i, m)$ indicates a more diverse user taste. The value of m is likely to be set as greater than 1 so that we could focus less on single or rare instances of songs being listened. Nevertheless, the exact value of m is likely to be question-specific to the application and sometimes it might even be more desirable to give a higher weight to the songs which are being listened to less often.

Consider the same 5 users discussed before and let's compare Hill number values for three different values of $m = 0.5, 2, 3$.

$$
\begin{aligned}
D^H(S_1, 0.5) &\approx 3.561 \\
D^H(S_1, 2) &\approx 2.777 \\
D^H(S_1, 3) &\approx 2.548
\end{aligned}
$$

$$
\begin{aligned}
D^H(S_2, 0.5) &\approx 4.159 \\
D^H(S_2, 2) &\approx 2.5 \\
D^H(S_2, 3) &\approx 2.132
\end{aligned}
$$

$$
\begin{aligned}
D^H(S_3, 0.5) &\approx 4.662 \\
D^H(S_3, 2) &\approx 3.846 \\
D^H(S_3, 3) &\approx 3.492
\end{aligned}
$$

$$D^H(S_4, \forall m : m \neq 1) = 5$$

$$
\begin{aligned}
D^H(S_5, 0.5) &\approx 3.187 \\
D^H(S_5, 2) &\approx 1.923 \\
D^H(S_5, 3) &\approx 1.7
\end{aligned}
$$

Hill number with $m = 2$ and $m = 3$ ranks the users in the same order by their specialist-generalist nature. However, only for $m = 0.5$ $D^H(S_2, 0.5) > D^H(S_1, 0.5)$ suggesting that user 1 is more generalist than user 2. This can be explained by the fact that since $m < 1$ exaggerates the small fraction weights, it blows up the Hill number score of the user number 2, who has more small mass points (such as $p_{2j} = 0.1$) in P_2 than the user 1. Neither of the orderings of user 1 relative to user 2 is wrong or correct and it solely depends on the application. Note, the same ordering as with $D^H(\cdot, 0.5)$ was also reached by Shannon entropy, yet not the Gini index.

Furthermore, the value of m determines how similar these different users are deduced to be based on the number of less often played songs in the playlist. For instance, increasing value of m

affects the value of $D^H(S_3, m)$ relatively a little and causes no change in $D^H(X_4, m)$ at all. These two users are generalists and thus the frequency of different songs being listened is rather even. On the other hand, the value of m has a big impact on the value of $D^H(S_5, m)$. Since this user is a specialist, her playlist contains songs that are not being played often. Consequently, a large value of m renders those small frequency songs increasingly less important and as a result, the value of Hill number falls. Thus, with $m = 0.5$ all of the users seem to be more similar (all need 3 to 5 equally popular songs to yield the same playlist pattern) than with larger m values. For instance, with $m = 3$ (where user 1 and 2 requires 3 equally popular songs, user 3 - 4 equally popular song, user 4 - 5 and user 5 - only 2 equally frequently listened songs). This suggests that a $m > 1$ might be a more suitable choice for playlist diversity measure owing to its ability to differentiate between the users better.

Similarly to all the other measures discussed before, Hill number is affected by both: frequency of the song being played and the number of them. A difference, which might be advantageous to some applications, is that we can adjust how important less popular songs are relative to the popular ones in the context of a single user. Another advantage is the simple interpretation as "the number of songs being listened at the equal frequency required to yield the same score as the one observed". This, however, is arguably less intuitive than the ones Gini-Simpson index or Shannon entropy offers.

However, the drawback of not recognising the similarity between the songs remains.

3.5 Measure adaptation to the hierarchical data

In the previous sections, we have analysed the statistical measures and their behaviour when considering song-level data only, i.e. all values of p_{ij} were calculated based on a frequency to which a song x_j is being listened to by the user i. All of those statistical measures had one major drawback - they are not able to recognise the similarity between the songs. One partial remedy to this issue would be to consider hierarchical data - where songs are grouped according to their metadata. For example, we could group songs according to their artists, genres, tempo, language, positivity or other musical features. This, in turn, would alter the interpretation and expected behaviour of the statistical measures discussed so far. To see this, let's introduce some new notation.

As before, we define $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ to be a total collection of songs. Now, let $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$ represent a collection of sets representing different music categories (genres, artists, etc.) where $Y_\alpha = \{x_j \in \mathcal{X} | x_j \in \alpha\} \subseteq \mathcal{X}$ - a set of songs falling under category α where $\alpha = 1, 2, ..., k$. Note, elements of $\mathcal{Y}$ form a partition of the space $\mathcal{X}$. Thus, $y_{i\alpha} = S_i \cap Y_\alpha$ - a set of songs belonging to a category α which user i has in her playlist. Finally, $z_{i\alpha}$ defines the relative frequency of category α in the playlist of the user i:

$$z_{i\alpha} = \frac{f_{i\alpha}}{\sum_{\alpha=1}^{n} f_{i\alpha}}$$

where $f_{i\alpha}$ stands for the number of times songs from the category α were played by the user i. Now, user's i playlist distribution in terms of the selection of the categories α will be represented

by Z_i. With this notation is mind we can consider how the statistical measures introduced would be altered by this hierarchical data representation.

3.5.1 Shannon entropy

Under the new representation, Shannon entropy can be rewritten as

$$D^{Sh}(Z_i) = -\sum_{\alpha=1}^{k} z_{i\alpha} log_2(z_{i\alpha})$$

The interpretation remains that of an expected surprise associated with user's i playlist, yet the surprise is now defined in terms of a category, not an individual song. This, for some applications, might appear to be a more intuitive way of defining entropy if we assume that a person is more likely to keep playing the songs from the same category, i.e. genre or artist, rather than the same song itself. As before, the entropy is affected by both: the relative frequency of $y_{i\alpha}$ and the number of different categories α a user is exposed to.

As before with the song-level data, Shannon entropy is unable to differentiate how similar different categories Y_α are. Hence, we are failing at the same step as before. However, by grouping the songs according to these categories we, arguably, partly limit the problem if we can assume that treating each genre or artist as a separate independent entity is a weaker assumption than treating each song as a separate independent entity.

3.5.2 Kullback-Leibler divergence - relative entropy

Kullback-Leibler divergence, similarly, can be expressed to incorporate hierarchical representation as:

$$D^{KL}(Z_i, Q_i) = -\sum_{\alpha=1}^{k} z_{i\alpha} log_2 \left(\frac{z_{i\alpha}}{q_{i\alpha}} \right)$$

where $q_{i\alpha}$ stands for the probability that a theoretical user i listens to a song in the category α. The theoretical playlist distribution can be expressed in the similar ways as before - by taking the average frequency of a category α being played across the users (similar to q_j^{II}) or defining it in terms of the number of people that have played a song in category α at least once (similar to q_j^I), etc. To put it more formally, we can adapt previously defined theoretical user distributions as:

$$q_\alpha^I = \frac{\sum_{i=1}^{n} \mathbb{1}\{|y_{i\alpha}| \neq 0\}}{\sum_{\alpha=1}^{k} \sum_{i=1}^{n} \mathbb{1}\{|y_{i\alpha}| \neq 0\}}$$

$$q_\alpha^{II} = \frac{\sum_{\alpha=1}^{k} z_{i\alpha}}{n}$$

In turn, the interpretation remains as a relative expected surprise associated with the empirical user playlist distribution Z_i when the expected distribution was Q_i. As before, using the q_α^I and q_α^{II} to define the theoretical, we would have the same theoretical distribution for each user i, thus we

can omit the subscript and instead define the theoretical distribution Q. Again, using the definition of Q we can adjust the entropy measure by the popularity of the category which might appear to be useful for some applications.

As with Shannon entropy, with this hierarchical notation, we partly control for the fact that original Kullback-Leibler divergence could not determine the relative similarity between the songs.

3.5.3 Gini-Simpson index

With the hierarchical data, the index becomes

$$D^G(Z_i) = 1 - \sum_{\alpha=1}^{k} z_{i\alpha}^2$$

and, similarly to before, can be directly interpreted as the probability that two songs chosen from the user's playlist are of different category. Such Gini index can be seen as a measure of how category-loyal (be it genre-loyal or artist-loyal) a person is with a higher value of Gini index indicating a more diverse taste and hence weaker category-loyalty. As before, the simple interpretation of the Gini index and no logarithmic transformation required which makes the calculation of the score less (albeit marginaly) computationally demanding serves as the two advantages over the simple Shannon's entropy.

3.5.4 Hill number

Hill number with hierarchical music data becomes

$$D^H(Z_i, m) = \left(\sum_{\alpha=1}^{k} z_{i\alpha}^m \right)^{\frac{1}{1-m}}$$

where m controls the weight of rare versus abundant categories in calculating the Hill number. Interpretation follows as the expected number of equally frequently played music categories needed to yield the same pattern as the one observed in the user data Z_i. The value of m is again likely to be set as $m > 1$ to exaggerate the weight of more abundant categories and ignore the less abundant one. This would, in turn, help to control the effect of the cases where someone listens to the music with the other user's account, or the user listened to some songs with the purpose of exploration which, finally, appeared to be not of her taste, etc. The role of m with hierarchical data, arguably, is more intuitive. Even though a person is a fan of some genre, she might not listen to the same song within that genre so many times as to significantly increase the relative frequency of it. Yet, if we consider a song as part of a bigger category, every single time when a user listens to the song within a specific category is reflected in the overall relative listening frequency of that genre.

4 Geometrical similarity measures: theory

Whilst the statistical diversity measures can exploit the information regarding the relative frequency to which a user listens to the songs, artists, genres in her playlist as well as how many of the songs, artists, genres are in the playlists, they cannot determine *how similar* those songs, artists or genres are. This is a major drawback since we would not consider a person who listens to Ludwig van Beethoven, Greta Van Fleet, and Arianna Grande at the equal frequencies to be as diverse as the person who listens to Armin van Buuren, Pleasurekraft and Flume at the same equal frequencies. One way of dealing with this problem is to introduce the geometry into the problem as previous research did [1]. To begin with, the songs have to be embedded in the metric space so that a geometrical distance could be applied between different sets of songs. Thus, in this section, the embedding method is briefly explained followed by a theoretical evaluation of the Generalist-Specialist score [1] and a TS-SS score [8].

4.1 Embedding

To create a metric space for the songs, a Hierarchical Poisson Matrix Factorisation (HPF) [7] was implemented on the user-song playcount data[1]. HPF was created specifically for the recommender systems which use large datasets of user behaviour data where each user engaged in only a small subset of the whole array of items. This makes it a desirable embedding method for the *EchoNest* dataset that is being used here. In short, HPF models a user's i song j playcount data - f_{ji} as the realisation of a Poisson random variable with rate parameter $\lambda_{ji} = \langle \vec{\mu}_i, \vec{x}_j \rangle$, where $\vec{\mu}_i$ and $\vec{x}_j$ are K-dimensional user preference and song vectors. Beyond that, user preference and song vectors are assigned Gamma priors to encourage sparse data representation. This allows to extract K-dimensional vectors for each song, i.e. $\vec{x}_j \in \mathcal{X},\ j = 1, ..., N$ where the dot product $\langle \vec{x}_j, \vec{x}_z \rangle$ captures the similarity in terms of how likely the two songs are to be listened by the same user.

This embedding space has one flaw that the popular songs will be assigned longer vectors as opposed to the non-popular songs. This leads to the case where the popular songs are going to be likely to be listened together with a plethora of other songs regardless of how truly similar they are in terms of their music properties. To put it more formally the dot product between the popular song $\vec{x}_p$ and any other song $\vec{x}_j$ is likely to be higher than the dot product between a non-popular song $\vec{x}_{np}$ and the same song $\vec{x}_j$. Nonetheless, normalising the song vector to have a unit norm can help to mitigate this issue. The dimension of the vectors was set to $K = 100$ due to the superior testing performance as opposed to a lower-dimensional embedding.

4.2 Generalist-Specialist score

The Generalist-Specialist score was introduced as a geometrical diversity measure and initially evaluated on the GitHub, Reddit and Spotify user data [19, 1]. At its core, the Generalist-Specialist

[1]The embedding was implemented and the vectors generated by Spotify using the Python package introduced in GitHub [6].

score is the cosine-similarity measure where for each user, her diversity score is calculated by taking the weighted average of the cosine-similarity between each song $\vec{x}_j$ in her playlist and the user's playlist centroid $\vec{\mu}_i$. The centroid is the average of the song vectors in the user's playlist weighted by the playcount f_{ij} of each song, thus it could be interpreted as the *average song* or *user preference*. Using the notation introduced before, the *user preference* vector of each user's i is defined as

$$\vec{\mu}_i = \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} f_{ij} \vec{x_j}$$

and consequently, the Generalist-Specialist score for the user i as:

$$D^{GS}(S_i) = \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} f_{ij} \frac{\langle \vec{x}_j, \vec{\mu_i} \rangle}{\| \vec{x}_j \| \, \| \vec{\mu}_i \|}$$

where $\| . \|$ denotes the Euclidean norm. At the extremes of $D^{GS}(S_i) = 0$ a user is an extreme generalist and at $D^{GS}(S_i) = 1$ - an extreme specialist. In theory, the cosine-similarity and the Generalist-Specialist score could range from -1 to 1, yet in our case due to the nature of the embedding and the fact that the count data is non-negative, the maximum angle between two songs can be 90 degrees and thus the range of cosine similarity, and thus the $D^{GS}(S_i)$, is reduced to $[0, 1]$.

Stemming from the fact that the norm of the embedded song vectors $\vec{x}_j$ correlates with the popularity of the song, it is better to normalise the vectors. Otherwise, the dot product between two songs would be larger when defined with any popular song as opposed to the less popular song despite of their true similarity. The normalisation of the song vectors exposes a new property of the Generalist-Specialist score as described in the Lemma 3 below.

Lemma 3. *Once the song vectors $\vec{x}_j$ are normalised to have a unit Euclidean norm, $D^{GS}(S_i)$ is equivalent to the Euclidean norm of the user's playlist centroid vector $\vec{\mu}_i$.*

Proof. *Assuming that $\forall j \;\; \| \vec{x}_j \| = 1$ and using the identities defined before we can rewrite $D^{GS}(S_i)$ as*

$$\begin{aligned}
D^{GS}(S_i) &= \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} f_{ij} \frac{\langle \vec{x}_j, \vec{\mu_i} \rangle}{\| \vec{x}_j \| \, \| \vec{\mu}_i \|} \\
&= \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} f_{ij} \frac{\langle \vec{x}_j, \vec{\mu_i} \rangle}{\| \vec{\mu}_i \|} \\
&= \frac{1}{\| \vec{\mu}_i \|} \left\langle \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} f_{ij} \vec{x}_j, \vec{\mu_i} \right\rangle \\
&= \frac{\langle \vec{\mu_i}, \vec{\mu_i} \rangle}{\| \vec{\mu}_i \|} \\
&= \frac{\| \vec{\mu_i} \|^2}{\| \vec{\mu}_i \|} \\
&= \| \vec{\mu}_i \|
\end{aligned}$$

The lemma above indicates that the Generalist-Specialist defined on the normalised song vectors not only helps to omit the implicitly defined popularity from the songs but would also save unnecessary computations since it can be simply be expressed as a Euclidean norm of the $\vec{\mu}_i$

Regarding the interpretation, the cosine similarity measures an angle between two vectors. Thus, similarly, the Generalist-Specialist score measures a weighted average angle between each song $\vec{x}_j$ in a user's i playlists and the *user preference* vector $\vec{\mu}_i$. This, however, is hard to translate to any other more meaningful interpretation as the ones we had in the context of the statistical measures. This stands as one potential drawback of the Generalist-Specialist score.

The other drawback of the cosine similarity and thus the Generalist-Specialist score stems from the fact that these measures only take the angle between two vectors into the account ignoring the magnitude of the vectors. This might be problematic in the cases when some songs are in parallel from the perspective of the origin.

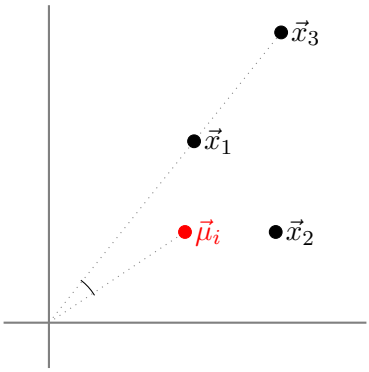


Figure 1: Generalist-Specialist score issues

Consider the Figure 1 where the $\vec{\mu}_i$ represents the *user preference* vector of the user i and the points $\vec{x}_1$, $\vec{x}_2$ and $\vec{x}_3$, a two-dimensional song vectors. From the perspective of the Generalist-Specialist score, there is absolutely no difference between the song $\vec{x}_1$ and $\vec{x}_3$ since the angle between either of these songs and the *user preference* vector is the same. However, this is wrong because the song $\vec{x}_3$ is clearly further away than the song $\vec{x}_1$ from the *user preference* vector in terms of the magnitude. This issue persists if the data is normalised to have a unit Euclidean norm. Normalisation forces all the points to lie on the n-dimensional unit sphere, thus the vectors $\vec{x}_1$ and $\vec{x}_3$ would simply overlap and be considered as the same point. It must be noted that the normalisation does not alter the angle between the vectors and thus no damage to the Generalist-Specialist score would occur. The situation illustrated in Figure 1 might not be an issue if such situations where the songs are in parallel from the perspective of the origin do not happen in the embedded metric space. This is an empirical question which will be evaluated in the Discussion section.

4.3 TS-SS

Motivated by the cosine-similarity problems as well as the weaknesses of the other geometrical measures such as the Euclidean distances, a new hybrid distance measure, called TS-SS, has been proposed [8]. TS-SS directly tackles the issue that the most commonly used geometrical distance measures exploit either the angle between the vectors or their relative magnitudes and not both. Thus, TS-SS aims to combine both of these attributes. In short, TS-SS is a simple multiplication of two other purely geometrical measures - Triangle's Area Similarity (TS) and the Sector's Similarity (SS). The Triangle Similarity simply computes the triangle area that two vectors create between them using the trigonometric identities as

$$TS(\vec{A}, \vec{B}) = \frac{\|\vec{A}\| \cdot \|\vec{B}\| \cdot sin(\theta')}{2}$$

where θ' is an angle between the vectors and can be derived using the cosine-similarity as

$$cos(\theta') = \frac{\langle \vec{A}, \vec{B} \rangle}{\|\vec{A}\| \, \|\vec{B}\|}$$

By itself, $TS(\vec{A}, \vec{B})$ has a drawback that it ignores the differences in the vector magnitudes and thus can lead to the wrong conclusions [8], thus another component - Sector's Area Similarity (SS) - is employed to correct for that. SS is constructed using the difference in magnitudes between the two vectors, Euclidean distance between the vectors and their angle. These measures are then manipulated using the area of a circle equation. More formally, let $MD(\vec{A}, \vec{B})$ to denote the vector magnitude difference as:

$$MD(\vec{A}, \vec{B}) = \big|\, \|\vec{A}\| - \|\vec{B}\| \big|$$

Then, denote the Euclidean distance between the vectors as $ED(\vec{A}, \vec{B}) = \|\vec{A} - \vec{B}\|$. The $SS(\vec{A}, \vec{B})$ can be defined using the area of a circle identity as

$$SS(\vec{A}, \vec{B}) = (ED(\vec{A}, \vec{B}) + MD(\vec{A}, \vec{B}))^2 \cdot \frac{\pi \cdot \theta'}{360}$$

where θ' is again the size of the angle between the two vectors.

Multiplying both, $TS(\vec{A}, \vec{B})$ and $SS(\vec{A}, \vec{B})$ leads to the TS-SS$(\vec{A}, \vec{B})$ score. Visually, the area TS-SS$(\vec{A}, \vec{B})$ score calculates can be represented as the grey shaded area in the Figure 2 borrowed from the original paper [8].

Adopting our notation we can implement TS-SS using similar methods as those proposed with the Generalist-Specialist score. In other words, each user's score is a weighted average of the TS-SS between each song x_j in their playlist and the *user preference* vector μ_i. The resulting score can be expressed as

$$D^{TS-SS}(S_i) = \frac{1}{\sum_{j \in S_i} f_{ij}} \sum_{j \in S_i} \Big(f_{ij} \cdot TS(\vec{x}_j, \vec{\mu}_i) \cdot SS(\vec{x}_j, \vec{\mu}_i) \Big)$$

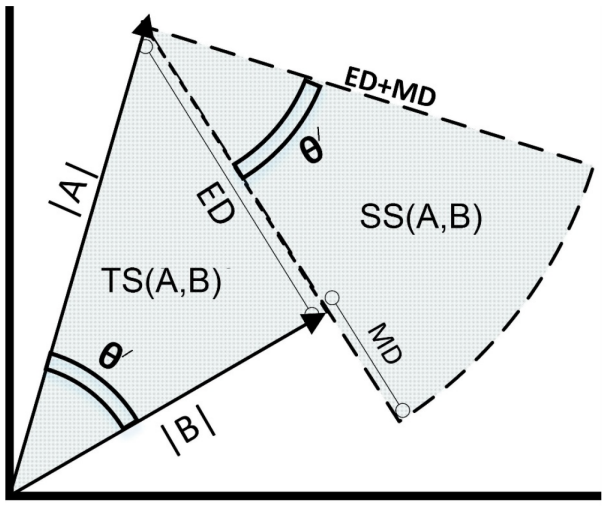


Figure 2: Visual TS-SS score representation

Even more so than with the $D^{GS}(S_i)$ score, the interpretation of $D^{TS-SS}(S_i)$ is awkward and stands as the average weighted TS-SS score which in itself is rather complicated. Thus, this presents one drawback of the measure.

Another drawback described by Mijalovic et al. [12] is that TS-SS score seems to be more prone to the *curse of dimensionality* than the cosine similarity when evaluated from the point of the classification as opposed to clustering as was done in the original paper. Bearing in mind that the embedded song vectors we are using a 100-dimensional, this might appear to be a problem.

4.3.1 TS-SS and normalisation

Normalisation has a significant effect on the measures that take the magnitude of the vector into account and literature advises against the normalisation for such measures [8, 5, 17, 18]. On the other hand, the embedding created by the Hierarchical Poisson Factorisation assigns longer vectors to more popular songs. It is thus likely that if a user listens to one popular song, her $D^{TS-SS}(S_i)$ is going to be higher purely due to the longer vector of the popular song. If we assume that the people who listen to at least one popular songs are likely to be more generalists, this might not be a problem. But if we are not ready to assume this, normalisation of the data would help to overcome this issue. However, normalisation comes with its cost and leads to the same problem as with the Generalist-Specialist score. Normalising the data to have a unit Euclidean norm forces the vectors to lie on the n-dimensional unit sphere. Thus, if two vectors share the same direction as in Figure 1, once the data is normalised, the vectors $\vec{x}_1$ and $\vec{x}_3$ would overlap and thus would be considered as identical. This would mean that we cannot distinguish these two songs and we are facing the same problem as with the Generalist-Specialist score.

However, the normalisation does not fully defeat the purpose of the $D^{TS-SS}(S_i)$ score. The *user preference* vector $\vec{\mu}_i$ will not necessary have a unit Euclidean norm even if it is calculated on

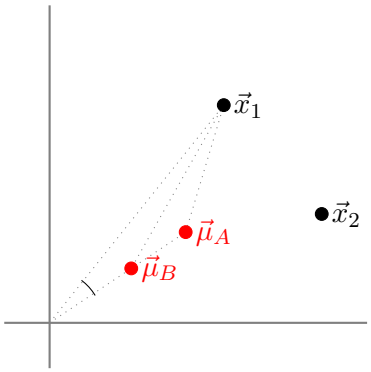


Figure 3: TS-SS score with the normalised data

the unit-norm song vectors $\vec{x}_j$. Thus, $D^{TS-SS}(S_i)$ would still enable us to take into the account the magnitude of the $\vec{\mu}_i$ and thus how far it is from the songs in the playlist whilst potentially gaining some more accuracy. To illustrate this, consider the subsample of a fictional user playlist in the Figure 3. For simplicity, we assume the embedding resulted in 2-dimensional song vectors $\vec{x}_j$ which due to the normalisation lie on the unit circle. We assume the two songs illustrated are only a subsample of the whole playlist with two theoretical *user preference* vectors $\vec{\mu}_A$ and $\vec{\mu}_B$. Regardless of the true *user preference* vector, the $D^{GS}(S_i)$ would assign an identical score to the diversity of this playlist subsample. This, however, is not the case with the $D^{TS-SS}(S_i)$ when implemented on the normalised data since the magnitudes of the *user preference* vectors differ. As a result, if a true *user preference* vector of this fictional user was $\vec{\mu}_B$, taking into account this subsample, she would be considered more generalist than if the true *user preference* vector was $\vec{\mu}_A$. This aligns with the intuition since the vector $\vec{\mu}_A$ is closer to both songs in terms of their magnitude.

Overall, implementing $D^{TS-SS}(S_i)$ score on both normalised and non-normalised data using our embedding comes with some drawbacks. It is therefore important to understand the limitations of both implementations or to consider a different embedding method which would not be prone to the issues that required normalisation.

5 Data

Data has been obtained from the publicly available Million Song Dataset. The Million Song Dataset has several different metadata datasets which could be used to enrich the analysis presented here. Yet, for the purposes of this project, primarily, the Echo Nest Taste Profile subset (referred to as the `triplet` data) together with supplementary data containing the track IDs and artists (referred to as the `bridge` data) were used [2].

The original `triplet` dataset contains $1,019,318$ unique users with $384,546$ unique `SongID` values and the respective counts of each song being played by each user (presented by the `Count` variable). The `Count` variable exhibits a substantial right skew (Table 1) with the 75% quantile

being 3 and the maximum value - 9,667. This indicates the presence of some users who listen to a few songs incredibly often. To be more precise, there are exactly 62 users who listened to some particular song more than 1,000 times. For our purposes of checking the quality and features of various diversity measure, these outliers are of no importance, hence we leave them in the dataset. Histogram of the `Count` data conditional on it being less than 40 is presented in Figure 4. This still exhibits a substantial right skew which is not surprising knowing that half of the `Count` values take value 1.

	`Count`
Count	483,73,586
Mean	2.8668
Stand. dev.	6.4377
Min	1
25 %	1
50 %	1
75 %	3
Max	9,667

Table 1: `Count` summary statistics

The bridge data, which maps `SongID` to `TrackID`, is constituted by exactly 1 million unique `TrackID` values together with the information about the artist (72,665 unique artists) and the song name each `TrackID` corresponds to. However, the number of the unique `SongID` values was only 999,056 indicating that 944 `SongID` values were mapped to more than one `TrackID` value. Having a closer look at those songs, it appeared that most of them were either a collaboration among several artists (1), had some spelling mistakes (2) or were assigned multiple `TrackID` values with no apparent reason (3). For example, the first case is illustrated by the Oasis And Friends Including Johnny Depp song *Fade Away* which under the different `TrackID` were instead labelled simply as Oasis song *Fade Away*. An example of the second case is Lily Allen whose name under the different `TrackID` was falsely written as *Lilly Allen*. The third case is the most suspicious one since different `TrackID` values are assigned to seemingly identical pieces of music, e.g Aerosmith song *Pink* has three different `TrackID` values, yet only one `SongID` value. This might be the case if there are different recordings of the same song which all receive their own `TrackID`. Bearing all the above points in mind, it was decided to simply discard duplicate values which resulted in the one-to-one mapping of `TrackID` to `SongID` with 999,056 unique instances of each. The duplicate removal also resulted in a loss of 13 unique `Artist` values. The list of them is available in the Appendix A.

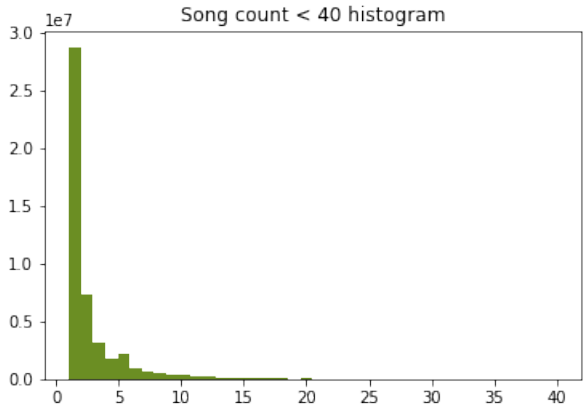


Figure 4: Song `Count` (less than 40) histogram

Merging the `triplet` data with the `bridge` data resulted in a combined dataset with the same 384,546 unique songs and 1,019,318 unique users as in the original triple dataset. Hence, no users has been lost. The merged data contains 6 different variables, namely, `UserID`, `SongID`, `TrackID`, `Artist`, `Song` and `Count`.

5.1 Subsample

To speed up the code running time, it was decided to work with the subsample of 10,000 users (seed - 123) which constitutes approximately 9.8% of the users in the original data. Bearing in mind that the purpose of this project is to analyse different similarity measures and not to derive accurate conclusions from inference using them, the usage of subsample is considered as an appropriate strategy. The resulting dataset has 10,000 unique users and 105,296 unique songs. To assess the quality of the subsample in terms of its ability to represent the original data `Count` variable summary statistics are analysed and presented in Table 2 and Figure 5.

	`Count`
Count	484,499
Mean	2.8854
Stand. dev.	6.5195
Min	1
25%	1
50%	1
75%	3
Max	1,222

Table 2: Subsample `Count` summary statistics

As in the original `triplet` data, half of the `Count` values in the subsample dataset take value 1. Furthermore, the mean and standard deviations in both dataset are very similar together with the same right skew underlying the data. Hence, we can assume that the subsample constitutes a representative sample of the original `triplet` data.

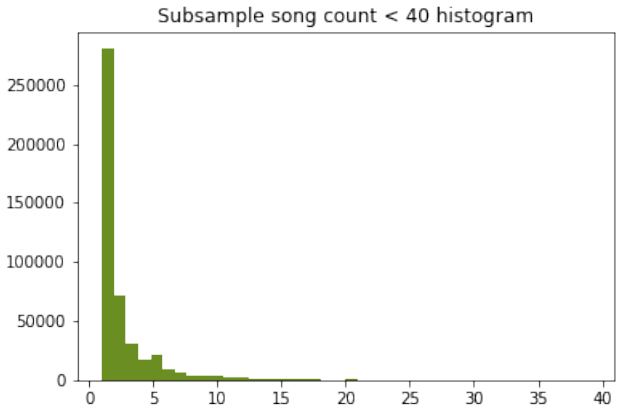


Figure 5: Subsample song `Count` (less than 40) histogram

Note, among the supplementary metadata datasets of the Million Song Dataset, there is also a dataset that maps songs to their genres called *tagtraum genre annotations for the Million Song Dataset* [14]. However, after merging this genre dataset with the `triplet` dataset, only 45% of the unique songs with 99.9% of the unique users remain. It thus means that on average each user loses a lot of information (songs) about her music listening pattern and thus we are left with very curtailed

information which could in turn, severely impact the values of the diversity measures. Furthermore, as stated in the article outlying the algorithm of genre annotation [14], some genres, such as Blues or Vocal, are not being separated individually and are instead considered as part of the Rock or Jazz genre respectively. Thus, in order to not introduce any more error than is necessary, it was decided to abstain from using the genre data. Consequently, for the hierarchical model, only the categorisation in terms of `Artist` variable as found in the `bridge` dataset will be used.

6 Evaluation

After implementing the diversity measures outlined before, the evaluation strategy needs to be adopted. In general, it is difficult to create an objective quantitative measure to evaluate the diversity scores, thus a more qualitative approach is undertaken. As a result, the diversity scores will be examined from several different perspectives and the conclusions derived consequently. The evaluation method consists of the following aspects:

1. *Uniqueness*: the purpose of this evaluation is to determine how well the measure is able to recognise minor differences between the users [8]. A desirable diversity measure would be able to assign a different value to as many users as possible. Thus, the *uniqueness* is defined as a fraction of unique score values out of all the score value each similarity measure yields.

2. *Relationships with the attributes*: the idea behind examining the dependence of the diversity measures on some specific attributes for the evaluation is that we expect the diversity measures to depend on the attributes in a specific way according to their theoretical properties. For instance, stemming from the previous work on the Spotify users behaviour [1], it is expected that more generalists users would be slightly more active on the platform. Following similar logic, it is likely that on average, a generalist user would listen to more artists and/or songs than their playlist specialist counterparts. In addition, Kullback-Leibler divergence with the distribution Q^I representing that of a mainstream generalist is expected to exhibit the strongest dependence on the number of popular songs in a user's playlist. If a measure violates these assumptions, it could be a sign of its drawback. The attributes chosen for the evaluation are as follows:

    - *Total user activity* defined as the total number of times a person played all the songs in her playlist. More formally, for each user i:

    $$\textit{Total user activity} = \sum_{j \in S_i} f_{ij}$$

    - *Coverage* defined as the number of distinct items (be it songs or artists). Mathematically,

for each user i:

$$Distinct\ songs = |S_i|$$
$$Distinct\ artists = \sum_{\alpha=1}^{k} \{|y_{i\alpha}| \neq 0\}$$

where $\alpha = 1, 2..., k$ denotes artists.

- *Popularity* is defined as the number of top 500 most popular songs that each user i listens to. The top 500 of the most popular songs are simply defined as the top 500 of the songs with the highest q_j^I count value described earlier. Let G denoted the set of those top 500 of the songs with the highest q_j^I count value. Then, for each user i, *Popularity* is defined as:

$$Popularity = |G \cap S_i|$$

3. *Concordance analysis*: the cases where the distinct measures do not agree will be evaluated to assess the quality and special features of them. Kendall's rank correlation coefficient, further referred to as Kendall's τ, alongside the manual check of the users with extremely different rank values will be evaluated. The Kendall's τ coefficient is defined as [13]

$$\tau = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\binom{n}{2}}$$

and takes the values in the range $[-1, 1]$. If two scores are identical, the Kendall's τ will take the value 1 and if the two scores are the exact opposites, $\tau = -1$. Thus any values in between the extremes show the degree of concordance or discordance.

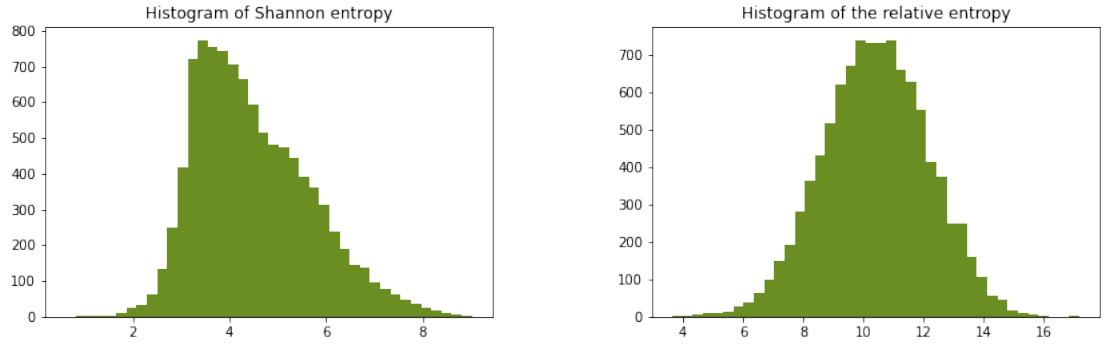
Furthermore, the ranking of all the users by all the scores will be constructed and the rank differences between the scores evaluated. Ranks as a metric are chosen due to their robustness and simplicity of interpretation.

4. *Specific cases*: some specific users will be chosen and their relative diversity rankings compared in the context of all the diversity measures to assess the differences between the diversity scores and how they label various types of user behaviour.

7 Discussion: Statistical Diversity Measures

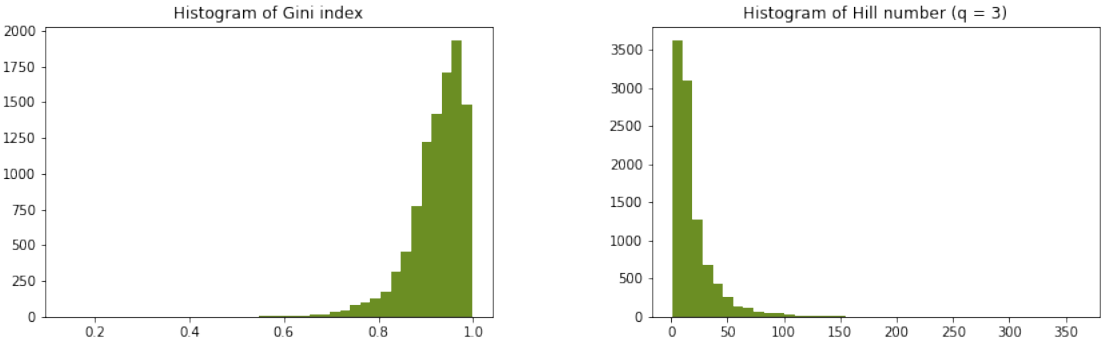
Implementing the statistical diversity measures on the $10,000$ user subsample yielded the following histograms of the diversity scores (Figure 6).

Shannon entropy (Figure 6 (a)) offers to some extent a round and right-skewed distribution with the mean and standard deviation taking values of 4.495 and 1.205 respectively. Minimum and maximum values are 0.608 and 9.016 indicating a good amount of spread of the score. This is



(a) Shannon entropy

(b) Relative entropy



(c) Gini index

(d) Hill number

Figure 6: Histograms of the statistical diversity scores

a desirable quality since it makes it more likely that the Shannon entropy is able to differentiate between similar user playlist distributions.

Kullback-Leibler divergence (Figure 6 (b)), or relative entropy, offers an even rounder distribution with mean 10.413 and standard deviation of 1.775. Minimum and maximum values are 3.645 and 17.18 respectively which also indicates a good amount of spread of the score values. The distribution resembles that of a Gaussian, thus it was decided to evaluate the normality and construct the confidence interval for the relative entropy. Bearing in mind that the scores are implemented on the subsample of the data, 95% confidence interval would indicate where the 95% of the score values are expected to lie given another sample from the same dataset or given yet unseen user data. The corresponding Q-Q plot can be found in Appendix B using which we conclude that the distribution of the relative entropy score can be approximated by a Gaussian distribution. Consequently, the confidence interval was found to be $[6.934, 13.892]$.

The spread, however, is rather poor in the context of the Gini index. Whilst the mean of the Gini index is 0.921, the standard deviation is only 0.066. This, together with the histogram in Figure 6 (c), shows that the Gini score values are very concentrated. Even though the minimum

and maximum values are 0.138 and 0.998, as much as 75% of the values are in the range from 0.896 to 0.998. This means that for the 75% of the users, probability of picking two different songs when sampling from the user's playlist with replacement is roughly at least 90%. From the perspective of the interpretation, it could be argued that all of these users ought to be labelled as specialists. One could try to re-scale the Gini index to offer a higher spread of the score, yet this would be done at a cost of the simple and straightforward interpretation of the Gini index. Thus, it seems that the Gini index is inferior to the Shannon and relative entropy in this specific context of the spread of the score values.

Regarding the Hill number, a positive value of m was chosen to exaggerate the weights of the more abundant songs. It was decided to proceed with $m = 3$ instead of $m = 2$ because in the latter case, the resulting Hill numbers would be perfectly correlated with the Gini index due to reasons explained in section 3.4. Thus, it would only give a different interpretation, but not the conclusions *per se*.

Hill number histogram in Figure 6 (d), similar to the Gini index in Figure 6 (c) shows a very skewed distribution. This time, the mean value is 21.373 and standard deviation 24.526 with minimum and maximum values of 1.119 and 362.329. Contrary to the Gini index case, the skew of the Hill number is not a substantial drawback due to the range of the values covered by the Hill number as well as the different interpretation. This time, 75% of the Hill number score takes the value less than or equal to 24.575. This implies that at most 24.575 equally abundant songs are needed to yield the average proportional patter of the playlist observed among 75% of the users. This is better coverage than the one in the Gini index case and thus, from this perspective, Hill number can be considered as a superior measurement to the Gini index in the playlist diversity measurement case. However, the skew might still appear to be a drawback if the Hill number fails to differentiate similar users.

The shape of the histograms of the statistical similarity measures directly translates into the *Uniqueness* (Table 3). Whilst relative entropy manages to assign a different diversity score to all of the user playlists, as many as a quarter of the users are sharing a score with at least one other user according to the Gini index. This strengthens the idea that the Gini index is an inferior diversity measure to the other three measures in the context of music diversity estimation. Shannon entropy and Hill number appeared to take the middle ground between the relative entropy and the Gini index with Shannon entropy being able to differentiate 1.83 percentage points more of the users than the Hill number.

	Shannon entropy	Relative entropy	Gini index	Hill number
Uniqueness	87.13%	100%	73.79%	85.3%

Table 3: Uniqueness Results

7.1 Dependence on the attributes

To facilitate the analysis of the attributes, scatterplots between the attributes and the diversity measures are plotted. To help to identify regions of overlapping points and thus high density, a colour gradient is employed with yellow regions indicating regions of high concentration.

Furthermore, in order to summarise a non-linear sparse data clouds, Locally Weighted Scatterplot Smoothing (LOWESS) [3, 15] is implemented. LOWESS is an iterative algorithm which works by taking the fraction of the closest points to the actual independent and dependent variables (x_i, y_i) based on their independent variable x. Then, LOWESS locally fits a low polynomial to the subset of the data and estimates y_i^{LW} using weighted linear regression. The weights take a traditional tri-cube weight function form as:

$$w(x) = (1 - |d|^3)^3$$

where d is a distance from a given data point x to the point where the curve is being fitted. The algorithm iterates until $w(x) < 1$. A fraction of 20% was chosen for this specific case. For the linear relationships - as is the case between the Hill number and all four attributes - a simple linear regression was used.

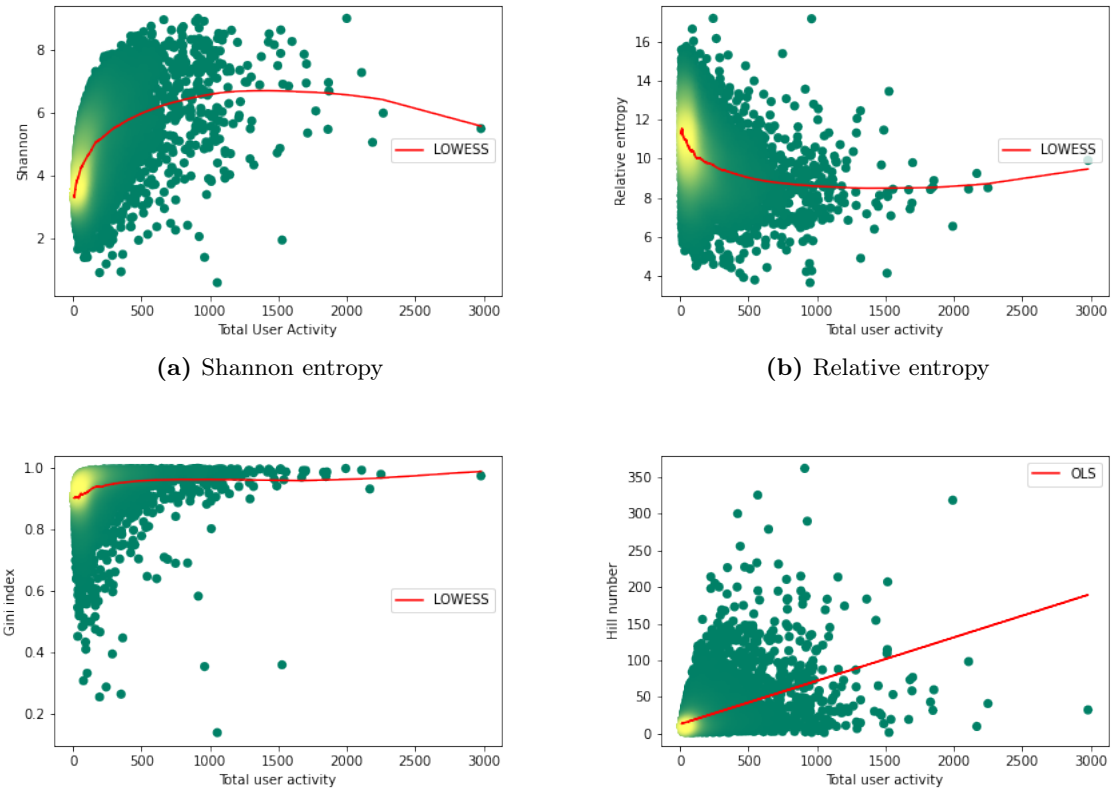
7.1.1 Total User Activity

Total user activity was defined as a sum of all the values of a `Count` variable given a specific user. In other words, it is a total playlist's playcount. The histogram of the variable indicates a right-skewed distribution with mean 136.493 and standard deviation of 185.97. The histogram is available in Appendix C.

Starting with the Shannon entropy, we see from Figure 7 (a) that there is a very weak sign of a positive dependence on the *Total User Activity*. This reflects the fact that the Shannon entropy takes into account the *relative frequency*, i.e. the probability, of a song being played rather than the absolute count. However, it must be noted that the low values of *Total User Activity* cover close to the full range of the Shannon entropy values whilst the large values do not. This creates a *triangle-relationship* when large values of the *Total user activity* suggest a high value of Shannon entropy (and thus generalist label), yet the reverse conclusion cannot be made.

The dependence becomes even weaker and almost non-existent when taking the Kullback-Leibler divergence into account. This is expected due to the *relative* nature of the $D^{KL}(S_i, Q^I)$ when the user with playlist distribution S_i is compared to the user with playlist distribution Q^I in terms of their relative song listening frequencies.

Gini index can be deduced to have close to none dependence on the *Total user activity* with LOWESS curve stemming vertically as indicated in Figure 7 (c). Thus, LOWESS curve here was implemented only for the consistency reasons and otherwise is not useful in determining the relationship between the Gini index and the attributes. One can see that the low *Total user activity* values cover almost the full range of the Gini index values. At the same time, from Figure 7 we



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 7: Diversity measures vs Total User Activity

can see that the triangle relationship is the strongest in the context of the Gini index. Thus, the high values of *Total user activity* can give a strong suggestion of a high Gini index value (and thus a generalist label), yet the reverse claim cannot be made.

Lastly, the Hill number vs *Total User Activity* scatterplot (Figure 7 (d)) indicates some weak presence of a linear relationship between the measures. The relationship indicates that to some (albeit weak) extent, higher user activity is positively associated with a high Hill number and thus the generalist label. Note, due to the uneven spread of the points in Figure 7 (d), the homogeneity of variance assumption essential for the ordinary linear regression estimation cannot be met. Thus, the OLS line is plotted only to indicate some direction of the relationship and is not meant to be analysed statistically any further.

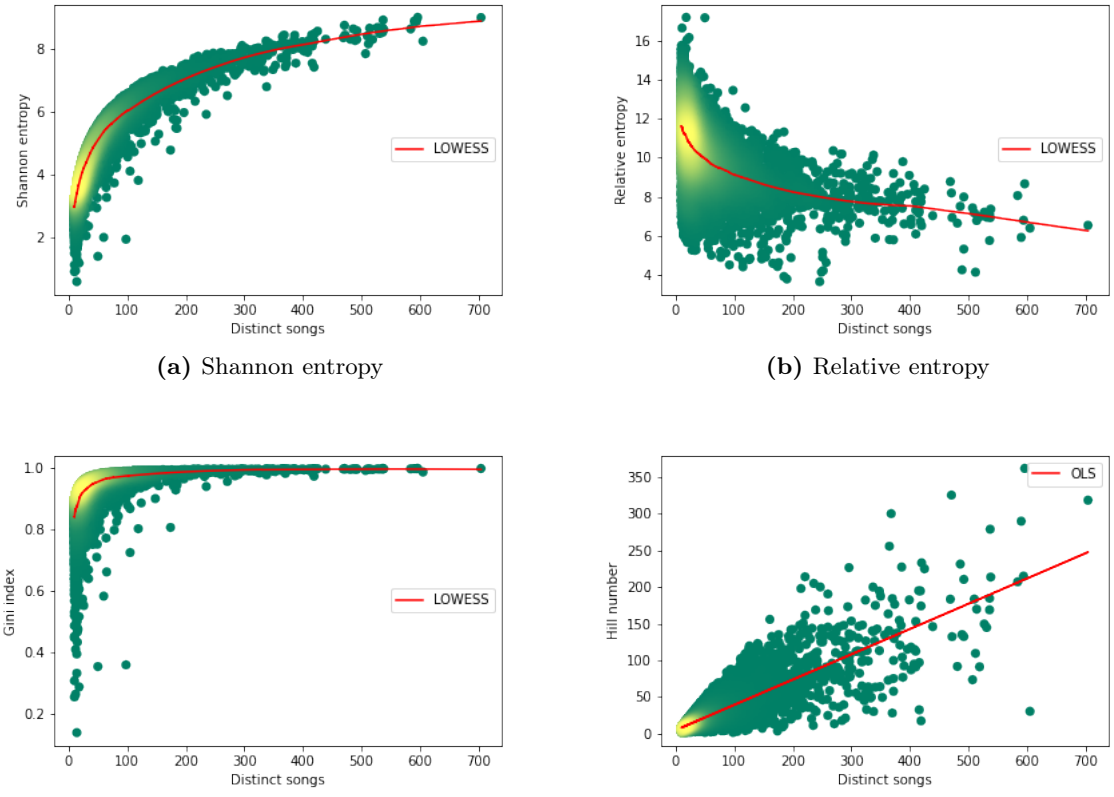
In general, none of the four statistical measures exhibits a clear dependence on the *Total User Activity*. This, arguably, suggests that those measures incorporate information beyond that. Another positive finding is that, albeit weak, all of the existing correlations signal the same conclusion - active users are more likely to be generalists than specialists. This, in turn, aligns with intuition and the findings in previous literature and gives some evidence that the four statistical measures

are appropriate diversity scores.

7.1.2 Coverage: distinct songs and artists

Distinct songs variable was defined as a total number of unique `SongID` values given each individual user. *Distinct artists*, similarly, was defined as the number of unique `Artist` values given each individual user. The distribution of the coverage in terms of the distinct songs is skewed to the right with a mean of 47.236 and standard deviation of 57.875. The distribution of the coverage in terms of the distinct artists played is similar, yet with the mean of 29.057 and standard deviation of 30.143. Both histograms are available in Appendix C.

Overall, the scatterplots of the coverage variable versus the statistical distance measures were very similar regardless of which coverage variable was used, hence here we illustrate the scatterplots with the *Distinct songs* variable owing to the stronger relationships presented. The scatterplot with the *Distinct artists* is available in the Appendix D. Nonetheless, the dependence of the statistical diversity measures on both coverage variables will be discussed in parallel.



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 8: Diversity measures vs Distinct Songs

The relationships between the coverage variables and the statistical diversity measures are

stronger than those encountered with the *Total user activity*. Shannon entropy seemingly exhibits the strongest dependence on both *Distinct songs* and *Distinct artists* variables followed by the Hill number (Figure 8). Note, for both of those measures, the dependence seems to be less noisy and thus stronger with the *Distinct song* variable which is expected knowing that each new song contributes an addition of a specific factor to the score. This time, the OLS in the Hill number scatterplot is more informative than with the *Total user activity* owing to the more compact spread of the points. Both scores suggest that users with a higher number of distinct songs and artists in their playlist tend to yield higher Shannon entropy or Hill number score and thus are perceived to be more generalists.

The relative entropy (Figure 8 (b)) now exhibits a clearer and stronger negative relationship with the number of distinct songs and artists than with the *Total user activity*. However, the relationship seems to be weaker than that offered by the Shannon entropy and the Hill number. This indicates that the relative entropy is different from those two measures, arguably due to the different benchmark category that the user playlists are compared to. This time the triangle-pattern is more pronounced and suggests that users with high coverage are likely to yield a low relative entropy score and thus would be perceived to be rather similar to the mainstream generalists. Yet, the reverse conclusion cannot be made.

The low values of *Distinct songs* and *Distinct artists* values corresponds to the full range of the Gini index score (Figure 8 (c)). Thus, apart from the triangle-pattern observed with *Total user activity*, no clearer relationship with the coverage can be deduced.

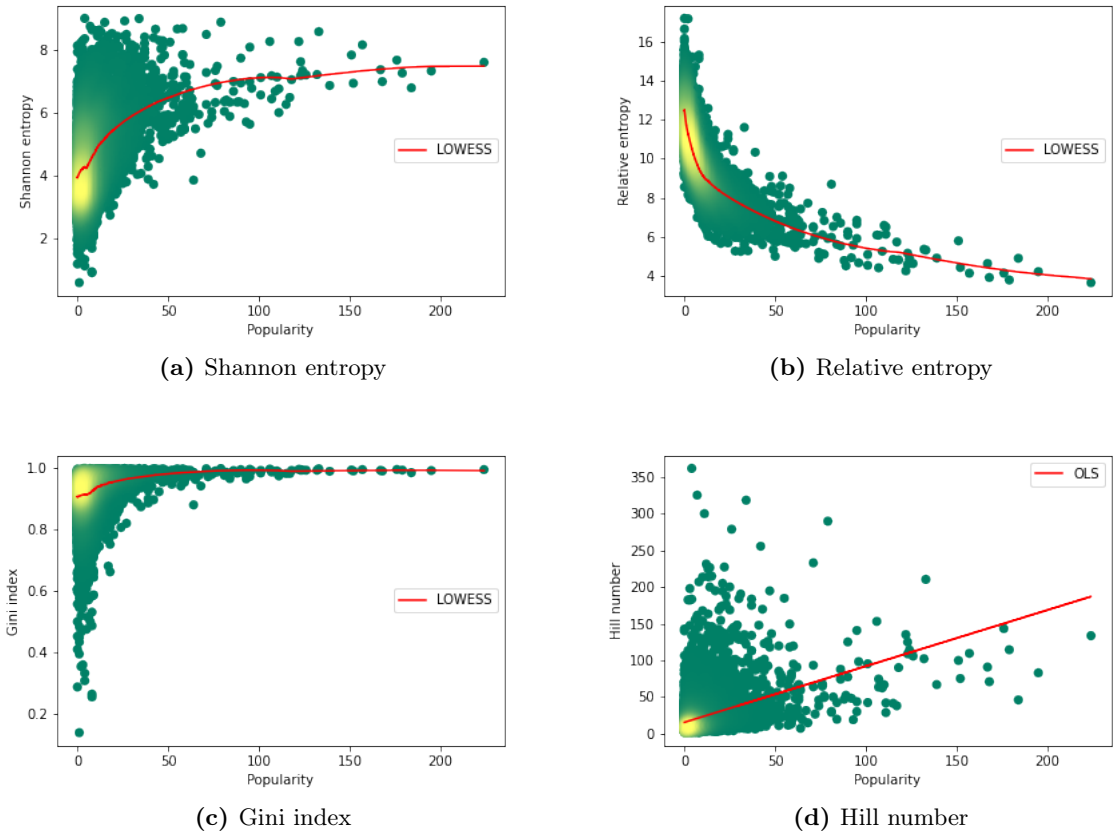
In general, all of the four statistical diversity measures point towards the same conclusion. To a certain extent, users who have higher artist or song coverage tend to have more diverse playlists. However, it must be noted that to some extent this conclusion follows because all of these measures assume that each and every song is completely distinct from another. Hence, it does not allow us to assess the correctness of the diversity measures since a user may listen to many songs by the same artist or many artists in the same genre.

7.1.3 Popularity

Popularity score was defined as the number of top 500 of the most popular songs that a user i listens to. The top 500 of the songs were defined as the 500 songs that had the highest Q^I `Count` value as defined by q_j^I described in the Kullback-Leibler divergence section. It is expected that the $D^{KL}(S_i)$ would have the highest dependence on this popularity measure as opposed to the other three measures.

The distribution of the popularity measure is illustrated in Appendix C. The range of the measure stems from 0 to 224 with a mean of 7.25 and the standard deviation as 11.928

Taking into the account Figure 9 we can see that the relative entropy has a substantially stronger relationship with the *Popularity*. Moreover, the relationship is now clearly less noisy than that observed with the other attributes. This suggests that $D^{KL}(S_i)$ to some extent is able to explain the variation in how many popular songs a person is listening to, especially considering close to non-



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 9: Diversity measures vs Popularity

existing relationship between the *Popularity* and the other three measures. This builds confidence in the reliability of this measure. Overall, it seems that users who have lower values of $D^{KL}(S_i)$ and are similar to the mainstream generalist, listen to a larger number of the top 500 of the popular songs.

The relationship between *Popularity* and the Shannon entropy is only informative based on the strong triangle-pattern present and uncovers an interesting feature of generalist users. It seems that users who listen to a lot of popular songs are likely to yield high Shannon entropy and thus be considered as generalists. Thus, this suggests that there is some relationship between the diversity of the taste and the "mainstream" music taste. If proven true, this finding might be useful in encouraging longevity of the users in the online platforms as well as could help in building better recommender systems. Nonetheless, the triangle-relationship means it would be wrong to claim that the generalist label according to the Shannon entropy suggest a strong affinity to the popular songs.

7.2 Concordance analysis

Beginning from the Kendall's τ coefficient in Table 4 we see that the extent to which the relative entropy $D^{KL}(S_i)$ is in an agreement with the other three measures is very limited. The negative value of the Kendall's τ is expected bearing in mind that the low values of the $D^{KL}(S_i)$ indicate a more generalist person, whilst the same conclusion comes with the high values of the other three measures. The limited degree of agreement is not surprising bearing in mind that $D^{KL}(S_i)$ uses mainstream generalist as the benchmark as opposed to the simple generalist used by the other three measures. Furthermore, Shannon entropy has a fairly strong concordance with both Hill number and the Gini index whilst the concordance between the Hill number and the Gini index is extremely strong and reaches 0.916. Thus, it is expected that the Gini index and the Hill number would yield very similar qualitative results.

Kendall's τ	Shannon entropy	Relative entropy	Gini index	Hill number
Shannon entropy	1	-0.36564	0.85214	0.77155
Relative entropy	-0.36564	1	-0.34627	-0.32770
Gini index	0.85214	-0.34627	1	0.91642
Hill number	0.77155	-0.32770	0.91642	1

Table 4: Kendall's τ

Albeit, the concordance between the Shannon entropy and Hill number is strong, it is not a perfect one. Thus, we should examine the cases where the two measures disagree. The same will be done with reference to the relative entropy.

Before plotting the ranks differences of the scores, it must be ensured that the low and high ranks correspond to the same conclusion, i.e. low Shannon entropy rank has to mean the same thing as the low relative entropy. It was chosen that the low rank has to correspond to the relatively specialist user and the high rank to the generalist user. As a result, the scores of the relative entropy $D^{KL}(S_i)$ rank had to be inverted since the high values of the relative entropy implies a non-mainstream specialist whilst the high values of the other measures - generalists. Taking the differences between the ranks of the scores were constructed as follows:

$$\begin{aligned}
\Delta_1 &= \boldsymbol{R}(D^{Sh}(S_i)) - [N - \boldsymbol{R}(D^{KL}(S_i))] \\
\Delta_2 &= \boldsymbol{R}(D^{Sh}(S_i)) - \boldsymbol{R}(D^{H}(S_i)) \\
\Delta_3 &= \boldsymbol{R}(D^{H}(S_i)) - [N - \boldsymbol{R}(D^{KL}(S_i))] \\
\Delta_4 &= \boldsymbol{R}(D^{Sh}(S_i)) - \boldsymbol{R}(D^{G}(S_i))
\end{aligned}$$

where N stands for the total number of users ($10,000$ in our case), $\boldsymbol{R}$ stands for the ranking of the data where the lowest rank is assigned to the lowest value of any score. In the case of ties, the

average of the ranks that would have been assigned to the tied values is taken.

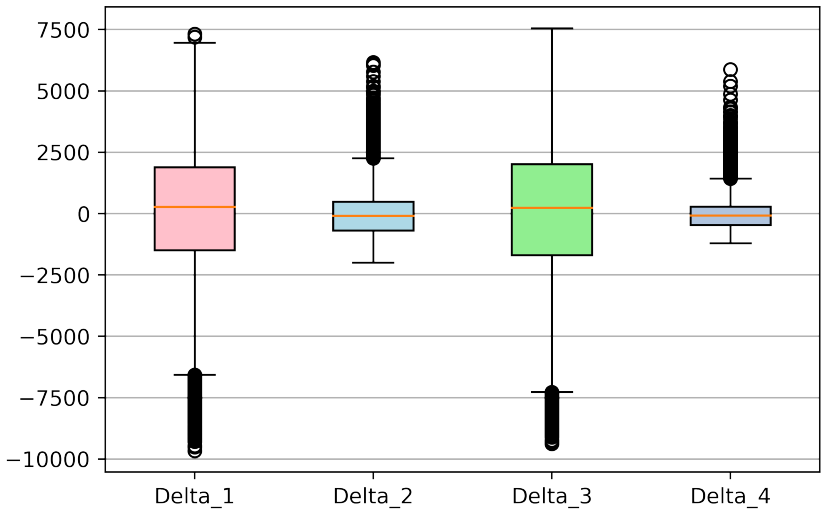


Figure 10: Box plots of the rank differences between the statistical measures

As expected, the rank differences involving relative entropy (Δ_1 and Δ_3) are the most pronounced (Figure 10). With the median of these two rank differences being higher than 0 (282 and 235 respectively), it seems that a median user has a higher Shannon entropy or Hill number score rank than the relative entropy rank. This suggests that a median user is perceived to be more generalist under the Shannon entropy and Hill number than under the relative entropy. It is important to note that the rank differences between these measures reach as much as $9,669$ in absolute terms suggesting that the scores disagree to a very significant extent. It is interesting to see which users are labelled so differently by the measures. For instance, let's consider the user $i = 5703$ whose Δ_1 has reached $-9,669$. The same user also had the largest Δ_3 of -9378. This user is considered relatively specialist according to the Shannon entropy ($D^{Sh}(S_{5703}) = 2.534$, $\boldsymbol{R}(D^{Sh}(S_{5704})) = 159$) and the Hill number ($D^{H}S_{5703} = 4.018$, $\boldsymbol{R}(D^{H}(S_{5703})) = 450$). At the same time, the same user is considered to be more similar mainstream generalist according to the relative entropy ($D^{KL}(S_{5703}) = 6.656$, $\boldsymbol{R}(D^{KL}(S_{5703})) = 9,828$). This user has a corresponding *Popularity* score of 11 meaning that she listens to more popular songs than an average user. In fact, the user listens only to the songs which are among the top 500 of the popular songs. Furthermore, those popular songs are of a rather diverse spectrum and include artists such as Björk, Florence & The Machine as well as Usher and Lil Wayne / Eminem. Thus, it seems correct for the relative entropy to label this user as a mainstream generalist. At the same time, listening to the popular songs only might suggest that a person is a specialist too. From the perspective of the Shannon entropy and Hill number, they labelled the user $i = 5703$ as a specialist because 79% of the total-playcount was spent listening to only three songs. The question whether the Shannon entropy and the Hill

number's conclusion is correct is open to debate and potentially requires a deeper music theory knowledge than we possess. The playlist of this user is available in Appendix E.

Moving towards the Shannon entropy and Hill number rank differences Δ_2 we see that the spread of the differences is much tighter than seen with the relative entropy. The median of -91 indicates that considering a median user, Shannon entropy places it closer to the specialist type than the Hill number. The maximum rank difference this time is $6,157$ which is still high, yet substantially smaller than seen with the rank difference involving relative entropy. This user $i = 6918$ yielded $D^{Sh}(S_{6918}) = 4.794$ and the Hill number $D^{H}(S_{6918}) = 3.492$ with the corresponding ranks of $8,205$ and $2,048$. Having a closer look at the user $i = 6918$ one feature become apparent. There exists one song, namely *On A Good Day* by OceanLab which was played a lot more (41 times) when compared to the other songs in the playlist. In fact, all the users who had a large positive difference in the rankings Δ_2 had at least one song which was listened disproportionately more than the other songs. On the other hand, the users who had a large negative difference in the ranking had a relatively even distribution of the `Count` variable in their playlist. This pattern can be explained by the theoretical properties of the Hill number with $m > 1$ which exaggerates the effect of the abundant songs as opposed to the rarely listened songs thus lowering the Hill number and creating a tendency for the user to be labelled as a specialist. From the interpretation point of view, Hill number embodies the idea that if most of the playcount mass lies on one or very few songs, the user is likely a specialist. However, looking at the playlist of the user $i = 6918$ (Appendix E) it seems odd to state that only 3.5 equally abundant songs are required to represent the same pattern. This user is listening to a variety of different artists (*Distinct artists*=50), leaning towards Pop and Electronic music, thus perhaps Shannon entropy's conclusion to rank it as $8,205$th - more generalist - seems more likely than the Hill number's ranking of $2,048$th. Yet again, this argument needs a deeper understanding of music theory to be made more constructive.

Rank differences between the Gini index and the other measures will not be analysed bearing in mind its similarity to the Hill number and inferior *Uniqueness* performance.

8 Discussion: Statistical Diversity Measures with hierarchical data

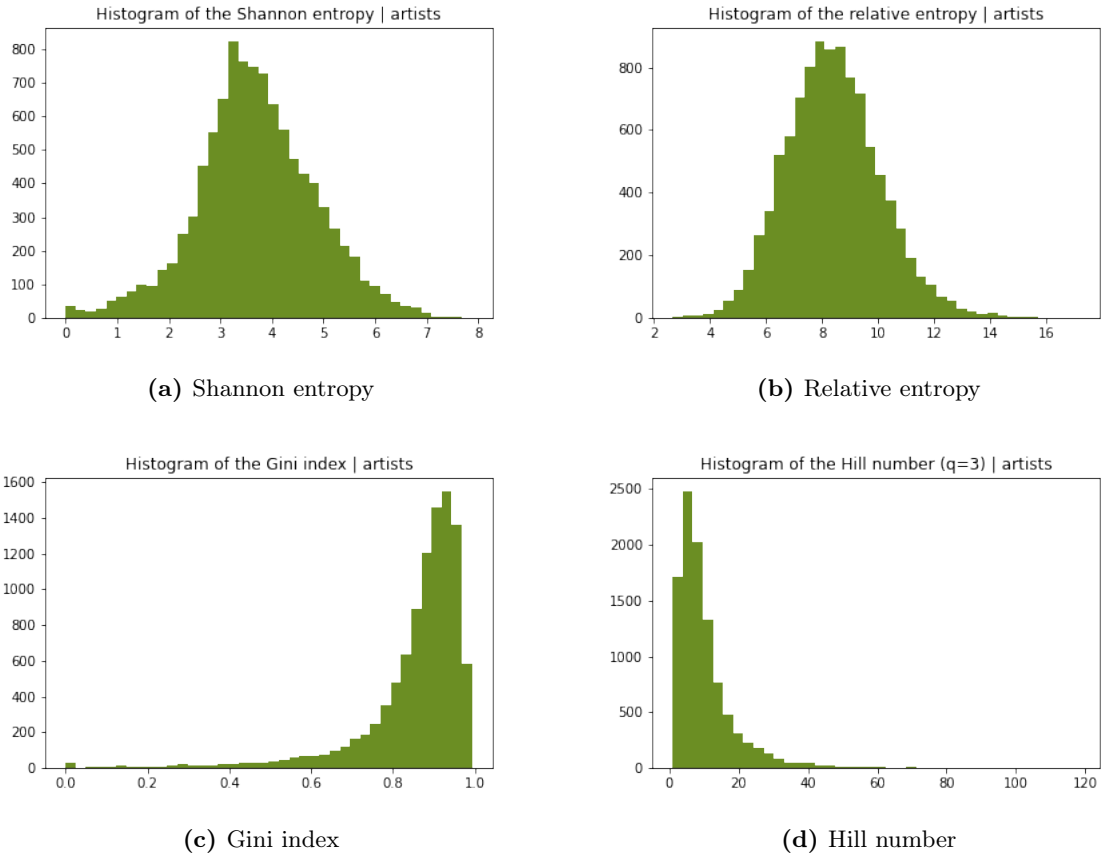
The hierarchical model, as discussed before, is implemented on the `Artist` level meaning that now not the different songs but different artists played by the user are considered in calculating the statistical diversity measures. The decision to categorise data based on the artists instead of the genres was made due to the substantial data loss occurring if the genre data was to be used as described in section 5.1. Using the notation introduced in section 3.5, let $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$ represent a collection of sets containing the songs by the each artist where $Y_\alpha = \{x_j \in \mathcal{X} | x_j \in \alpha\} \subseteq \mathcal{X}$ - a set of songs played by the artists α. Then, $y_{i\alpha} = X_i \cap Y_\alpha$ is a set of songs played by the artist α which user i listens to. Finally, $f_{i\alpha}$ denotes a number of times artist α was played by the user i and $z_{i\alpha}$ - the relative frequency (probability) of the artist α being played by the user i.

We now repeat the same evaluation analysis done with the statistical diversity measures as in the previous section, yet now we employ hierarchical structure described in section 3.5. Our subsample of users listened to $21,360$ different artists (thus $k = 21,360$) with the top 10 artists according to the sample wide q_α^I frequencies listed in Table 5.

Note, that the Million Song Dataset is prone to matching errors between the `TrackID` and `SongID`, thus errors such as a band Harmonia being among the top 10 artists appear [2]. In reality, Harmonia corresponds to Katy Perry. For our purposes, matching errors are not too important - especially because only 0.6% of the whole Million Song Dataset is certainly mismatched - and thus will be ignored.

Artist	q_α^I
Kings Of Leon	0.004684
Coldplay	0.004316
Harmonia	0.003610
Björk	0.003534
Florence + The Machine	0.003503
OneRepublic	0.003328
Train	0.003287
The Killers	0.003108
Radiohead	0.002825
Dwight Yoakam	0.002760

Table 5: Top 10 artists



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 11: Histograms of the hierarchical statistical diversity scores

Starting from the histograms in Figure 11 we see that the distributions of the statistical diversity

scores implemented on the hierarchical data are similar in shape to the ones denoting the distributions of the same scores implemented on the song-level data. Shannon entropy offers a rather round distribution with a mean of 3.690 and standard deviation - 1.153. Due to its close resemblance of the Gaussian distribution, Q-Q plot was plotted which suggested that the Shannon entropy score follows the Gaussian distribution well and deviates mildly at the left tail only (Appendix B). Thus the confidence interval for the Shannon score was constructed suggesting that 95% of the scores are expected to lie within $[5.047, 11.744]$ interval given a different sample of the data.

The artist-level relative entropy score this time had a longer right tail than when implemented on the song-level data. This might suggest that on the artist-level data, there seems to be more extreme specialist than under the song-level data. As a result, the distribution differs rather substantially from the Gaussian distribution when evaluated by the Q-Q plot. Despite that, distribution is still round with a mean of 8.395 and standard deviation of 1.709.

The Gini index score shows a vast improvement when comparing it to the song-level data. This time, albeit the distribution is skewed to the left, it offers a greater and less concentrated range of score values. This is because once the playlist is considered from the perspective of the artist, a probability of an artist α being played by the user i - $z_{i\alpha}$ - becomes greater than the probability of a particular song j being played by the user i - p_{ij}, thus the Gini-index on average will be smaller than with the song-level data. The mean of the distribution is now 0.854 and the standard deviation is 0.137. Note, there are 33 users who yielded the Gini score of 0 and all of them had only one artist in the playlist. Intuitively, it seems natural to label them as an extreme specialist as it is done by the Gini score. Nonetheless, even if the range of the score is better, it is still far from ideal considering the interpretation of the score. Now, 75% of the playlist have a probability of selecting two different artists with replacement larger than 82.663%. It is better than roughly 90% probability of selecting two different songs with replacement, yet it still seems that most of the users to some extent would be labelled as generalists.

The histogram of the Hill number also offers a less concentrated distribution of the score with the same skew to the right. Now, the mean becomes 10.249 and the standard deviation - 8.724. As before, the skew is less of a problem in the context of the Hill number than it is in the context of the Gini index due to the different interpretation.

	Shannon entropy	Relative entropy	Gini index	Hill number
Uniqueness	92.80%	99.96%	80.52%	91.06%

Table 6: Uniqueness Results with the hierarchical statistical diversity measures

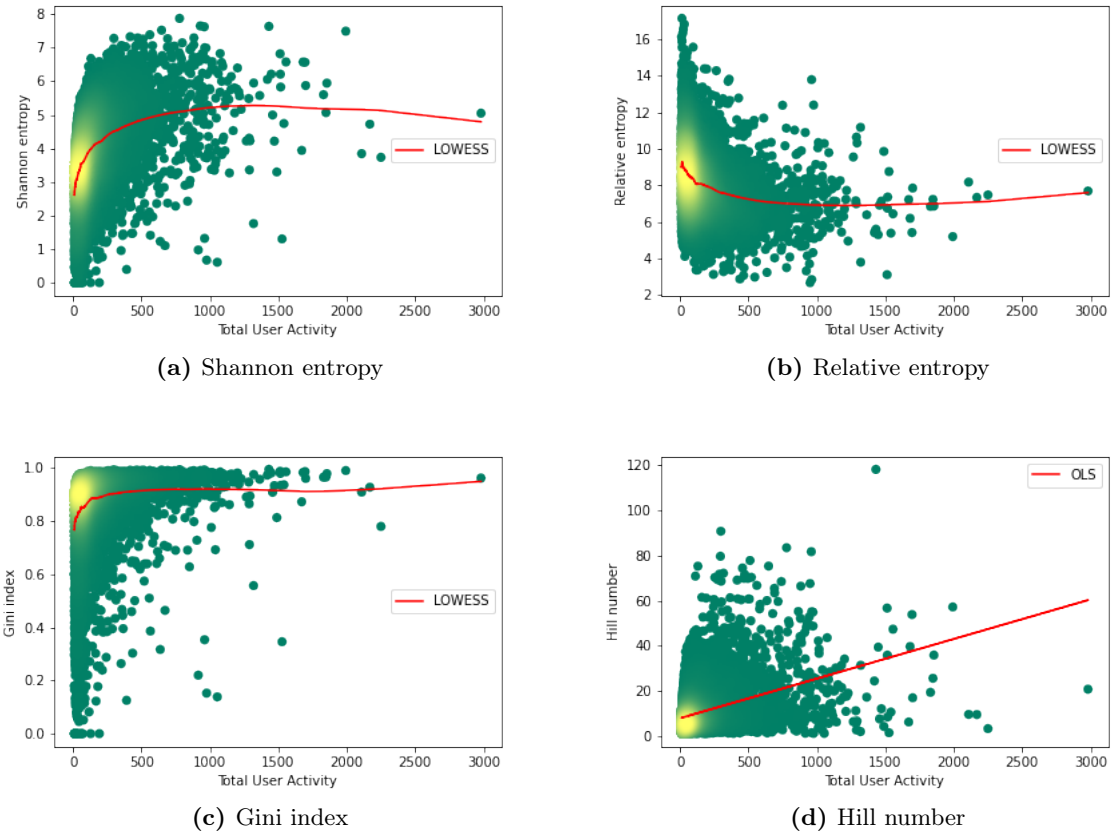
In terms of the *Uniqueness* in Table 6, all of the measures excluding the relative entropy scored better with the artist-level data than with the song level data. *Uniqueness* of the Shannon entropy rose by 5.67 percentage points, of Gini index rose by 6.73 percentage points and of Hill number - by 5.73 percentage points. The relative entropy experienced a marginal decrease in *Uniqueness* equal to 0.04 percentage points and still remains the best measure in terms of its ability to differentiate

between all the users. The overall better *Uniqueness* results suggest that with the hierarchical data, the statistical diversity measures can distinguish between the similar users better. This stands as one practical advantage of the hierarchical diversity representation.

8.1 Dependence on attributes

An identical approach to that in section 7.1 is employed to evaluate the dependence of the hierarchical statistical diversity measures on the attributes. As before, LOWESS curve is implemented to summarise non-linear relationships and OLS, to summarise linear relationships.

8.1.1 Total User activity



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 12: Diversity measures vs Total User Activity

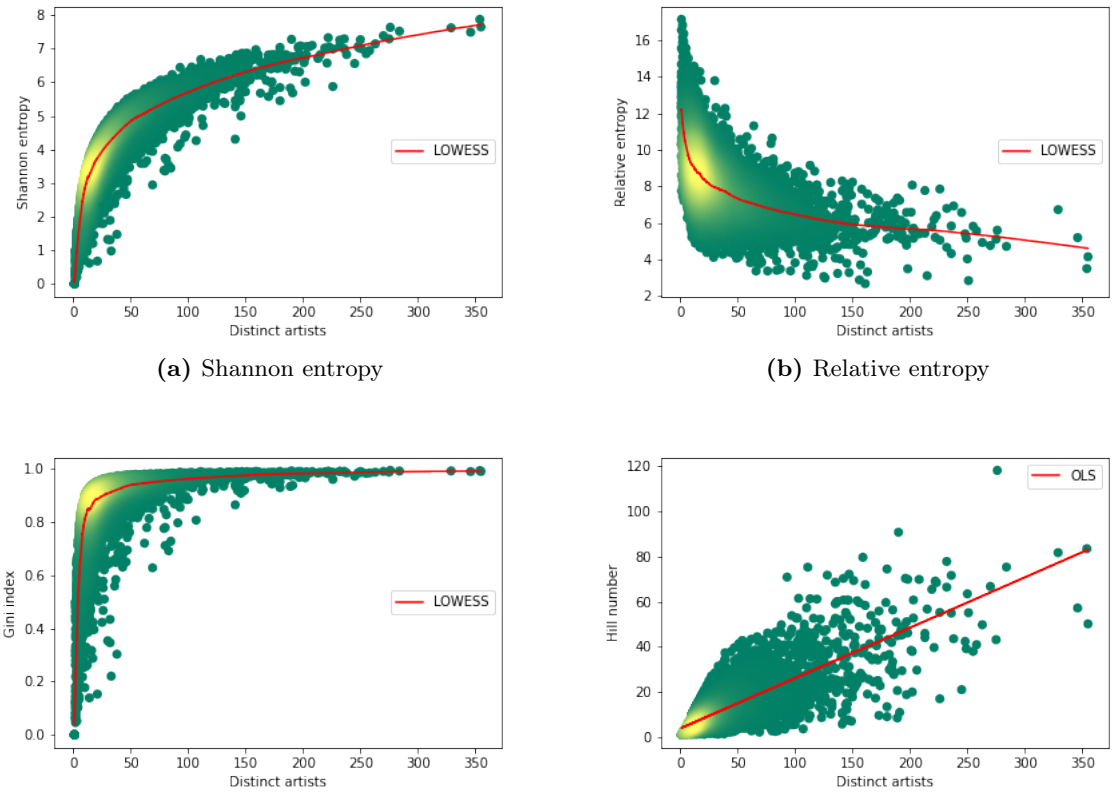
Starting from the Shannon entropy in Figure 12 (a), we see that there is close to non-existent relationship between the score and the *Total user activity*. The relationship follows the same pattern as with the song-level data, yet now is noticeably weaker. Furthermore, the *triangle-pattern* is somehow weaker than with the song-level data indicating that with the hierarchical representation, active users are more likely to be labelled as specialists than with the non-hierarchical representation.

The relationship between the relative entropy and the *Total user activity* is again very weak, thus we proceed to the Gini index. Gini index, as before, offers a strong *triangle-pattern* with otherwise non-existent relationship between the score and the *Total user activity*. LOWESS score here becomes meaningless and is only kept due to the reasons of consistency. The only information Gini index offers about the activity of a user is that the very active users are unlikely to be strong specialists.

The relationship between the *Total user activity* and the hierarchical Hill number is very similar to the one with the song-level Hill number. It is a positive one, yet rather weak. As before, due to the large level of heteroskedasticity, the regression line is uninformative.

Overall, as before there is a tendency for the more active users to be considered as having a more diverse taste in music, yet this relationship is even weaker when compared to the song-level statistical diversity measures.

8.1.2 Coverage: distinct number of artists



(a) Shannon entropy

(b) Relative entropy

(c) Gini index

(d) Hill number

Figure 13: Diversity measures vs Distinct Artists

Plotting the hierarchical statistical diversity measures against the *Distinct songs* variable showed

that the relationship between the scores and the coverage in terms of song variable is weaker than with the coverage in terms of artists variable. This is to be expected since the hierarchical statistical diversity measures virtually ignore the number of distinct songs and only consider them implicitly via the number of distinct artists. Thus, we will focus on the *Distinct Artists* variable instead. The scatter-plots between the scores and the *Distinct Songs* are available in Appendix F.

As was expected, the relationships between almost all the scores and the *Distinct artists* variable is fairly strong. Shannon entropy in Figure 13 (a) shows a strong logarithmic relationship with the number of distinct artists. The strong relationship is expected bearing in mind that the entropy was directly calculated on the number of distinct artists a user is listening to. This indicates that according to the Shannon entropy the more distinct artists a user has in her playlist, the more likely she is to be a generalist.

The relative entropy offers a weaker relationship with the *Distinct artists* than the Shannon entropy, yet it is stronger than with the song-level relative entropy. Together with strong *triangle-pattern* it seems that users who listen to many distinct artists are more likely to be mainstream-generalist or at least unlikely to be non-mainstream specialists.

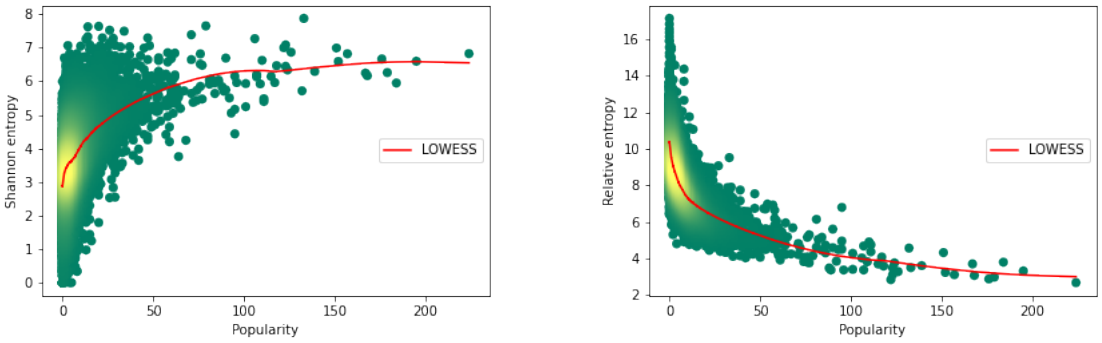
Gini index, as before, offers no meaningful relationship apart from the *triangle-pattern* leading to the same before: users listening to a lot of artists are unlikely to be perceived by the Gini-Simpson index as specialists.

Hill number shows a positive relationship with the *Distinct artists* variable. However, the relationship between the song-level Hill number and the *Distinct songs* still appears to be stronger (Appendix F). This might be the case if users tend to listen to different artists at more equal frequencies than they listen to different songs. This, in turn, would lead Hill number with $m = 3$ to exaggerate the weight of those abundant artists even more leading to a more significant deviation from the simple *Distinct artists* variable.

In general, all the four measures point to the same conclusion that users listening to more distinct artists are, to some extent, more likely to be considered as generalists.

8.1.3 Popularity

The relationships between the hierarchical similarity measures and the *Popularity* follows closely the patterns seen with the non-hierarchical diversity measures (Figure 14). As before, only the relative entropy has a strong relationship with the *Popularity* with Gini index and the Hill number measures having a weak or non-existent relationship (thus their scatterplots are omitted). Nonetheless, there is some weak positive relationship between the *Popularity* and the Shannon entropy indicating again that the *Popularity* itself might be correlated with the generalist nature at least to some extent. Overall, it seems, as expected, that the more often a user is listening to the popular songs, the more likely she is to be a mainstream-generalist and - according to the Shannon entropy - unlikely to be a specialist.



(a) Shannon entropy

(b) Relative entropy

Figure 14: Diversity measures vs Popularity

8.2 Concordance analysis

Identical concordance analysis to that seen before is conducted with the hierarchical statistical diversity measures.

Kendall's τ	Shannon entropy	Relative entropy	Gini index	Hill number
Shannon entropy	1	-0.43092	0.84233	0.76844
Relative entropy	-0.43092	1	-0.41819	-0.40140
Gini index	0.84233	-0.41819	1	0.92273
Hill number	0.76844	-0.40140	0.92273	1

Table 7: Kendall's τ for the hierarchical statistical diversity measures

When defined on the artist-level data, the Kendall's τ between the relative entropy and the other three measures has in absolute terms increased (Table 7). Nonetheless, it still reaches at most -0.43092 with the Shannon entropy and is still considered as a weak discordance. As before, the negative sign of Kendall's τ is expected since the low values of $D^{KL}(Z_i)$ imply similar conclusion to the high values of the other three measures. Thus, the higher agreement between the relative entropy and the other three statistical measures might imply that the extent to which a user has a mainstream generalist taste in music in terms of artists is better approximated by the extent to which she is a generalist with the artist-level as opposed to the song-level data. The Kendall's τ among the Shannon entropy, Gini index and Hill number remained virtually the same.

The rank differences were constructed in the same way as with the song-level diversity measures: the relatively most specialist users according to each measure are assigned the lowest rank and the most generalists - the highest. To put it formally:

$$\Delta_5 = \boldsymbol{R}(D^{Sh}(Z_i)) - [N - \boldsymbol{R}(D^{KL}(Z_i))]$$
$$\Delta_6 = \boldsymbol{R}(D^{Sh}(Z_i)) - \boldsymbol{R}(D^{H}(Z_i))$$
$$\Delta_7 = \boldsymbol{R}(D^{H}(Z_i)) - [N - \boldsymbol{R}(D^{KL}(Z_i))]$$
$$\Delta_8 = \boldsymbol{R}(D^{Sh}(Z_i)) - \boldsymbol{R}(D^{G}(Z_i))$$

The resulting rank difference distributions in Figure 15 indicate that the rank differences involving relative entropy, Δ_5 and Δ_7, seem to be less skewed than with the song-level diversity measures (Δ_1 and Δ_3). This suggests a more even spread of the users who are considered more/less non-mainstream specialists according to the relative entropy than considered as specialists according to the Shannon entropy or Hill number. The median values of 139 (Δ_5) and 106 (Δ_7) still imply that a median user is considered to be more generalist under the Shannon entropy or the Hill number than she is considered as a mainstream generalist by the relative entropy. Furthermore, the spread of the rank difference is narrower with the maximum value of Δ_5 in absolute terms being $8,288$ as opposed to $9,669$ with the Δ_1. Yet, the same user $i = 5703$ had the maximum rank difference in artist-level and song-level measures owing to the same reason of frequent listening of very popular songs. Thus, this user will not be analysed again. The maximum value of Δ_7 is $8,249$. This time the users with the largest rank differences between the Shannon entropy - relative entropy (Δ_5) and Hill number - relative entropy (Δ_7) rank differences measures are not the same and will be analysed separately.

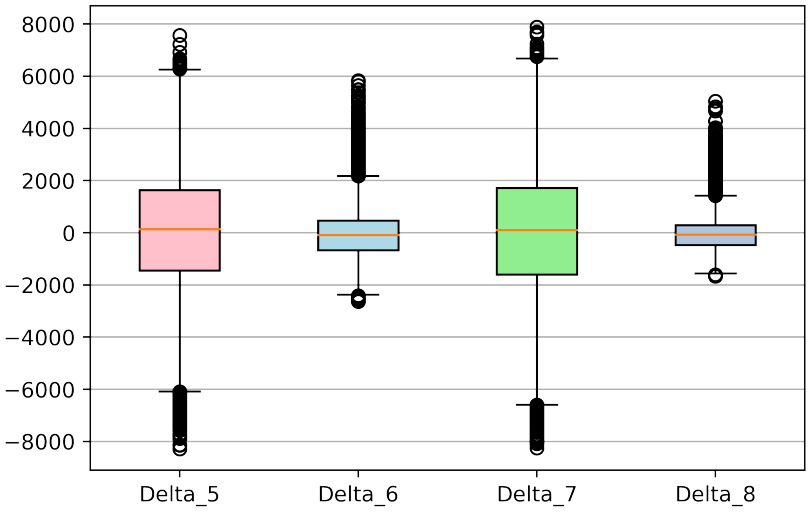


Figure 15: Box plots of the rank differences between the hierarchical statistical measures

The user $i = 3075$ had the largest rank difference between the Hill number and the relative entropy ($\Delta_7 = 8,249$). This person yields $D^H(Z_{3075}) = 2.5$ and $D^{KL}(Z_{3075}) = 6.321$ thus it is considered a rather specialist by the Hill number (rank 729/10,000) and close to the mainstream generalist by the relative entropy (rank 8,977/10,000). Looking closer at the playlist of this user (Appendix E) the reason behind such rank disparity becomes obvious. Starting from the perspective of the Hill number, 54% of the time the user was listening to the *Killers* songs with the other artists being played only once each. Since the Hill number with $m = 3$ exaggerates the often-played artists, the resulting low Hill number is natural. From the perspective of the relative entropy, 10 out of 13 songs played by the user were among the top 500 the most popular ones, thus the mainstream nature of the user dominated and led to the low $D^{KL}(Z_{3075})$ score and thus close proximity to the mainstream-generalist label. However, it is debatable to what extent this user is a generalist with all of its playlist showing a strong inclination towards Pop-Rock/Rock style of music, albeit of a popular type. This shows the limitation of the $D^{KL}(Z_i)$ score: two forces - mainstream nature and the generalist nature - both push the score down and the very strong mainstream specialist can yield a comparable score as a slightly less mainstream generalist. Thus, this score is perhaps most useful when both the degree of the mainstream and generalist nature are important.

Coming to the Shannon entropy and Hill number rank difference Δ_6, the maximum rank difference is $5,828$ which again is smaller than that seen with the song-level data. The median now is -76 which is again, in absolute terms smaller than with the song-level data, yet still negative suggesting that the Hill number tends to label a median user more generalist as opposed to the Shannon entropy. The user with the highest rank difference ($i = 8536$) is not the same as with the song-level measures Δ_2 and yielded $D^{Sh}(Z_{8536}) = 4.579$ (rank 7,860), $D^H(Z_{8536}) = 4.287$ (rank 2,032). Thus the user is closer to the generalist under the Shannon entropy and closer to the specialist under the Hill number. The reason is as expected: there is one artist *Juanes* who is listened by the user 38% of the time. On the other hand, there are 79 distinct artists in the playlist of this user. The former reason leads to a smaller Hill number and the latter to a larger Shannon entropy. The list of artists listened by this user is rather diverse and ranges from Latin music to a slightly heavier Rock music bands such as Limp Bizkit (Appendix E), thus, arguably, it perhaps is more natural to consider this person a more generalist than the specialist. However, a deeper music theory knowledge is required to make this argument more constructive.

9 Discussion: Geometrical Diversity Measures

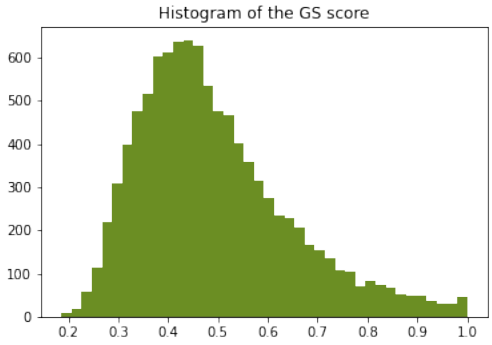
We adopt the same evaluation strategy as with the statistical diversity measures when analysing the geometrical diversity measures. The Generalist-Specialist score is implemented on the normalised vectors which, as explained in the Lemma 3 is equivalent to taking the Euclidean norm of the user's centroid vector μ_i. TS-SS, on the other hand, will be implemented on both: normalised and non-normalised vectors in order to assess the quality of diversity when the magnitude of the song vector incorporates popularity and when it does not. We denote the non-normalised TS-SS score as

$D^{TS-R}(S_i)$ (R standing for "raw") and the normalised as $D^{TS-N}(S_i)$ (N standing for normalised).

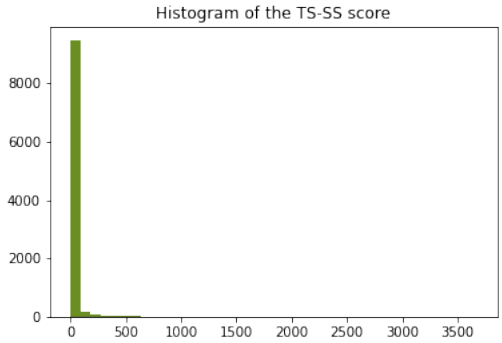
The histogram of the Generalist-Specialist score appears to be slightly skewed to the right (Figure 16 (a)) with a minimum value of 0.186, maximum - 1. The mean and standard deviation of the Generalist-Specialist score are 0.492 and 0.153. The histogram of the two TS-SS scores appeared to be extremely different. The non-normalised $D^{TS-R}(S_i)$ score presents an extremely right-skewed distribution (minimum and maximum values of $1.760873 \cdot 10^{-15}$ and 3675.942 repetitively). The corresponding mean and standard deviation are 27.8 and 155.004. The normalised TS-SS score was scaled by multiplying by $10,000$ to have a nicer representation without changing the meaning of the results. The histogram in Figure 16 (c) of this score appears to be skewed to the left, yet to a substantially lesser extent than the same score on the non-normalised data. The values of the scaled score range from $1.852093 \cdot 10^{-7}$ to 1.457 with the mean of 1.151 and the standard deviation of 0.262. This difference in the distributions of the normalised vs non-normalised TS-SS scores reflects the immense importance of the magnitude of the vectors. Not only the scale is narrower, which is expected, but the skew is different. The non-normalised TS-SS score has a lot of relatively large extreme values which might be due to the popularity of the songs. A user with the largest $D^{TS-R}(S_i)$ score of 3675.942 is an already discussed user $i = 5703$ whose playlist consists only of the songs which are among the top 500 most popular songs (Appendix E). The same user yielded the scaled $D^{TS-N}(S_{5703})$ score of 0.740 placing it in the lowest quartile of the normalised score distribution and thus making it more likely to be a specialist. Thus, this gives one piece of evidence to the hypothesis that with the non-normalised data, users who listen to popular songs are automatically likely to be labelled as more diverse.

Nonetheless, the different shapes of the histograms do not translate to the *Uniqueness* as it was the case with the statistical similarity measures. All three scores yield the *Uniqueness* of 100% indicating that they manage to distinguish each and every user. This automatically serves as an advantage of the geometrical measures over the Shannon entropy, Gini index and Hill number as defined on both song and artist level data.

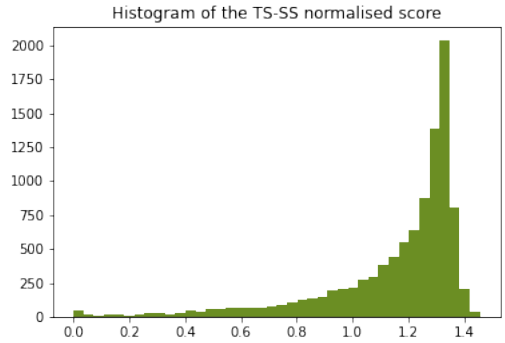
Furthermore, the empirical existence of a theoretical flaw of the Generalist-Specialist score was examined. As explained in section 4.2, Generalist-Specialist score does not take the magnitude of the vector into account, thus making it possible for the situations where two vector overlap in terms of their angles from the perspective of the centroid μ_i to occur. In such a case, the songs would be label as being identical (Figure 1). To check the validity of this concern, the cosine similarity between each song could be calculated. If the score between any two songs is 1, then those song vectors are overlapping. However, to speed up the code running-times, the cosine-similarity was defined for each user between the songs of her playlist and her *user preference* vector μ_i. Then, the number of songs which had an identical cosine similarity for each user is calculated. Being aware that the identical cosine similarity score can occur if the songs simply form the same angle but do not take the exact same position, those songs are examined closer. There appeared to be 66 users who had at least two identical cosine similarity scores among the songs of their playlist. Having a closer look at those songs, it appeared that they were assigned identical embedding vectors $\vec{x}_j$



(a) Generalist-Specialist histogram



(b) TS-SS non-normalised histogram



(c) TS-SS normalised histogram

Figure 16: Histograms of the geometrical similarity measures

by the Hierarchical Poisson Factorisation. For instance, two songs by the artist Dismantled called *Fields* and *Thanks For Everything (Album Version)* were assigned the same embedding vector albeit being two different songs. For a non-professional ear, those songs are of a very similar style, yet clearly, they are not identical. Thus, how appropriate it is to treat them as equivalent is debatable. Similarly, two songs by the artist Big Boss Man called *Trilby of Fun* and *VIP 233* were also given the same embedding vector. This, however, points to the potential drawbacks of the embedding rather than the Generalist-Specialist score itself. Naturally, the exact same users had more than one identical TS-SS scores between the *user preference* vector and at least two of the songs as the Generalist-Specialist score. This indicates that all of the overlapping songs share the same vector $\vec{x}_j$ with at least one other song. Consequently, no overlapping songs for two distinctly embedded songs were found at least with this subsample. Nevertheless, this might depend on the embedding and subsample, thus the potential existence of the Generalist-Specialist flaw in terms of overlapping vectors cannot be fully ruled out.

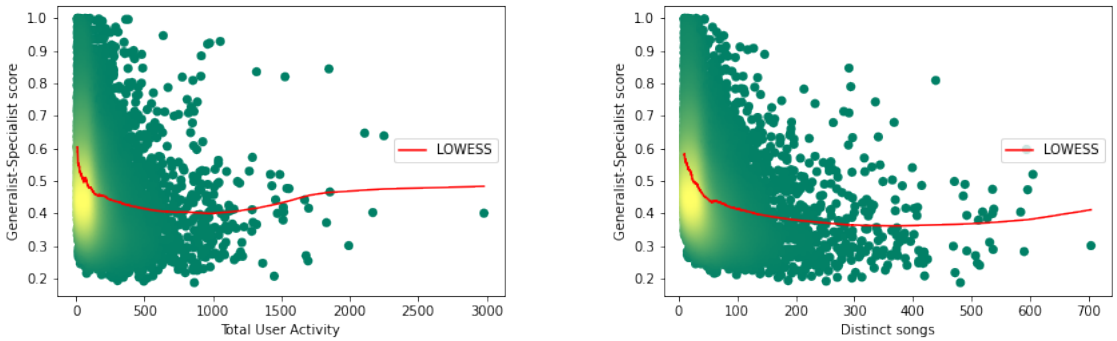
9.1 Dependence on attributes

Plotting all three geometrical scores against the attributes showed that the non-normalised TS-SS $D^{TS-R}(S_i)$ score showd no relationship with any of them, even the *Popularity*. This is potentially the case because only the users whose playlist consists mostly of the popular songs are considered as of generalist regardless of the actual number of songs they listened, i.e. the fraction of popular songs, not their absolute number *per se* matters. For example, a user with the largest number of popular songs (*Popularity* = 224) yielded the $D^{TS-R}(S_i)$ score of 74.194 which is in the top quartile, yet not the largest value. This user's playlist was 91% made of out the popular songs, whilst the users with the top three $D^{TS-R}(S_i)$ scores had playlists which consisted solely out of the popular songs. Thus, it might be the case that the higher percentage of popular songs in the playlist creates a tendency for a higher $D^{TS-R}(S_i)$ score. Thus, a new variable denoting the fraction of the popular songs in the playlist was created. The scatterplot between this variable and the non-normalised TS-SS score is available in the Appendix G and it shows that there is only a very mild relationship between the score and the fraction of the popular songs listened by the user. Thus, it seems that the $D^{TS-R}(S_i)$ score incorporates more information in itself than just the mere popularity. Nevertheless, the scatter-plots containing $D^{TS-R}(S_i)$ will not be analysed further.

Looking at Figure 17 we see that the Generalist-Specialist score has some, albeit extremely weak, relationship with the *Distinct Songs* and *Distinct Artists* variables. The relationship between the score and the *Distinct Artists* variable seems to be stronger and thus suggests that the number of artists might be a more appropriate crude measure to assess the diversity of a playlist. This, in turn, gives some support for using hierarchical data representation if opting to the statistical diversity measures. Overall, there seems to be some negative relationship between the score and the *Distinct artist* variable when the number of artists increases from 0 to 30 only.

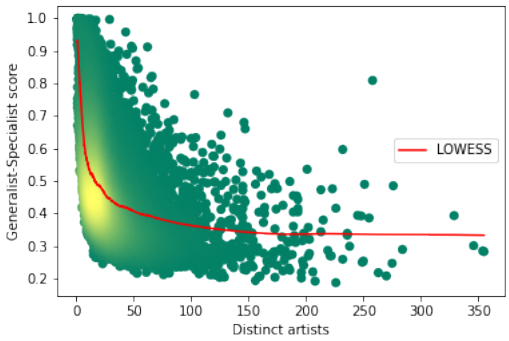
The other two variables, namely *Total user activity* and *Popularity* show close to none relationship with the score. Both coverage variables seem to be negatively related to the $D^{GS}(S_i)$ score with users who have a lot of distinct artists or distinct songs in their playlists unlikely to yield high $D^{GS}(S_i)$ score and thus not likely to be specialists. However, users who have a small number of the distinct artists or songs in their playlist can take any place in the generalist-specialist spectrum as perceived by the $D^{GS}(S_i)$ score leading to the same *triangle-pattern* as observed with the statistical diversity measures.

Figure 18, in turn, shows the scatterplots of the $D^{TS-N}(S_i)$ score with the four attributes. In a nutshell, there is no clear relationship with the score and any of the attributes as the shapes of the LOWESS curves are rather similar to the ones seen with the Generalist-Specialist score in Figure 17. As with the $D^{GS}(S_i)$, the most pronounced relationship is with the *Distinct Artists* variable. Nevertheless, apart from the *triangle-pattern* showing that users listening to many different artists are likely to be considered as generalists by $D^{TS-N}(S_i)$ score, no further meaningful conclusions can be drawn.

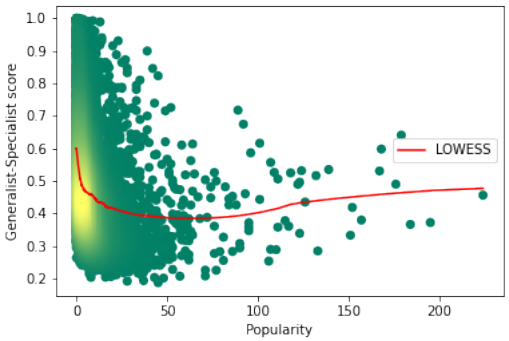


(a) Total User Activity

(b) Distinct Songs



(c) Distinct Artists



(d) Popularity

Figure 17: Generalist-Specialist score vs attributes

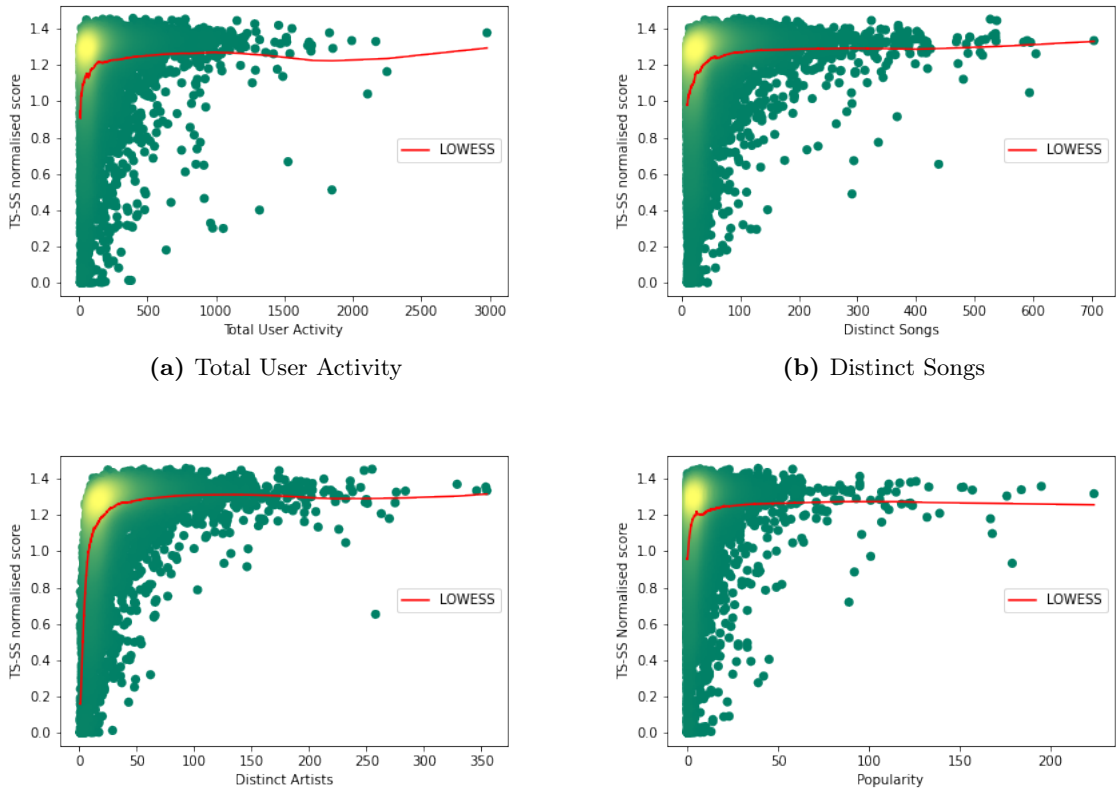
9.2 Concordance analysis

Same three geometrical diversity measures are compared in terms of their concordance. Beginning from the Kendall's τ in Table 8, we see that the Generalist-specialist and $D^{TS-N}(S_i)$ scores have a clear tendency towards ranking the users as the opposites in terms of their ranking with Kendall's τ reaching -0.6980. The negative sign is expected bearing in mind that the higher number of the $D^{GS}(S_i)$ score means closer proximity to the specialists and the higher number of the both $D^{TS-SS}(S_i)$ score - higher proximity to the generalists.

Kendall's τ	Generalist-Specialist	Normalised TS-SS	Non-normalised TS-SS
Generalist-Specialist	1	-0.6980	-0.16756
Normalised TS-SS	-0.6980	1	0.15924
Non-normalised TS-SS	-0.16756	0.15924	1

Table 8: Kendall's τ for the geometrical diversity measures

On the other hand, the concordance between the $D^{TS-N}(S_i)$, $D^{GS}(S_i)$ scores and $D^{TS-R}(S_i)$



(a) Total User Activity

(b) Distinct Songs

(c) Distinct Artists

(d) Popularity

Figure 18: Normalised TS-SS score vs attributes

is very small ($\tau = 0.15924$ and -0.16756 respectively) and again signals immense importance of the magnitude of the vector in the TS-SS score and its measure of diversity.

In order to examine such disparity in rankings closer, rank difference measures were constructed. As before a low rank was chosen to refer to the most specialist user, thus $D^{GS}(S_i)$ score ranking was inverted. The rank difference plots in the Figure 19 formally refers to these transformations:

$$\Delta_9 = (N - \boldsymbol{R}(D^{GS}(S_i))) - \boldsymbol{R}(D^{TS-R}(S_i))$$
$$\Delta_{10} = \boldsymbol{R}(D^{GS}(S_i)) - \boldsymbol{R}(D^{TS-N}(S_i))$$
$$\Delta_{11} = \boldsymbol{R}(D^{TS-N}(S_i)) - \boldsymbol{R}(D^{TS-R}(S_i))$$

where $\boldsymbol{R}$ stands for the ranking where the lowest value for each measure is assigned a low rank and N - the total number of users (10,000 in this case). Since the *Uniqueness* reached 100% for all the geometrical measures, no ties occur and thus no tie strategy needs to be described.

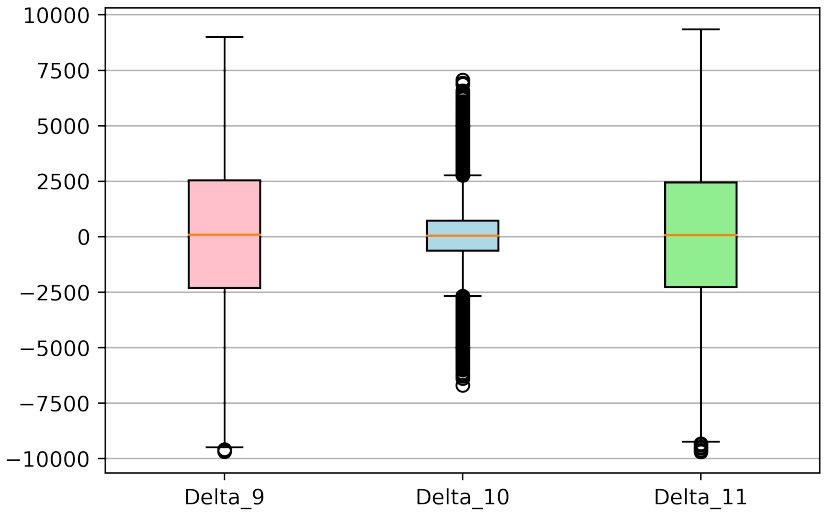


Figure 19: Box plots of the rank differences between the geometrical diversity measures

Before proceeding to the analysis of specific users it must be noted that unlike with the statistical measure, now it is not straightforward why the geometrical diversity measures rank the users differently. It depends heavily on the embedding and the angle and/or the magnitude of the song vectors. The fraction of the popular songs in the playlist is perhaps the only attribute which would be able to explain discordance between the non-normalised TS-SS score and the other two scores. Thus, the quality of the explanation of why the scores give different results is expected to be inferior to that seen with the statistical diversity measure.

From the plot in Figure 19, we see that the Δ_9 covers almost the full range of the values with the median of 103 indicating that a median user tends to be considered more generalist by the Generalist-Specialist score than by the non-normalised TS-SS score. The highest rank difference in absolute terms reaches $-9,680$. The user with this values of Δ_9, namely $i = 8243$, is considered as a strong generalist by the Generalist-Specialist score ($(N - \boldsymbol{R}(D^{GS}(S_{8243})) = 9,478$) and a specialist by the non-normalised TS-SS score ($\boldsymbol{R}(D^{TS-R}(S_{8243})) = 471$). Looking closer at the playlist of this user (Appendix E), we see that she listens to all the songs at the equal frequencies and does not listen to any of the top 500 popular songs. The latter fact might be the reason why the user is placed so low according to the $D^{TS-R}(S_i)$ score. The artists listened by this user cover a very wide array of the genres, namely Pop-rock, Punk rock, Avant-garde, Pop, House, Indie rock, etc., thus it seems extremely odd to consider this person a specialist. This gives one piece of evidence that the Generalist-Specialist score is better than the non-normalised TS-SS score.

The Δ_{10} rank difference covers a narrower range than the Δ_9 with the median value of 58 again showing that a median user is considered slightly more generalist under the Generalist-Specialist score than by the TS-SS score implemented on the noramlised data. The maximum value of Δ_{10}

is $7,064$ carried by the user $i = 8600$ who is considered a very strong generalist by the Generalist-Specialist score ($N - \boldsymbol{R}(D^{GS}(S_{8600})) = 9,996$) and a slightly specialist by the normalised TS-SS score ($\boldsymbol{R}(D^{TS-N}(S_{8600})) = 2,932$). This user listens to a plethora of different songs (*Distinct Songs* = 481) and artists (*Distinct Artists* = 226) thus, this gives a sign that it might be a generalist. At the same time, the artist to whom the user listens the most are mostly playing Alternative Rock, Electronic and Electronic Rock music which pushes it slightly to a more specialist side. The conclusion which measure is correct is hard to attain and will likely depend on the purpose and needs of the application. If a variety of the artists outweighs the fact that the artists who receive most of the playcount are from two or three main genres, then the Generalist-Specialist score might be preferred. However, considering the user's artist genre distribution, it might seem wrong to label her as virtually the most generalist user. Especially, taking into the account that the user $i = 8243$ who was described above is considered more specialist despite its equal listening frequencies to the artist from a very wide array of genres, yet who listens to fewer artists in total. However, the validity of such a statement is open to debate.

Lastly, looking into the Δ_{11} we see that the median value of this rank difference is 80, thus again the TS-SS score implemented on the normalised data considers a median user to be more generalist than the TS-SS score implemented on the non-normalised data. The maximum absolute rank difference value is now $-9,693$. This user $i = 9416$ is considered as a strong generalist by the non-normalised TS-SS score ($\boldsymbol{R}(D^{TS-R}(S_{9416})) = 9,976$) and a specialist by the normalised TS-SS score ($\boldsymbol{R}(D^{TS-N}(S_{9416})) = 283$). This is likely to be the case because 100% of the user's songs are among the top 500 of the popular songs. Looking closer into the playlist of this user (Appendix E) we see that all of these songs are either in Pop or Alternative Rock genres. Thus it seems incorrect to label the user as such a strong generalist. Again, it seems that for the HPF embedding, normalisation is crucial if we want to exclude all the information of popularity in the measure of diversity. In a sense, TS-SS on a non-normalised data becomes closer to the Kullback-Leibler divergence variations described before, with, in some sense, the reference group being that of a mainstream-generalist in terms of the proportion of the popular songs in the playlist.

10 Cross-analysis of all diversity measures

To compare all the diversity measures introduced, Kendall's τ metric will be constructed as well as the specific users compared. This time, the users who yielded the largest rank differences so far will be evaluated only. The reason for this is that the label of those users already appeared to be prone to the most disagreement between the measures. Thus, those users are the edge cases which could help to identify the flaws or special features of the measures. Hence, the users 3075, 5703, 6918, 8243, 8536, 8600 and 9416 (Appendix E) will be evaluated using all the diversity measures introduced. To simplify the notation, the users are named in letters from A to G according to the alphabet in the order as presented above.

Furthermore, the *correctness* or *accuracy* of these measures will not be evaluated due to the lack

of music theory knowledge available to the author as well as the inherent ambiguity of the diversity labelling. Instead, the disagreements between the measures will be discussed in the light of their special features thus deepening an understand on the expected behaviour of the measures and their potential applications. For the sake of brevity, only the best performing measures will be evaluated, i.e. Gini index will not be discussed owing to its inferior *Uniqueness* and distorted interpretation as compared to the other measures. Lastly, the relative entropy defined on the song as well as on the artist level will be compared to the TS-SS score defined on the non-normalised data owing to the notion of the popularity defined implicitly in these three scores.

10.1 Description of the users

To facilitate the evaluation, we first describe the playlists of these users in Table 9.

User	*Total User Activity*	*Distinct Songs*	*Distinct Artists*	*Popularity*
A	13	13	7	10
B	76	11	10	11
C	156	100	50	14
D	20	20	19	0
E	201	137	79	8
F	864	481	226	45
G	18	13	12	13

Table 9: Summary of the users' playlists

Users A, B and G appeared to have rather short playlists which are dominated by the Pop or Alternative Rock popular songs. User A, however, listened mostly to one artist Killers with some appearance of other Pop, Alternative Rock and Soul music whilst user B listened the most to the three artists, namely Björk, Dwight Yoakam and Kings Of Leon, which fall under Contemporary Electronic, Country and Alternative Rock genres. User G had the list of songs in her playlist very similar to that of user B, yet with more evenly distributed `Count` variable and no songs outside of the Pop, Alternative Rock genres.

Users E and F had the largest number of various artists and songs in their playlist. User E distinguished by having a large portion of her playlist constituted by the Latin artists mostly in the Latin Pop, Latin Rock and Salsa genres with some appearance of Alternative rock. One artist - Juanes - was listened substantially more than the other artists. User F, on the other hand, listened mostly to the Alternative Rock artists, especially the ones known for the Emo rock, Indie rock and Punk rock, as well as some Electronic, House, Techno music. However, both users E and F have a variety of songs from other genres, such as Pop, although they were listened on very low frequencies (usually listened to only once).

User C appeared to listen to mostly two genres: Pop and Electronic (including House, Trance). As with the user E, one artist OceanLab was listened to substantially more than the others. Finally, user D distinguishes by having an almost uniform distribution of its `Count` variable. The playlist

consists mostly of some subgenre of Rock music, dominated by Indie and Folk Rock songs, yet also incorporates Electronic music as well as Pop to a relatively large extent. In general, it is expected for the users A, B and G to be considered as more specialist and users C, D, E and F - more generalists.

10.2 Analysis

Kendall's τ analysis (Table 10) uncovers one important feature of the hierarchical statistical diversity measures. Hierarchical Shannon entropy and the Hill number scores seem to be in a higher absolute concordance with the geometrical diversity measures than the non-hierarchical statistical scores. This gives some evidence to the reliability of the hypothesis that measuring users diversity in terms of the different artist instead of songs gives us a partial remedy to the fact that otherwise, the statistical measures are unable to recognise how different or similar songs in the playlist are. We can claim this because we know for a fact that the geometrical directly consider the similarity of the songs via the embedded space. Nonetheless, the Kendall's τ between the statistical and geometrical diversity measures is rather weak and in absolute terms reaches at most -0.4616 with the Generalist-Specialist score and the hierarchical Shannon entropy.

	Song-level		Artist-level	
Kendall's τ	$D^{Sh}(S_i)$	$D^{H}(Z_i)$	$D^{Sh}(Z_i)$	$D^{H}(Z_i)$
$D^{GS}(S_i)$	-0.2925	-0.2440	-0.4616	-0.4478
$D^{TS-N}(S_i)$	0.3115	0.2390	0.45297	0.378952

Table 10: Kendall's τ between the geometrical and statistical diversity measures

Starting from the Shannon entropy, Hill number, Generalist-Specialist score and the TS-SS implemented on normalised data we see the corresponding scores in Table 11. The number in brackets refers to the rank of the score where the lowest rank implies the most specialist user in the group. The $D^{TS-N}(S_i)$, as before, was scaled as $D^{TS-N}(S_i) = D^{TS-N}(S_i) \cdot 10^3$.

	Statistical		Hierarchical statistical		Geometrical	
User	$D^{Sh}(S_i)$	$D^{H}(3, S_i)$	$D^{Sh}(Z_i)$	$D^{H}(3, Z_i)$	$D^{GS}(S_i)$	$D^{TS-N}(S_i)$
A	3.700 (3)	13.000 (4)	2.188 (1)	2.509 (1)	0.585 (3)	1.011 (3)
B	2.534 (1)	4.018 (1)	2.507 (2)	4.018 (2)	0.663 (2)	0.740 (2)
C	5.644 (5)	7.406 (2)	4.685 (6)	7.304 (4)	0.481 (5)	1.168 (5)
D	4.321 (4)	20.000 (5)	4.221 (4)	17.541 (6)	0.293 (6)	1.305 (7)
E	6.828 (6)	70.090 (6)	4.579 (5)	4.287 (3)	0.578 (4)	1.266 (6)
F	8.318 (7)	91.544 (7)	6.993 (7)	55.182 (7)	0.187 (7)	1.121 (4)
G	3.530 (2)	9.000 (3)	3.419 (3)	8.646 (5)	0.857 (1)	0.417 (1)

Table 11: Diversity scores of the specific users (rank)

Starting from the most specialist user ranking, we see that all three groups of measures (Statistical, Hierarchical statistical and Geometrical) chose the most specialist user differently. Song-level statistical diversity measures ranked the user B with the least number of *Distinct Songs* as the most specialist whilst the artist-level statistical diversity measures chose user A with the least number of *Distinct Artists*. Geometrical measures both chose user G as the most specialist. This is probably the case because this user listens only to the popular songs in a Pop - Alternative Rock spectrum whilst the users A and B, have at least one song in a rather different genre such as Soul, Folk (user A) and Country (user B). Overall, both Shannon entropies and both geometrical measures in different orders named the users *A, B, G* as the most generalists, agreeing with our intuition.

Geometrical measures consider user B to be more specialist than the user A potentially because user B listens only to the popular songs, which are, arguably, rather similar. However, some might not agree with this, especially, since 53% of the time user A listens to the same artist (Killers). The Hill number, however, had a rather different ranking which directly reflects existence or absences of few songs being played substantially more by the same user as discussed in sections 7.2 and 8.2.

As with the specialist users, Shannon entropies and the geometrical diversity measures places users *C,D,E,F* in various orders as more generalist which agrees with our intuition. Hill number, again gave different rankings for the same reason as before. Regarding the most generalists user, all the measures, apart from the $D^{TS-N}(S_i)$ agree that the user F is the most generalist. $D^{TS-N}(S_i)$, on the other hand, ranks the user D as the most generalist. From the perspective of the statistical measures, user F had the largest number of *Distinct Songs* and *Distinct Artists*, thus this conclusion is not surprising. The interesting part is the different conclusion reached by the two geometrical measures. On one hand, the user F listens to more songs and artists of various style, yet she listens mostly to some subgenre of Rock music with lesser appearances of Electronic music and then much less often - Pop music. User D, on the other hand, listens to fewer different songs and artists, yet more evenly. The genres of these songs are in rather equal amounts Rock, Electronic and Pop. Thus, one could consider the TS-SS with the normalised song vectors as more correct if the even listening pattern of the distinct type of songs is a feature one would describe a canonical generalist with. If instead, a wide coverage of distinct styles of music, albeit with less even spread, is preferred as a feature of the generalist, then the Generalist-Specialist score might be a better option.

Overall, the reason behind the different ranking of users by the statistical measures is clear and easy to understand. When it, however, comes to the geometrical measures, we are faced with a "black-box problem" common in Machine Learning field when it is hard to understand what exactly leads the algorithm to label the users differently. Nonetheless, we can use the theoretical properties to at least grasp some meaning behind the disparity in the results. Owing to the fact the $D^{TS-N}(S_i)$ score takes the magnitude differences of the $\vec{x}_j$ and $\vec{\mu}_i$ into the account, it is likely that the scores would have a higher chance of labelling a person who listens to a lot of similar songs as well as some different songs (like users E and F) as more generalist. This is because the presence of different style songs might push the centroid to be further away from the cluster of the similar songs in the embedded space and albeit the angle between the centroid and the song vector might remain small,

the difference in magnitudes could, in theory, lead the songs in a cluster to contribute higher weights thus increasing the difference in magnitudes between $\vec{x}_j$ and $\vec{\mu}_i$. Visually, the described information could lead the *user preference* vector to move from $\vec{\mu}_A$ to $\vec{\mu}_B$ as in Figure 3. From Table 11, we see that $D^{TS-N}(S_i)$ labels user D, who did not show some loyalty to one particular type of songs, as the most generalist. Users E and F, on the other hand, had some expressed affinity to the Latin or Alternative Rock music thus forming the song clusters within their playlists.

In order to test the hypothesis that the $D^{TS-N}(S_i)$ score is more sensitive to song clusters than $D^{GS}(S_i)$, some fictional playlists are created. The first pair of the playlists contains the following songs:

$\boldsymbol{A_1}$ playlist:

- Madonna - *Holiday*
- Madonna - *Hang up*
- Beyoncé - *If I was a boy*
- Lady GaGa - *Alejandro*
- Sepultura - *Dead Embryonic Cells*

Playlist $\boldsymbol{A_2}$ contains all the same songs but Sepultura - *Dead Embryonic Cells* is replaced with Beyoncé - *Sweet dreams*. This thus makes $\boldsymbol{A_1}$ playlist to contain mostly female Pop singer songs with one Brazilian thrash metal song. $\boldsymbol{A_2}$ playlist, on the other hand, has only female Pop singers thus forming a stronger cluster than $\boldsymbol{A_1}$.

The second pair of playlist mostly consists of Rap music:

$\boldsymbol{B_1}$ playlist:

- 2pac - *Staring through the rear window*
- 2pac - *So many tears*
- Eminem - *The real slim shady*
- Public Enemy - *By the time I get to Arizona*
- Igor Stravinsky - *Le Sacre du Printemps*

Now, playlist $\boldsymbol{B_2}$ replaces Igor Stravinsky with Public Enemy - *Rebel without a cause* thus forming a stronger cluster than $\boldsymbol{B_1}$.

We now calculate the $D^{GS}(S_i)$ and $D^{TS-N}(S_i)$ scores for both pairs of the playlist and compare how the score change when the cluster is strengthened. We assume that all the songs in all the playlists are listened to exactly once. The resulting scores and the percentage changes are presented in Table 12.

Playlist	$D^{GS}(S_i)$	% more specialist	$D^{TS-N}(S_i)$	% more specialist
$\boldsymbol{A_1}$	0.5272	2.71%	1.0472	4.60%
$\boldsymbol{A_2}$	0.5415		0.9990	
$\boldsymbol{B_1}$	0.5305	6.05%	1.0529	10.02%
$\boldsymbol{B_2}$	0.5626		0.9474	

Table 12: Changes in the geometrical scores

We can see that the elimination of one song outside the cluster has a substantially larger percentage effect on the $D^{TS-N}(S_i)$ score than on the $D^{GS}(S_i)$ thus giving some evidence that the TS-SS score implemented on the normalised song vectors is more sensitive to the clusters of the songs than the Generalist-Specialist score. This is not surprising knowing that the odd songs outside the cluster tend to move the *user preference* vector away from the cluster thus increasing the difference in magnitudes as well as angles. Thus, this must be taken into account when deciding to use either of the scores.

10.3 Diversity measures with implicit popularity

We now repeat the same analysis as above for the song-level relative entropy, artist-level relative entropy and the TS-SS score implemented on non-normalised song vectors. We note that the Kendall's τ between the $D^{TS-R}(S_i)$ and $D^{KL}(S_i)$ as well as between $D^{TS-R}(S_i)$ and $D^{KL}(Z_i)$ was limited and took the values $\tau = -0.4575$ and $\tau = -0.4169$ respectively. Contrary to the measures discussed before, this time the hierarchical relative entropy was in lower absolute concordance than the non-hierarchical relative entropy. The resulting scores for each user are presented in Table 13.

User	$D^{KL}(S_i)$	$D^{KL}(Z_i)$	$D^{TS-R}(S_i)$
A	7.4629 (4)	6.3210 (4)	622.2460 (5)
B	6.6560 (5)	5.6904 (5)	3675.9422 (7)
C	8.9850 (2)	6.3556 (3)	0.0069 (2)
D	12.9286 (1)	10.1730 (1)	0.0000001 (1)
E	8.9202 (3)	7.7301 (2)	7.7705 (4)
F	6.6422 (6)	5.0454 (7)	0.1748 (3)
G	6.1419 (7)	5.0795 (6)	1586.1161 (6)

Table 13: Diversity score with popularity for the specific users (rank)

Beginning from both relative entropy measures, these measures disagree on the ranking of the users G, F and E, C. However, the $D^{KL}(S_i)$ score of users C and E as well as the $D^{KL}(Z_i)$ score of the users F and G are only marginally different and thus the difference in order does not imply a strong difference between the users. User D was chosen to be the most non-mainstream specialist by both relative entropies. This conclusion was most likely reached because user D has indeed a non-mainstream taste and does not listen to any popular songs. Yet, the user seems to be more

generalist than specialist thus empirically showing limitations of the $D^{KL}(\cdot)$ scores. Interestingly, user A was not labelled among the top 3 most mainstream generalist users suggesting that the Kullback-Leibler divergence measures contain more information than solely the fraction of popular songs in the playlist.

The ranking of the users is slightly more different with the $D^{TS-R}(S_i)$ score. Firstly, we can see a clear split between relatively big values (users A, B, G) and small values (users C, D, E, F) in terms of the proportion of popular songs in the playlists. The largest disparity in the ranking between the relative entropy and the TS-SS score comes with the user F who is ranked as being close to the mainstream generalist by both Kullback-Leibler divergences and ranked as closer to the non-mainstream specialist according to the TS-SS score. This labelling is despite the fact that the user F, in general, listens to a lot of songs and among those songs are a great number of popular songs. User E is considered to be more mainstream generalist than the user F despite having a smaller number of both songs and popular songs in the playlist. This is a somehow surprising result and leads to questioning the reliability of the $D^{TS-R}(S_i)$ score. Nonetheless, $D^{TS-R}(S_i)$ seems to recognise the pure generalist nature of the users better than the Kullback-Leibler divergence with the top 3 most generalist users being A, B and G as expected. However, it could also be the case that the TS-SS score recognises the fraction of popular songs and not the generalist nature *per se*.

11 Application: Recommender System

At the crux of the motivation behind the diversity measures is the idea that in general, the recommender systems work poorly for the generalist users. A team working with the Spotify data has observed that once a user becomes more diverse, she shifts from algorithmic recommendations towards organic individual song exploration [1]. At the same time, it has been observed that people using recommender systems tend to become less diverse in terms of their music taste. This is an undesirable consequence bearing in mind that the longevity of a user is strongly associated with the music taste diversity of a user, at least in Spotify. Thus, there is a need to create recommender systems that would offer a suitable algorithm for people already having a diverse music taste. Contrary to the standard recommender systems, a desirable recommended system would have to promote a more diverse taste. Whilst the construction and the evaluation of the recommender systems is beyond the scope of this work, we could use finding by Anderson et al (2020) [1] to evaluate the diversity measures proposed and thus show them in reference to an application.

The evaluation method described below aims to determine how well the diversity measures explain the extent to which it is difficult to *guess* or recommend a set of songs a user is likely to listen to. This would serve as an evaluation of our diversity measures using a very naive recommender system. We expect the users to whom it is difficult to recommend a song to also be of a generalist type. It must to noted, that the results are likely to depend on the recommender system itself, thus another purpose of this application is to propose a potential pipeline of evaluation which could be replicated using any other recommendation system. To measure the difficulty to recommend, a very

simple recommender system prototype is exploited:

1. Given the playlist of each user i, split that playlist into the testing and training datasets holding an 80/20 % split approximately.

2. Using each user's training dataset, calculate a *user preference* vector of this user. This *user preference* vector is defined in the same way as the centroid $\vec{\mu}_i$ in section 4.2.

3. For each user, simple cosine similarity is calculated between the user's *user preference* vector $\vec{\mu}_i$ and all the songs $\vec{x}_j$ in her testing dataset.

4. The average of the cosine similarities for each user would denote the *Difficulty to recommend* variable which we name `Proximity`.

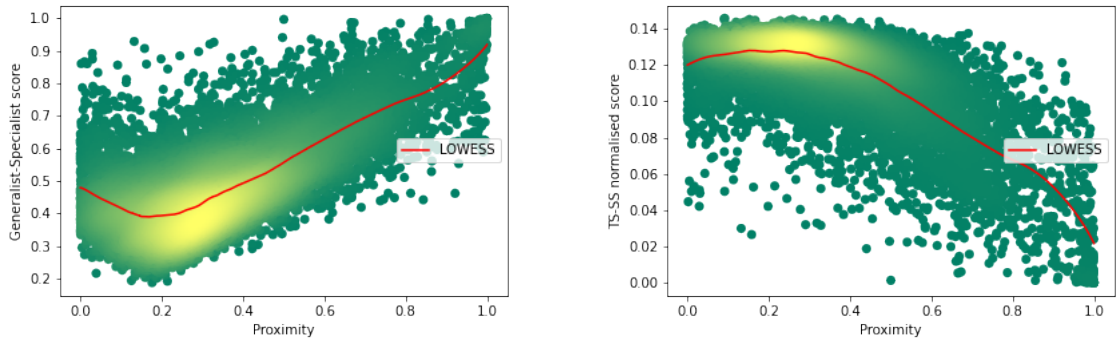
The same embedded song vectors $\vec{x}_j$ are used as before in this project. Moreover, the normalised vectors $\vec{x}_j$ are used owing to the implicit song popularity associated with the length of a song vector as well as arguably better TS-SS score performance using the normalised data as opposed to the non-normalised data.

Once the `Proximity` variable is constructed, we then regress it on all the diversity measures introduced. Note, owing to the nature of the $D^{GS}(S_i)$ score and the underlying cosine similarity in its definition, it is expected that the regression would be the strongest and the most significant with this score. Thus, we should make any conclusions regarding the Generalist-Specialist score carefully.

After plotting the scatter plots between the `Proximity` variable and all the diversity measures, it was observed that only the geometrical measures constructed on the normalised data, namely $D^{GS}(S_i)$ and $D^{TS-N}(S_i)$, had some correlation. Thus, the statistical diversity measures as well as $D^{TS-R}(S_i)$ will not be analysed in the regression context. Nonetheless, the scatter plots between these measures and the `Proximity` variable are available in Appendix H.

The scatterplots between the $D^{GS}(S_i)$, $D^{TS-N}(S_i)$ and the `Proximity` indicated a non-linear relationship. Various preprocessing techniques such as logarithmic, Box-Cox transformation, different degrees of polynomials of the scores were tested, yet none of them managed to lead to a linear relationship with `Proximity`. This rendered the usage of a linear regression inappropriate. Thus, a Locally Weighted Regression referred previously as LOWESS is implemented to summarise the relationship between the scores and the `Proximity` variable. The resulting scatter plots together with the LOWESS curves are available in Figure 20.

From Figure 20 we see that, in general, the larger value of `Proximity` the more likely it is that the person is considered as relatively specialist by both scores. In other words, the better the *user preference* vector of each user helps to detect songs the user listens in the testing set, the more likely it is for that user to be generalist, as expected. Interestingly, the LOWESS curves are not monotonic suggesting that the relationships between the scores and the `Proximity` are more complex. In fact, the positive association between `Proximity` and the generalist nature is only present when the `Proximity` takes a value larger than 0.2. In the case of the $D^{TS-N}(S_i)$, the slight curvature of

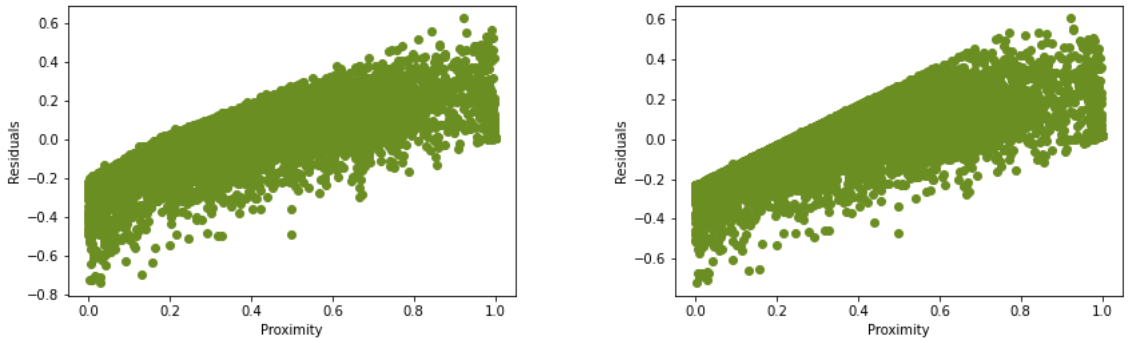


(a) Generalist-Specialist score **(b)** TS-SS score on normalised data

Figure 20: `Proximity` vs the geometrical diversity measures

the LOWESS line can be due to the lower mass at the low regions of `Proximity` with the mass gradient in the scatterplot showing a rather monotonic relationship. Yet, with the $D^{GS}(S_i)$ score, the non-monotonic relationship is clear. This could imply that the Generalist-Specialist score is weaker at detecting extremely generalist users.

The relationships are, however, not tight and present large residuals. The Root Mean Squared Error (RMSE) for the $D^{GS}(S_i)$ and $D^{TS-N}(S_i)$ scores are 0.1544 and 0.1548 respectively. This indicates that to some limited extent both of these scores explain the degree of difficulty to recommend a song to a user, yet not fully. On average, both scores would miss-predict users `Proximity` value by 0.155 points. Furthermore, the RMSE values, we see that both of the measures explain the difficulty to recommend a song to a similar extent. Bearing in mind that the `Proximity` variable was constructed using the average cosine similarity, and thus resembles the $D^{GS}(S_i)$, the fact that $D^{TS-N}(S_i)$ achieved very similar RMSE can be perceived as a sign that the latter score is at least not inferior to the former in estimating the average difficulty to recommend a song.



(a) Generalist-Specialist score **(b)** TS-SS score on normalised data

Figure 21: `Proximity` vs residuals

Examining the residuals versus true value plots in Figure 21 gives us slightly more information. It seems that $D^{TS-N}(S_i)$ has a narrower residual range than the $D^{GS}(S_i)$ when considering the users to whom it is especially hard to recommend songs (low values of `Proximity`). This agrees with the clear absence of monotonically increasing relationships between the $D^{GS}(S_i)$ score and `Proximity` (Figure 20) and the idea that the perhaps Generalist-Specialist score struggles to detect extremely specialist users. At the high end of the `Proximity` values, the range of residuals yielded by using LOWESS with both scores seems to be similar. Thus, this gives some evidence that the $D^{TS-N}(S_i)$ score is marginally more able to detect generalist users than $D^{GS}(S_i)$. However, this can be claimed only if we consider this particular prototype of a naive recommender system. In addition, the residuals, as well as RMSE, depend on the modelling choices and it is possible that other non-linear models such as Support Vector Regression, K-nearest neighbours regression, etc. would be able to use the information given by the diversity scores better and thus yield higher power in explaining the difficulty to recommend a song for each user.

12 Conclusions

The success of algorithmic recommendations is influenced by the diversity of a user taste thus creating a need to understand the spectrum of the user taste diversity better. However, a plethora of ideas of what are the canonical features of a diverse user taste directly translates into an abundance of measures aiming to calculate the degree of diversity. In this work, four different statistical diversity measures and two geometrical measures were evaluated. Statistical measures have an advantage over the geometrical ones because the definition of diversity a measure is adopting is very clear. For instance, Shannon entropy defines diversity in terms of the *surprise* associated with a playlist distribution. Kullback-Leibler divergence, or relative entropy, similarly, defines diversity as a *relative surprise* when we expect to see some other playlist distribution instead of the one actually observed. For Gini-Simpson index the diversity is simply the probability of getting two items of a different type when sampling from the playlist with the replacement. Finally, Hill number understands diversity as a number of equally abundant classes needed to yield the average proportional abundance obtained in the dataset. Nonetheless, none of these definitions of diversity manages to measure how similar elements in a set (i.e. playlist) are. To partly control for this issue, the statistical measures could be constructed on the hierarchical data. We thus implemented them on the artist-level data and from their relationship with the geometrical diversity measures noticed that the hierarchical statistical diversity measures managed to understand better how similar songs in the playlist are.

To tackle the issue and make sure that the diversity measure can recognise how similar songs within one playlist are, geometrical diversity scores were introduced. Generalist-Specialist (initially introduced by Anderson et al [1]), as well as TS-SS scores, work by measuring the geometrical distance between the songs once the songs are assigned some vectors. The key difference between the two geometrical diversity measures is that the Generalist-Specialist score only takes the angle between the two vectors into account whilst the TS-SS score also considers the difference in mag-

nitudes between the vectors. Bearing in mind the properties of the song vectors constructed using Hierarchical Poisson Factorisation, it was decided to implement the diversity measures on unit-norm vectors. This, as it was explained, limits the power of TS-SS score, yet not fully defeats it bearing in mind that the distance is calculated between each song in the user's playlist and the *user preference* vector, which does not have to be of a unit-norm. During the evaluation, it was observed that the TS-SS is more sensitive to the clusters of song vectors than the Generalist-Specialist score. This implies that the TS-SS score is expected to label a person as more diverse if only one of the songs in her playlist is of a different type than the Generalist-Specialist score. The key drawback of these measures is a very awkward interpretation as well as a dependence on the quality of the embedding space used.

Overall, most of the measures described above are good or correct for certain applications. Thus, it is always a context-specific question which measure ought to be used. For example, Kullback-Leibler divergence as defined here will incorporate a notion of how many popular songs a user has in her playlist whilst Shannon entropy is the most associated with the number of distinct songs or artist in the playlist. Hill number, on the other hand, is very sensitive to the cases when one song or one artist is listened to substantially more than the others thus labelling a user as having a less diverse taste. However, in the context of recommender systems and understanding which users are difficult to recommend to, the geometrical diversity measures are likely to work the best. Yet, for some custom recommender systems, this might not be the case.

A great share of ideas was beyond the scope of this project. For example, other hybrid diversity scores could be constructed by using the statistical and geometrical scores in combination or combining the geometrical measures such as Euclidean distance and the cosine-similarity only. Furthermore, simple Machine Learning technique Shared K Nearest Neighbours [11] has the potential to serve as an interesting diversity measure and could be evaluated in the future. More advanced statistical techniques such as the Earth Movers Distance, sometimes called Optimal Transport, should also be attempted in the diversity measurement problem. These measures are likely to give us yet another way of thinking about the diversity and thus creating more possibilities to detect the diverse sets and use this information to improve the recommendation system algorithms and offer a better online platform experience for those of us who like to explore.

References

[1] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, pages 2155–2165, 2020.

[2] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[4] V. Crupi. Measures of biological diversity: Overview and unified framework. In *From Assessing to Conserving Biodiversity*, pages 123–136. Springer, 2019.

[5] S. Das, O. Egecioglu, and A. El Abbadi. Anonymizing edge-weighted social network graphs. *Computer Science, UC Santa Barbara, Tech. Rep. CS-2009-03*, 2009.

[6] david cortes. *Hierarchical Poisson Factorization.* `https://github.com/david-cortes/hpfrec`.

[7] P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

[8] A. Heidarian and M. J. Dinneen. A hybrid geometric approach for measuring similarity level among documents and document clustering. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 142–151. IEEE, 2016.

[9] M. O. Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[10] S. L. Hill, M. Harfoot, A. Purvis, D. W. Purves, B. Collen, T. Newbold, N. D. Burgess, and G. M. Mace. Reconciling biodiversity indicators to guide understanding and action. *Conservation Letters*, 9(6):405–412, 2016.

[11] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *International conference on scientific and statistical database management*, pages 482–500. Springer, 2010.

[12] M. Mihajlovic and N. Xiong. Finding the most similar textual documents using case-based reasoning. *arXiv preprint arXiv:1911.00262*, 2019.

[13] R. N. (originator). Kendall tau metric. In *Encyclopedia of Mathematics*, 2001. `http://encyclopediaofmath.org/index.php?title=Kendall_tau_metric&oldid=50721`.

[14] H. Schreiber. Improving genre annotations for the million song dataset. In *PProceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[15] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

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

[17] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Acm sigir forum*, volume 51, pages 176–184. ACM New York, NY, USA, 2017.

[18] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (AAAI 2000)*, volume 58, page 64, 2000.

[19] I. Waller and A. Anderson. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, pages 1954–1964, 2019.