Imagine All the People: Characterizing Social Music Sharing on Reddit

Abstract

Social groups form and develop their collective identities through cultural activity, such as music sharing. Although online community platforms have been extensively studied, the social context of music sharing in online communities has been difficult to include. Datasets derived from online platforms often do not contain structured data corresponding to the music being shared on them, and music datasets are mostly either acoustic in nature or contain traces of private listening behavior. In this work, we join these two literatures with a large-scale analysis of social music sharing on Reddit. We find significant patterns in how artists are publicly discussed, and cluster them into "social genres". These often follow musical genres, but there is also a significant amount of extramusical sharing. We characterize this by developing two scores, the GS-score and a meme score, that capture the extramusicality and meme-iness of an artists' sharing, respectively. Finally, we study the cultural contexts that musical artists are shared in, and find fine-grained distinctions between social genres and artists that align with widely held intuitions.

Introduction

Music binds us together. It is universal, being present in every observed society (Mehr et al. 2019); it spans the full range of the human experience (Blacking and Nettl 1995); and it is a critical component in the formation and maintenance of social identity (Tarrant, North, and Hargreaves 2002; Gregory 1997). As the sociomusicologist Simon Frith contends, "Social groups... only get to know themselves *as groups* through cultural activity." (Frith 1996).

Online social groups also get to know themselves through cultural activity. The sharing of content, memes, jokes, etc., is a main way in which online communities express their identities. However, while the use of text, memes (Chen 2012; Bauckhage 2011), and images (Hu et al. 2014) in online group contexts have been the focus of rich strands of literature, the sharing of music has been less well-studied. How are musical artists and genres shared in online communities?

The study of music, on the other hand, has recently been augmented with computational techniques and models, as well as the availability of large-scale datasets of music consumption behavior. Millions of songs can be categorised into fine-grained musical genres using automated and semi-automated techniques, signal processing methods can perform complex tasks on acoustic data, and higher-level semantic qualities of songs can be extracted from the music alone. However, the *social contexts* in which music is invoked has not been subject to the same level of analysis. How can we enrich our understanding of music by examining the contexts in which it is shared?

In this work, we join these two literatures by conducting a large-scale analysis of the online social sharing of music. We study invocations of music on Reddit, one of the world's largest online social platforms, in which people publicly discuss various topics in communities called "subreddits". We focus on three main areas of interest. First, how are artists shared on Reddit in relation to each other? Are specific artists shared in similar ways to others? We show that sharing patterns cluster artists together into novel "social genres". The similarities and differences between how musical artists are invoked in public discussions are related to, but sometimes distinct from, their musical similarities and differences. In many cases, musically dissimilar artists appear in similar social contexts and are thus grouped together by our method. Second, can we infer cultural features associated with musical artists by harnessing information about the underlying communities where they are shared? For example, if a specific artist is shared largely in left-wing communities, it can be inferred that the artist has some association with the left-wing, whereas if an artist is shared largely in communities for high-school students, then they are likely to have a younger following. By building on previous work using neural community embeddings to identify cultural axes of meaning in Reddit, we quantify the cultural contexts artists are shared in. These features can act as a new source of information for both the artists and researchers to deepen their understanding of music listeners. Finally, how can we measure the extent to which the social sharing of certain music is "extramusical", i.e. driven by factors that seem unrelated to the music itself? Are there artists that are shared in vastly different ways than what would be expected from their music? This relates to the broader question of what artists became culturally significant outside their musical domain and had an impact in a broad range of areas.

One manifestation of this is when an artist becomes associated with a meme or a joke online. We develop a method to quantify how musically-driven or socially-driven the sharing of an artist's music tends to be.

Related Work

Our research aims to connect the study of online communities with music sharing behavior. We thus draw on two areas of existing work: large-scale analyses of online communities and behavioural analyses of music.

The current efforts of understanding human-music interaction have been largely limited to survey-based analyses of music. In (North and Davidson 2013) the authors attempts to add to the broad literature sociologists have been developing relating how music taste differs between social groups (Fox and Wince 1975). The authors add nuance to how musical taste differs by employment, education, and global region, developing off of their past work comparing how music affects travel, money, and health (North and Hargreaves 2007). Other academics go even further in analyzing social interaction with music by attempting to answer the age-old question of whether or not music is universal amongst cultures, and the extent to which music is diverse (Mehr et al. 2019).

In addition to comparing cultural factors, research has developed an understanding of how listening habits correlate with behavioural habits. In (Krause and North 2018) the authors analyze how music listening habits change by season, and demonstrate that different moods of music are listened to in different weathers. A larger scale analysis was done to measure how diurnal and seasonal patterns affected someone's music preferences and compared how changes in time, age, and location of the listener affected music listening decisions (Park et al. 2019). An analysis of how music tastes evolve over a person's life was done on Spotify, where the authors analyzed how moving from city to city affected ones music tastes (Way et al. 2019). Our work complements these efforts by contributing a large-scale behavioral analysis of how music is *publicly* discussed and shared, as opposed to how it is consumed in private.

Data

Our main data source is all Reddit submissions and comments from 2006 to the end of 2019 (Baumgartner et al. 2020). We focus on the subset of comments and submissions which contain music sharing. While doing a search for posts containing artist names is the simplest approach, this risks finding mentions of other concepts with the same name as the artist. For example, a search for the famous rapper Pitbull provides an instance of someone sharing a photo of the pitbull bread of dog. For this reason, we filter comments and submissions to shares that link to a Spotify track or album, or the YouTube video for an artist's song.

While extracting posts containing Spotify links is easily achieved by searching for the `open.spotify.com` domain, the same method cannot be used for YouTube, as many YouTube links point not to music videos but other types of video. To create a list of YouTube links which point to a song, we first collect only Spotify shares from our dataset. We then use the Spotify API to extract artist and track information. This included genres, popularity, track name, artist name, album. For the album shares we collect all tracks in the album and all metadata for those tracks. In total, we end up with 536,860 unique tracks from the Spotify shares.

To link each of these tracks with a YouTube video we scraped the associated Last.fm link and extracted the YouTube link. Last.fm has links to 194,067 of our tracks (36%). Some links did not align with the artist due to the scraping occasionally picking up the wrong link from the Last.fm website. For this reason, if a YouTube video was associated with more than one track, we validated the link by extracting the title of the YouTube video and checking if it matches the song title.

Using this list of YouTube music videos, we find all comments and submissions on Reddit containing a link to a music video or Spotify. After removing artists that were shared less than 20 times, because too few data points risked painting an inaccurate picture of the artist, our final dataset contained 1.3 million music shares on Reddit.

Characterizing Sharing with Embeddings

Reddit Embedding

Reddit consists of many communities called 'subreddits' which act as forums for users to discuss, share and present ideas. Communities exist for a wide array of topics, ranging from sports to music, politics to mathematics, and jokes to photos, and music sharing occurs not only in music-dedicated communities but is widespread across the platform. To understand the semantic relationships between the communities in which music is shared, we use a 'community embedding', an adaptation of word embeddings used in natural language processing. We develop our community embedding according to the method outlined by Waller and Anderson (Waller and Anderson 2019). Each community is represented as a 150-dimensional vector in a Euclidean space, and the similarity between two community vectors is related to the similarity of the user bases of the two communities. Notably, textual content of comments and submissions is not used in the creation of this community embedding, so the similarity of two communities is related only to *who* comments in a community, not *what* they comment. Usefully, prior work has found semantic relationships are preserved in this space– for example, simple vector arithmetic will correctly answer the analogy `Torontobluejays - toronto + nba` $\approx$ `torontoraptors`. A community embedding is thus a useful tool to quantify the similarity of disparate communities on the site.

Artist Embedding

This community embedding is used to create a vector for each of the artists in our dataset using the following process. Let P represent the dataset of all music sharing posts (where posts can be either comments or submissions). P_a is a subset of P restricted to sharing of artist a. Then for each $p \in P_a$ we define $\vec{p}$ to be the vector of the community in which the

comment or submission was posted. Then, an artist's vector $\vec{a}$ is:

$$\vec{a} = \sum_{p \in P_a} \frac{\vec{p}}{|P_a|} \tag{1}$$

i.e. the arithmetic mean of the vectors of the communities where it was shared, weighted by the number of shares in each community.

This artist vector is a function of only where the artist is shared and how often they are shared there. Thus, these vectors can be used as a new method of comparing artists, with the goal of measuring which artists are shared similarly to each other, and which are shared very differently. The cosine similarity metric is used for this purpose, since it calculates the similarity of the vectors, outputting a value between -1 and 1; -1 if the vectors are opposite, 0 if they are orthogonal, and 1 if they are equal. Thus, for any two artists a_1 and a_2 the similarity between them equals the following:

$$\cos(\vec{a_1}, \vec{a_2}) = \frac{\vec{a_1} \cdot \vec{a_2}}{||\vec{a_1}||||\vec{a_2}||} \tag{2}$$

This social similarity metric facilitates several new levels of analysis that we use to uncover important social relations.

Artist Similarity. Figure 1 displays the similarity of the 30 most shared artists on Reddit. Examining the similarity between artists in this space reveals new relationships that are not obvious by performing a purely sonic comparison. For example, it is clear that Linkin Park, Daft Punk, Rebecca Black, and Foo Fighters are similar in the space. However, they share almost no Spotify genres in common. However, artists that belong to the same musical genre are often also very similar to each other. For instance, artists like Kanye West, Eminem, and Childish Gambino have a very high similarity, and Kendrick Lamar lights up with them on the off-diagonal. Moreover, starting from Tame Imapala to Red Hot Chili Peppers there is a group of similar classic rock and pop artists. The colour bar on the side and top represents the age of the most popular song by an artist; the darker the colour, the newer the artist. Interestingly, similar artists tend to belong to similar periods.

Social Genres

It is clear that the similarity of artists using this method is related to, but sometimes distinct from, their musical similarities and differences. We cluster artists based on this similarity to create groupings of artists which are similar in how they are *invoked*, which we term 'social genres', and examine how these genres differ from traditional musical genres.

Creating social genres. To find social genres, we use a hierarchical clustering algorithm with fifty clusters. Silhouette scoring and the elbow-method was used to determine the number of clusters. Hierarchical clustering was used for clustering since it provides an intuitive result on how clusters are formed, with a clear decision tree.

We named social genres based on three properties of the cluster: the top communities, the top Spotify genres, and

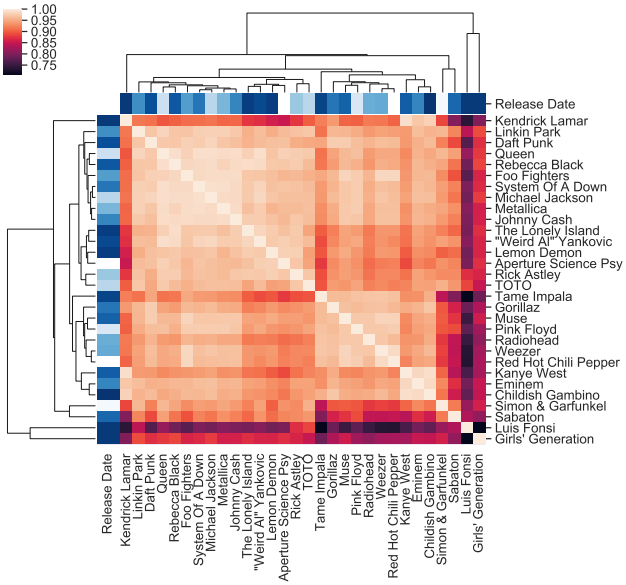

Figure 1: Heatmap of artist similarity (as measured by cosine similarity of the artist vectors) for the top 30 most shared artists on Reddit, including a dendogram of our clustering. The blue bars reprseent the age of the most popular song by an artist (darker is newer.)

Social Genre 7: pop

Spotify genres		Reddit communities		Artists	
Name	Count	Name	Count	Name	Count
pop	20471	Music	26141	St. Vincent	9096
dance pop	17539	popheads	21044	CHVRCHES	8953
electropop	13118	AskReddit	15234	Calvin Harris	7744
post-teen pop	8932	tipofmytongue	11461	Lorde	7497
art pop	7865	listentothis	8267	Grimes	6990
indietronica	7588	indieheads	5809	Azealia Banks	6171
indie pop	7438	hiphopheads	5112	Kylie Minogue	5660
metropopolis	6008	ifyoulikeblank	3581	Taylor Swift	5217
escape room	5804	videos	3257	Janelle Manáe	5130
indie poptimism	5068	CasualConv.	1926	Perfume Genius	3492

Table 1: The Spotify music genres, subreddits, and artists that occur most often within the social genre we call "pop".

the artists in the cluster. First, the traditional genres that each artist belongs to were extracted from Spotify. Next, we examined the top ten artists, genres, and communities that made up each social genre in a plot similar to Table 1, which was used to name social genre 7. If one Spotify genre dominated the social genre, then that genre name was taken as the name of the social genre. Frequently more than one genre was dominated by the same top Spotify genre, for example pop. These genres were then named sequentially, as in pop1 and pop2. However, often there was a diverse array of Spotify genres present in a social genre; at this point, a more in-depth analysis was done of the genre, looking at the communities and artists it is dominated by. A name was then given to the genre based on this assessment.

Analysis of social genres. Examining these social genres uncovers new connections that are difficult to see through

a conventional music analysis. Social genres fall into two broad categories; those that align with traditional musical genres, and new categorizations that expose different relations. To begin, we found many social genres correspond directly with traditional music genres. Pop punk, alternative rock, metal, metalcore, electronic dance music (edm), lofi hip hop, trance, and some examples of pop tended to cluster together in their respective musical genres. Most of the sharing in these categories were dominated by communities that are dedicated to these genres, for example pop punk has a community called "PopPunkers" and metalcore has a community called "Metalcore". However, not all social genres have a clear correspondence to a traditional genre. Often, the social genres were able to capture something that regular genres would have difficulty measuring. For instance, one social genre featured new music makers who are sharing their music to try to gain exposure in communities like `WeAreTheMusicMakers`, `ThisIsOurMusic`, and `Songwriters`. Another social genre featured drag music, and was dominated by music sharing in drag communities on Reddit, namely `RuPaul`. In addition to finding these new cultural aspects, these social genres added nuance to some mainstream genres. For instance, there are a few social genres that were dominated by edm music, but a closer analysis revealed that each of these classes actually represented a unique type of edm. Most of the sharing for the Monstercat edm genre was done in the Monstercat community, which is a Vancouver based edm producer and has a unique flavour of music. In other words, artists in this community were shared because they were either produced by Monstercat, or developed some social association with the brand. There are two last social genres worth noting here. First, genre 21 was able to represent comedic music, it grouped communities that shared memes together. This genre was dominated by the sharing of Luis Fonsi's recent success, Despacito, which developed a meme meaning on Reddit. Finally, genre 48 captured sharing from two communities called `CasualConversation` and `infp`. Both of these communities are platforms to lessen stress and relax; the about section of `CasualConversation` calls it "The friendlier part of Reddit. Have a fun conversation about anything that is on your mind." So, this genre represents this casual type of music. Our full dataset of social genres is available in our data repository[1].

Social genres can also be compared to one another to understand the overall structure of music sharing on Reddit. Figure 2 is a similarity heatmap between the various genres. In the bottom-right corner we notice a rap cluster, which also lights up on the off-diagonal along the other rap genre. Further up we see a brightly lit cluster that consists of electric music. One interesting thing that this heatmap points to is that the international, Asian music tends to cluster together. Starting from Japanese to Philippines there is a square that features Japanese, Korean, and Philippine music. The square right in the middle represents artists from the rock and pop categories who are famous. Finally, in the top left corner, we see singletons, for example underground hip hop (represented as ind. hip hop) that has a unique sharing tendency (though it does light up a little with rap2 and uk hip hop). Additionally we see a few clusters for metal music.

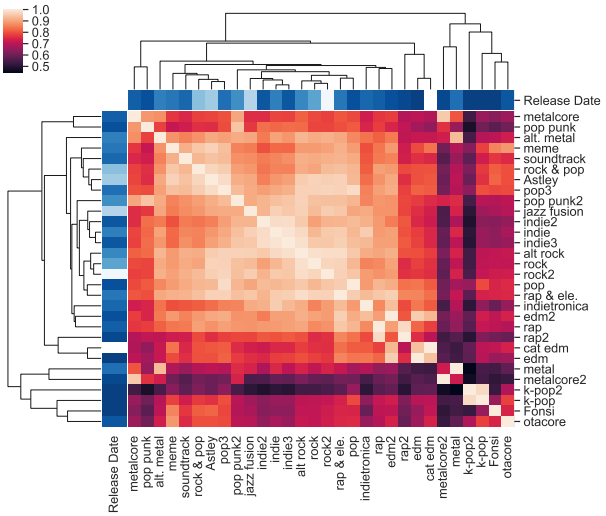

Figure 2: Heatmap of social genre similarity (as measured by cosine similarity of the genre vectors) for all social genres. The blue bars reprseent the age of the most popular song of the artists within the genre (darker is newer.)

Linguistic Analysis

A useful property of cultural sharing is that it includes a linguistic component. Specifically, by embedding the language used in Reddit in the same space, we can learn about the words that are associated with both artists and social genres. One of the most insightful cultural cues is the language we use when describing an idea or topic. Each community on Reddit has a set of words that occur more frequently in it than the average community. These words become signals for that community and can be used to associate words with artists or genres.

Embedding Words. We embed the words using a very similar process to artists. Instead of creating a new embedding, we give each word a vector which is a weighted average of all the communities in which the word is used. Since each word vector exists in the same space as the community and artist vectors, it is possible to directly compare them using cosine similarity. Vectors are created for the top 10,000 words by number of usages on Reddit, and the average is weighted by the positive pointwise mutal information (PMI) of each word with each community, such that each word is a weighted average of the communities in which it appeared more than expected.

Comparing Words to Genres. As both the word embedding and the artist embedding are in the same space, a similarity score can be calculated between these sets of vectors. For example, using the vector for the word "concert", we can calculate the cosine similarity between it and the vector of each of the social genres, and arrive at at an ordered

[1](Link to data repository omitted for anonymous submission)

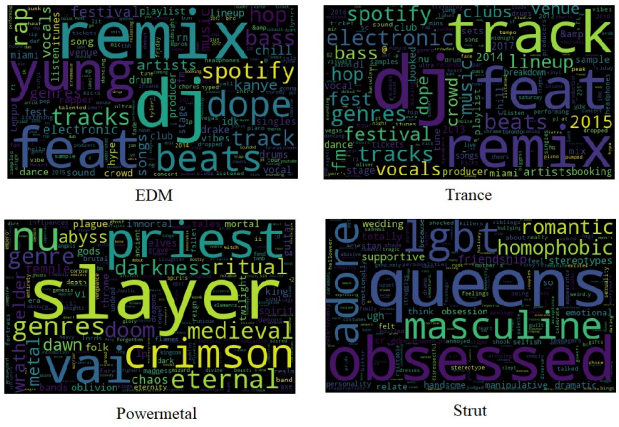

Figure 3: Word clouds of the most similar words to four social genres. Word-genre similarity is calculated by taking the cosine similarity of the word vector and the artist vector.

list of social genres by similarity to the word ”concert”. We perform this calculation for all of the top 10,000 words and present the results of this comparison in Figure 3.

For example, notice the differences between the word clouds of the EDM and trance genres in Figure 3. Both of these genres of music are similar and therefore share a lot of the same words (”dj”, ”remix”, ”beats”, ”spotify”, etc.) However, many words also differ between the two genres. Trance has more words that are connected to festivals and live music, whereas EDM does not. Namely, trance has the words ”festival”, ”crowd”, and ”hop”, which all relate to live music.

In addition to differentiating between two similar social genres, this word embedding enriches our cultural understanding of some social genres with important language. For example, the words most associated with the ‘strut’ or ‘drag’ genre are “queens”, “masculine”, “obsessed”, “adore”, and “lgbt”. In contrast, the words most associated with power metal, a faster and lighter sub-genre of heavy metal, are ”slayer”, ”crimson”, ”priest”, ”medieval”, and ”eternal” (Figure 3).

Cultural Contexts of Online Music Sharing

As previously discussed, community embeddings preserve semantic relationships between communities. This capability can be applied beyond solving analogies by creating cultural axes that can be used to measure the association between a community and a cultural concept. Cultural axes have been previously demonstrated to be an effective and accurate tool in word embeddings (Kozlowski, Taddy, and Evans 2019), and are especially useful in this case as we can project not only communities but artists and social genres onto these axes as well.

Creating Cultural Dimensions

We create six cultural axes: age, gender, sociality, affluence, partisan, and partisan-ness. Each axis is created using two input ‘seed’ communities, which are subtracted to obtain a vector for the axis. The seeds are teenagers and RedditForGrownups for age; AskMen and AskWomen for gender; democrats and Conservative for partisan; vagabond and backpacking for affluence; and nyc and nycmeetups for sociality. We also calculate a partisan-ness axis, which represents how political a community is, by adding the vectors of the two seeds for partisan (democrats and Conservative). We score communities (and artists, and social genres) on these axes by taking the community’s projection on the axis vector. This score reflects how similar it is to the seeds; if it is far more similar to the first seed, it will have a negative score, and if it is far more similar to the second seed, it will have a positive score. A community with a score of zero on an axis is equidistant between the two seeds. We use the seed augmentation and validation strategies outlined in (Omitted 2020) to increase the robustness of the axes and validate their accuracy.

These cultural axes permit us to deepen the analysis of artists and genres by measuring their association with cultural features. Projecting artists onto these cultural axis vectors (age, gender, partisaness, politcalness, affluence) give insights into the associations of the communities in which their music is shared. We conducted two analyses on this cultural level: macro and micro. On the macro level, a comparison was done of how the music sharing community on Reddit differs from the typical user and how the cultural features of music sharing evolve over time. On the micro level we specifically analyzed how various artists and genres projected onto these vectors, giving every artist a score for each of the cultural features.

Macro Analysis. Figure 4 features three rows. The bottom row shows the typical distribution of music sharing. The comment row shows the typical distribution of how all users comment. Finally, the all users row shows the difference between the two distributions; the area in green is where music distribution dominates the typical sharing distribution. In this plot it shows that the music sharing community tends to be less political (partisan-ness), the gender differences are more polarized, and they represent a younger more affluent demographic.

Micro Analysis. In this section we analyze how artists and social genres project onto the various cultural factors. Since each artist vector is normalized, to calculate the similarity between an artist (or a social genre) and a cultural vector requires simply to use the cosine similarity. A higher projection on affluence, for example, would mean an artist tends to be shared in more affluent subreddits, and a lower projection would mean less affluent.

Figure 5 is a ticker plot that shows how the social genres project onto the features. From the age dimension it is clear that high-energy electronic music (hardstyle, edm, monstercat emd) has a younger audience than indie, electronica, and soundtrack (here, electronica differs from edm in that it is more down-tempo and relaxed). On the partisan axis, strut music (associated with drag) is the farthest left. This aligns with the language most associated with strut, featuring words like “lgbt”, “drag”, etc., which also lean to the

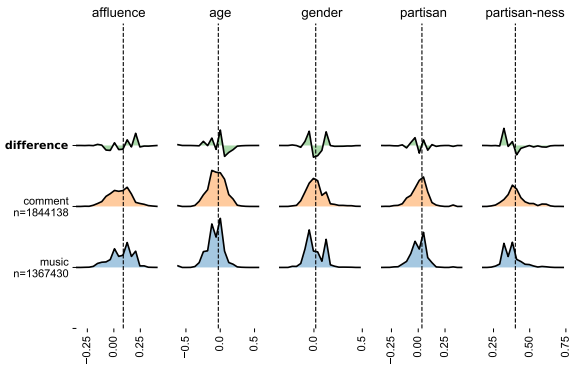

Figure 4: The typical distribution of activity along cultural axes for all comments and instances of music sharing. Two million comments from all comments and two million instances of music sharing are randomly sampled from the dataset; comments for communities not included in the embedding (as they were too small) are discarded. The top line shows the subtraction of the music sharing distribution from the all comment distribution.

left on Reddit. The genre pop4 projected low on the partisan axis, and after diving into this genre it was revealed that the artists were not only shared in the pop communities, but also in the drag and lgbt communities, and so this genre represents more left-wing pop. On the other end of the spectrum, the metal and edm genres tended to be most linked with the right. On the gender dimension, we found that strut also projected most in the female direction. Another interesting phenomena was that Korean and Philippine music was most feminine, and pop4 also projected high on this axis. In line with music on the right, we found metal and electric music tended to lean more male. Finally, it's worth noting that the genre for Casual Conversations was highest on the sociality dimension, which reaffirms the social aspect of this genre.

Moreover, in addition to comparing genres, we projected artists onto the various dimensions. Below we share two of the plots showing audience differences on the age and affluence, and partisan and partisan-ness dimensions. In this visual it is clear that there is a positive correlation between affluence and age. Namely, as the affluence of an artists audience goes up, the audience is more likely to be older. Rappers like Lil Pump, Travis Scott, BROCKHAMPTON, A$AP Ferg, Chief Keef, tend to have a younger less affluent audience. Some alternative and pop artists have an audience with an age that is more centered and a little higher on the affluence axis, artists here include Twenty One Pilots, Arctic Monkeys, Ed Sheeran. Finally, it's interesting to observe that Killer Mike, Dead Kennedys, The Coup, Living Colour, Eric Idle, and Frank Zappa are both older musicians and bands and have an older following.

We also projected the artists onto the partisan partisan-ness axes. The Coup's position farthest on the left and most political aligns with their music and social presence where they have been known for their political activism and criticism of the racial divides in America. Killer Mike also projects on the left and high on the partisan-ness axis becoming an outspoken critic of systemic racism present in the US. Alternatively, the Dead Kennedy's are a little more centric but still very political, which aligns with their music that criticizes and satirizes politicians on both the left and the right. The artist that project farthest on the right were Mike Diva and Knife Party. Analysis into both artists revealed that their shares were mostly in `The_Donald`, a community dedicated to Donald Trump. The shares for Mike Diva were largely driven by his song/video "Japanese Donald Trump Commercial". Knife Party similarly released a song called Centipede which became associated with Donald Trump. Lastly, the least political and partisan is Above & Beyond who produce trance music.

Now that we have analyzed music in the embedding and added the additional cultural features, it makes sense to take a step back and capture this phenomena we've alluded to throughout the paper. Specifically, the emergence of the idea that there were some artists who played a roll within their music genre, but then there was a collection of other artists who had an impact beyond their musical context.

Musical and Extramusical Sharing

As we have seen in the analysis of social genres, how music is shared is not always purely driven by the music itself. Often, there are other meanings that become attached to certain pieces of music, and the patterns in who invokes them in public discussions differ greatly from the patterns in who listens to them in private. We call this important behaviour *extramusical sharing*, and seek to rigorously quantify it here.

A concrete example of extramusical sharing is a song that becomes a meme. The last ten years saw the rise of online memes, which are most often represented in an image format. However, some musical artists have also become associated with memes. This occurs frequently and explains a large amount of music sharing on Reddit, and so we also aim to quantify the "meme score" of an artist. The prototypical example of this is likely Rick Astley's Never Gonna Give You Up. In 2007, the practice of "rickrolling" became popular on 4chan, which consists of making a post that purports to link to something of interest, but which actually directs users to the music video of Never Gonna Give You Up. This caused Rick Astley's song from the 1980s to become an internet sensation in the 2000s, with most of the sharing being driven by the meme association the song developed. Here we presents a novel method of developing this meme metric that can be used to classify artists on how "meme-y" their sharing is.

Measuring extramusicality

To measure the extramusicality of how an artist is shared, we first observe that extramusical sharing is more likely than ordinary musical sharing to occur in dissimilar communities. When artists are straightforwardly shared for their music, as opposed to invoking them at another level of meaning, the sharing often happens in related subreddits. For example,

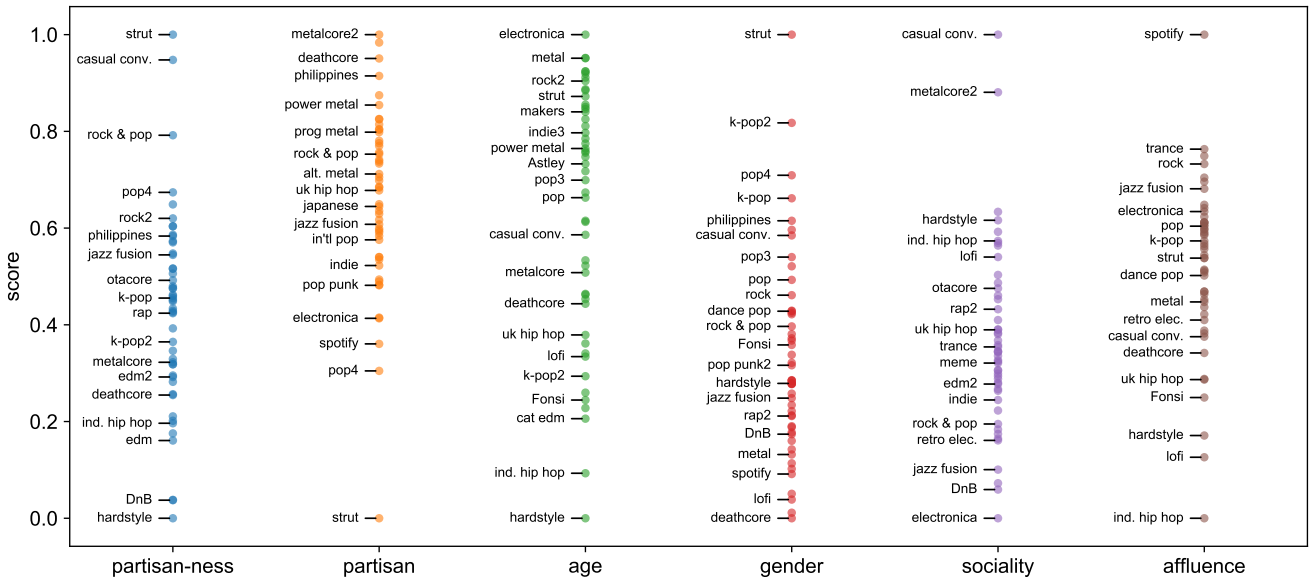

Figure 5: Social genre scores on each of the cultural axes.

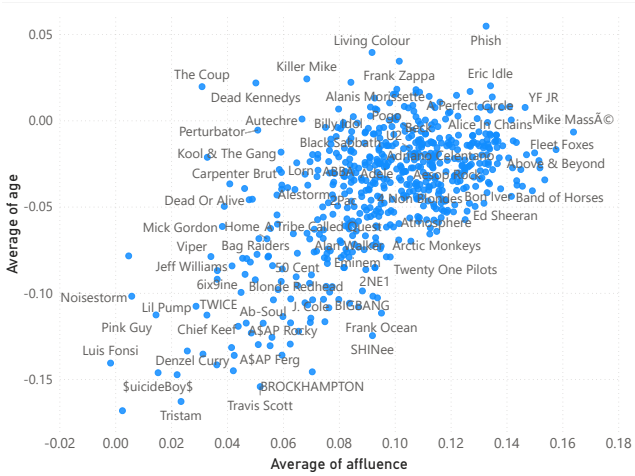

Figure 6: The joint distribution of artists along the age and affluence cultural axes.

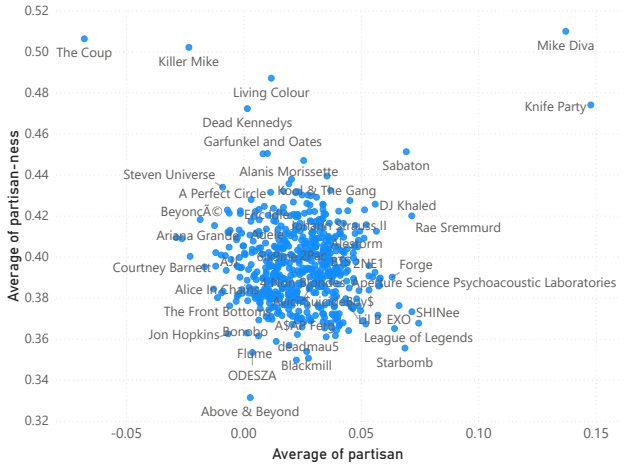

Figure 7: The joint distribution of artists along the partisan and partisan-ness cultural axes.

fans of contemporary rap music will share rap songs in rap-related subreddits. But when there are extramusical factors that affect the sharing of an artist, the communities they are shared in can be quite different from each other, since the reason behind the sharing is decoupled from the music itself. For example, those who share Never Gonna Give You Up only need to be aware of a popular internet phenomenon, and these kinds of users are in a wide variety of communities.

Our approach in this section follows this insight: artists being shared for reasons beyond the music are more likely shared among a diverse range of communities, whereas artists being shared straightforwardly are more likely shared in a more concentrated set of communities. Oftentimes entropy is used to measure diversity, but the issue with it here is it does not factor in how far away communities are within our embedding. For example, let us say we have two distinct artists, artist a_1 and a_2, and that both are shared in two communities. Artist a_1 is shared in a classical community and Mozart community whereas a_2 is shared in a classical community and a sports community. Entropy will measure each of these artists as the same even though the second artists sharing is more spread out within our embedding. For this reason, we use a different measure of diversity that captures community similarity (the GS score (Waller and Anderson 2019)). In this formula, a_i represents the artist vector, J rep-

resents the cardinality of the set of communities the artist is shared in, w_j is the number of times the artist is shared in community j, finally $\frac{\vec{c_j} \cdot \vec{a_i}}{||\vec{a_i}||}$ represents the cosine similarity between the artist vector and the community.

$$GS(a_i) = \frac{1}{J} \sum_j w_j \frac{\vec{c_j} \cdot \vec{a_i}}{||\vec{a_i}||} \quad (3)$$

The GS-score is the average cosine similarity between the artist and all the communities the artist was shared in, weighted by the number of shares in that community and normalized by total shares. After calculating the GS-score for each artist, we restricted our attention to artists in the 70th percentile of sharing popularity or above. However, since artists with more shares have a natural tendency to have a lower-GS score, we use an activity-adjusted GS-score to correct this bias. This activity-adjusted GS Score is calculated as the percentile GS Score within each activity level (for the 80th, 90th, and 100th percentile groups).

The activity-adjusted GS-score presents a very coherent picture of musical versus extramusical sharing. The top and bottom 10 artists are shown in Table 3. The top artist is Kool & The Gang, whose song Celebration is often invoked in online discussion to congratulate someone, or to celebrate a noteworthy event. Some of the most extramusical artists wrote popular songs that were made for films, including Dwayne Johnson's You're Welcome and Kenny Loggins' song from Top Gun. Although Simon % Garfunkel are often shared straightforwardly, "The Sounds of Silence" has become an internet meme due to the lyric "Hello darkness my old friend.", and thus they score very high on extramusicality.

On the other end are artists whose sharing is very specialized. These artists are dominated by rap and international music, since a lot of this sharing is very subreddit-specific. For example, the most specialized artist is SHINee, a k-pop group, whose sharing is almost entirely located in k-pop communities, which occupy a small niche in the community embedding. The other artists tend to be rappers whose sharing is mostly concentrated in hiphop-related communities. This aligns with some of the very specific hip hop social genres we found above. Despite this, some rappers are shared extramusically. These include DJ Khaled, Lil Pump, and Pitbull, who all score within the top 150 artists by extramusical sharing.

Meme Scoring

In this section, we quantify how meme-oriented an artists' sharing is. To measure this abstract concept, we first chose a set of artists whose mentions are almost always made in a meme context (e.g. Rick Astley). We then found the communities that disproportionately shared these artists, and used their corresponding community vectors to define a pro-meme vector. Similarly, we found communities that almost never shared these artists, and used their corresponding community vectors to define an anti-meme vector. By subtracting the pro-meme vector from the anti-meme vector, we arrived at a meme vector where artists that project highly onto it are associated with memes, and artists that project low on it are not. Explicitly, the calculation was done as follows:

	Artist	Top Track	GS Score
Top ten	Kool & The Gang	Celebration	0.15
	Dwayne Johnson	You're Welcome	0.30
	Simon & Garfunkel	The Sounds of Silence	0.45
	Noisestorm	Crab Rave	0.60
	Kenny Loggins	Danger Zone - "Top Gun"	0.75
	Haddaway	What Is Love	0.90
	Tegan and Sara	Closer	1.20
	ABBA	Money, Money, Money	1.34
	The Weather Girls	It's Raining Men	1.49
	LL Cool J	Mama Said Knock You Out	1.64
Bottom ten	SHINee	View	100
	Jay Rock	Vice City	99.9
	Immortal Technique	Dance With The Devil	99.7
	Isaiah Rashad	Heavenly Father	99.6
	Gucci Mane	Both	99.4
	Freddie Gibbs	Bout It Bout It	99.3
	Young Thug	Best Friend	99.1
	Denzel Curry	ULT	99.0
	Mac Miller	Donald Trump	98.8
	The Underachievers	Herb Shuttles	98.7

Table 2: Top and bottom artists ranked by our extramusical sharing score.

1. For each artist-subreddit pair calculate the proportion of shares in that subreddit the artist represents.
2. For each artist calculate the mean and variance of the proportions it represents in various subreddits
3. For each artist-subreddit pair calculate the z-score.

With these z-scores calculated we can find which artists are disproportionately shared in what communities. Since Rick Astley is associated with music meme-iness, we used him to create a music meme vector by taking all subreddits that have a z-score with Rick Astley greater than 2. Then we took the average of those vectors and called it v^+. Then we took subreddits with a z-score less than -1 and take the average of those subreddits and called it vector v^-. By letting $v^m = v^+ - v^-$ we defined the meme vector, and then projected the artists onto this vector.

Table 3 shows the top ten artists by meme score, restricted to artists whose sharing falls in the 90th percentile for ease of interpretability. We discuss the top five artists in depth to demonstrate their respective meme context. The top artist is Luis Fonsi, whose famous song Despacito became a meme on Reddit. Analyzing shares reveals a habit of users replying to sad news by posting something along the lines of "So sad, Alexa play despacito." The vast majority of this sharing took place in the communities `Darkjokes` and `memes`. The song DuckTales by Video Game Players has also become a meme online, especially within the video game playing community. Looking at the shares for this song the contexts range widely, but in general this artist is shared in gaming communities including `gaming`, `GamePhysics`, `Overwatch`, and `Warframe`. One instance of the post was a response to question asking "Which internet meme song

	Artist	Top Track	GS Score
Top ten	Luis Fonsi	Despacito	0.0168
	Video Game Players	DuckTales	−0.150
	Blonde Redhead	For the Damaged Code	−0.161
	Starbomb	It's Dangerous to Go Alone	−0.193
	Rick Astley	Never Gonna Give You Up	−0.202
	Pink Guy	Stfu	−0.214
	Noisestorm	Crab Rave	−0.222
	Kirin J Callinan	Big Enough	−0.228
	Thanks, Computer!	I'm Shithead	−0.231
	Tristam	Frame of Mind	−0.232
Bottom ten	Pavement	Cut Your Hair	−0.534
	Deafheaven	Dream House	−0.525
	Autechre	feed1	−0.525
	Oneohtrix Point Never	Sticky Drama	−0.524
	Janelle Monáe	Make Me Feel	−0.520
	Slowdive	Alison	−0.520
	Purity Ring	bodyache	−0.519
	St. Vincent	Digital Witness	−0.516
	Swans	The Seer	−0.513
	Björk	It's Oh So Quiet	−0.512

Table 3: Top and bottom artists ranked by our meme sharing score.

do you actually like." Blonde Redhead's song For the Damaged Code was sampled in a song that appeared in a popular comedy TV show, and has become frequently used in videos in the community `WatchPeopleDieInside`. Similar to DuckTales, Starbomb's It's Dangerous to Go Alone is also widely within the gaming community, and is a comedic song about Legend of Zelda. Finally, Rick Astley has already been discussed and his song Never Gonna Give You Up was used as the definition for a meme song.

The artists that projected low on the meme vector tend to be award-winning rock, critically-acclaimed singer-songwriters, or electric bands. Pavement, Slowdive, and Swans produce indie and experimental music. St. Vincent and Björk are singer-songwriters, and Autechre and Purity Ring are electronic bands.

Conclusion

In this work, we have conducted a large-scale empirical analysis of social sharing of music on Reddit. We found that artists cluster together into "social genres" based on how they are invoked in public discussions. These social genres often follow musical genres, but there is also a significant amount of extramusical sharing. We characterize these more socially-driven artists by developing two scores, the GS-score and a meme score, that capture the extramusicality and meme-iness of an artists' sharing, respectively. We also enrich our understanding of social genres by connecting them with the words that are shared most similarly. This helps us disambiguate between genres with similar artists but which are shared in different types of communities and contexts. Finally, we study the cultural contexts that musical artists are shared in, and find fine-grained distinctions between social genres and artists that align with widely held intuitions.

This paper helps augment the study of online communities by analyzing the patterns in how they share music, a fundamental component of group identity formation. We also shed light on the online cultural contexts that music appears in, which was previously difficult to measure with only acoustic data or private consumption traces. As music is fundamentally social, and social communities depend on cultural activities to define themselves, we others build on our work to further tighten the link between these two perspectives.

References

Bauckhage, C. 2011. Insights into internet memes. In *ICWSM*, 42–49.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 830–839.

Blacking, J., and Nettl, B. 1995. *Music, culture, and experience: Selected papers of John Blacking*. University of Chicago Press.

Chen, C. 2012. The creation and meaning of internet memes in 4chan: Popular internet culture in the age of online digital reproduction. *Habitus* 3(1):6–19.

Fox, W. S., and Wince, M. H. 1975. Musical taste cultures and taste publics. *Youth & Society* 7(2):198–224.

Frith, S. 1996. Music and identity. *Questions of cultural identity* 1(1):108–128.

Gregory, A. H. 1997. The roles of music in society: The ethnomusicological perspective.

Hu, Y.; Manikonda, L.; Kambhampati, S.; et al. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Icwsm*.

Kozlowski, A. C.; Taddy, M.; and Evans, J. A. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84(5):905–949.

Krause, A. E., and North, A. C. 2018. 'tis the season: Music-playlist preferences for the seasons. *Psychology of Aesthetics, Creativity, and the Arts* 12(1):89.

Mehr, S. A.; Singh, M.; Knox, D.; Ketter, D. M.; Pickens-Jones, D.; Atwood, S.; Lucas, C.; Jacoby, N.; Egner, A. A.; Hopkins, E. J.; et al. 2019. Universality and diversity in human song. *Science* 366(6468).

North, A. C., and Davidson, J. W. 2013. Musical taste, employment, education, and global region. *Scandinavian journal of psychology* 54(5):432–441.

North, A. C., and Hargreaves, D. J. 2007. Lifestyle correlates of musical preference: 3. travel, money, education, employment and health. *Psychology of music* 35(3):473–497.

Omitted. 2020. Community embeddings reveal large-scale cultural organization of online platforms. unpublished.

Park, M.; Thom, J.; Mennicken, S.; Cramer, H.; and Macy, M. 2019. Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour* 3(3):230–236.

Tarrant, M.; North, A. C.; and Hargreaves, D. J. 2002. Youth identity and music. *Musical identities* 13:134–150.

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

Way, S. F.; Gil, S.; Anderson, I.; and Clauset, A. 2019. Environmental changes and the dynamics of musical identity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 527–536.