Understanding Musical Diversity via Online Social Media

Minsu Park[*1], **Ingmar Weber**[2], **Mor Naaman**[1], **Sarah Vieweg**[2]
[1]Jacobs Institute, Cornell Tech
[2]Qatar Computing Research Institute (QCRI)
{minsu, mor}@jacobs.cornell.edu
{iweber, svieweg}@qf.org.qa

Abstract

Musicologists and sociologists have long been interested in patterns of music consumption and their relation to socioeconomic status. In particular, the Omnivore Thesis examines the relationship between these variables and the diversity of music a person consumes. Using data from social media users of Last.fm and Twitter, we design and evaluate a measure that reasonably captures diversity of musical tastes. We use that measure to explore associations between musical diversity and variables that capture socioeconomic status, demographics, and personal traits such as openness and degree of interest in music (into-ness). Our musical diversity measure can provide a useful means for studies of musical preferences and consumption. Also, our study of the Omnivore Thesis provides insights that extend previous survey and interview-based studies.

Introduction

The cultural and social significance of music is universal; music is found in every known human culture, and plays a role in rituals, wars, ceremonies, work, and everyday life (Wallin, Merker, and Brown 2001). Tia DeNora (DeNora 2000) noted that "Music is not merely a meaningful or communicative medium. It does much more than convey signification through non-verbal means. At the level of daily life, music has power. It is implicated in every dimension of social agency." As social media become more ingrained in our lives, it follows that connections between social media use, and habits and norms regarding music consumption, will occur. In this paper, we present an empirical analysis of social media data as they relate to and reveal details of users' musical tastes.

A person's musical consumption can reveal a lot about their personality, preferences, and sense of self. One can have limited tastes; they may listen to a single genre like pop or rap, and not diverge into other genres. On the other hand, another individual may be eclectic in their musical choices and have a playlist filled with jazz, hip-hop, indie rock, classical, and so forth. We often think of such differences as a matter of individual choice and expression; however, to a great degree, it is hypothesized and tested that the diversity of musical tastes can be explained by external factors. For example, previous research has identified a relationship between musical tastes and social factors, and produced the *cultural omnivore thesis*. This thesis describes "a shift in the orientation of high-status individuals toward an inclusive range of musical preferences that traverses the traditional boundaries between *highbrow*, *middlebrow*, and *lowbrow* genres (Peterson 1992; 1997; 2005)." However, symbolic boundaries between musical genres have been eroding (Goldberg 2011) in recent years, which provides an opportunity to rethink the high-to-lowbrow cultural categories in relation to musical diversity. This can lead to a better understanding of the impact of social conditioning on diverse musical tastes, and by proxy, a better understanding of the connection between socioeconomic status, demographics, and the diversity of musical preferences.

To date, the social computing community has examined online listening activity as source of information and recommendations for music (Bu et al. 2010; Zheleva et al. 2010; Farrahi et al. 2014; Turnbull et al. 2014). However, computational tools and online outlets such as social media can make further contributions toward understanding human behavior related to musical consumption and help to elaborate user-centric music retrieval systems by analyzing personal characteristics. We focus on exploring a new means of measuring the diversity of individual musical tastes by using data collected from social media, and examine the relationship between musical diversity and various individual factors including socioeconomic and demographic information, as well as social and individual information that can be collected from social media.

Through a multi-platform analysis of a dataset of U.S. Last.fm[1] users and their corresponding Twitter accounts, we examine music consumption together with demographics (e.g., age and gender) and other descriptive variables for a community music fans who have an online presence. Using

[*]The majority of this work was done while Minsu Park was a research intern at Qatar Computing Research Institute.

[1]Last.fm is a music recommendation service. The site builds a detailed profile of each user's musical consumption by recording details of the tracks the user listens to, either from Internet radio stations, or the user's computer or many portable music devices. It also offers some social networking features such as recommending and playing artists to Last.fm friends (Wikipedia 2015).

Twitter-derived information for these users, we inferred their socioeconomic information (e.g., income, education level, and area of their residence) as well as other social and personal variables (e.g., how diverse their friends and interests are, and how 'open' and 'into music' they are). We then defined a measure for musical diversity by applying the notion of shared understanding as socially perceived distances between genres. We suggest that designing a diversity measure can provide a useful means for studies in recommendation systems. Moving from designing a measure to analysis of associations between diversity and individual factors, we suggest this type of analysis can provide meaningful insights that are complementary to those provided by previous survey and interview-based studies regarding the musical omnivore thesis. Our main contributions therefore are as follows:

- We propose and validate a novel diversity measure that borrows the concept of Rao-Stirling diversity for music consumption. While recent studies (Hurley and Zhang 2011; Farrahi et al. 2014) define diversity (as it relates to music consumption) as the total number of unique genres associated with all artists listened to, we go into more detail, and define diversity as a multidimensional property that has three main attributes: *variety* (the number of unique genres one listened to), *balance* (the listening frequency distribution across these genres), and *disparity* (the degree of distance between musical categories).
- We investigate the relation between musical diversity and various other variables including socioeconomic factors. In particular, we find that followers of high-profile news media are more likely to have diverse musical tastes. We also consistently find a weak, but robust trend for people who are more 'into' music to have less diverse tastes. Along with these findings, our results also show that demographic factors such as age and gender are associated with musical diversity rather than conventional socioeconomic status such as income and education level.

We begin by reviewing the primary key research around the diversity of musical tastes, and then identify possible challenges for developing better measures of diversity.

Related Literature

Disciplines such as sociology and social computing addressed the notion of *cultural omnivorism* and the importance of understanding the musical diversity. Given the wealth of related work on these topics, our review focuses on what could be tested by complementing the limitations of previous studies through social media data and how we can design a meaningful measure for the diversity of musical tastes.

Changing Status of the Omnivore Thesis

Since the publication of Bourdieu's seminal work *Distinction* (Bourdieu 1984), in which he explains the notion of cultural capital and exhibits how access to education, knowledge of the arts, and familiarity with other highly regarded aspects of western culture lead to a 'highbrow' status, copious research has investigated the relationship between socioeconomic position and musical tastes (Coulangeon and Lemel 2007). The majority of the current studies on the *omnivore thesis* in relation to musical tastes, proposed by Richard Peterson (Peterson 1992) show that people with a higher socioeconomic status have broader (omnivorous) musical tastes than those with a lower socioeconomic status who have limited (univorous) musical preferences in lowbrow music. There are generally two definitions of omnivorousness, referred to as the *volume* and the *compositional* definitions (Warde, Wright, and Gayo-Cal 2007). The first refers to higher socioeconomic status people favoring more musical genres than those of lower socioeconomic status. The second refers to the situation that people with higher socioeconomic status tend to have more eclectic tastes across the spectrum of high-to-lowbrow music than people with lower socioeconomic status.

More recently, however, Peterson (Peterson 2005) conducted comparative research and noted that "despite the attention paid to the concept by numerous scholars, the subtypes of omnivorousness suggested by them were diverse and fall into no recurrent patterns due to changes in the socio-cultural world." Indeed, though there is a little disagreement that the contemporary era has witnessed shifts in the ways cultural preferences and practices are mapped onto social locations, the extent to which this implies changes in the functioning of cultural capital remains unclear (Rimmer 2012). In addition, Peterson (Peterson 2005) raised a question regarding the traditional measurement of omnivorousness, and recent qualitative studies identified a number of limitations in conventional survey-based studies (Warde, Wright, and Gayo-Cal 2007; Rimmer 2012): First, the simple or compositional volume of genres preferred by an individual is insufficient to show the full picture of one's form of engagement and social status since different conceptual frameworks may provide different understandings. Second, there is a tendency to discriminate genres within preferred genres (i.e., even though one answers 'rock' as a preferred genre, it does not mean that one likes *all* kinds of rock; therefore, it is possible that someone who likes a Heavy Metal, a subgenre of rock, says "I like rock," and someone who likes the same subgenre says "I don't like rock"). This inability to discriminate genres, or lack of knowledge regarding how to best express what genres one prefers, can create confusion (Rentfrow and Gosling 2003). This gap may bring inconsistency in the preference scoring across survey participants. Finally, the high-to-lowbrow scheme should be reconsidered in contemporary social contexts as Peterson (2005) argues that there is no consensus. In addition, a lot of research has used inconsistent levels of genres, e.g., a questionnaire of preferences for opera, jazz, rock, and heavy metal may be used in these types of surveys, even though heavy metal is often considered a subgenre of rock.

We believe online social media data can help rectify some of these limitations and provide a unique and useful perspective on the musical omnivore thesis: data collected from social media sites can provide a unique capacity to (i) reduce the inconsistency of preference scoring (which may differ

across people due to their inability to discriminate) by systematically classifying the genres consumed by users, (ii) explore a different level of relationship between social status and musical tastes by accessing the subgenres of choice among users, which are more fine-grained than higher-level genres, and (iii) analyze data on a consistent level of genre-hierarchy. Further, social media data can provide users with open-ended spaces (Lewis, Gonzalez, and Kaufman 2012) in which to list their favorite music, concert attendance, and direct/indirect musical information sources, which offers an unprecedented opportunity to examine how tastes are associated with various individual factors. Up to now, the majority of research on musical tastes has relied on closed-ended surveys typically measuring preferences in terms of genres, and our aim is to contribute a new way to look at the relationship between musical preference and various social and individual factors.

Technology and Music Listening Practice

Exploring musical diversity is an interesting challenge in social computing, as well as music information retrieval (MIR); it also has many applications in real-life scenarios. In MIR, some researchers have explored to achieve the optimal balance between the two objectives on recommendation, similarity and diversity, because it has been recognized that being accurate with similarity metric alone is not enough to judge the effectiveness of a recommendation system (McNee, Riedl, and Konstan 2006; Chen, Wu, and He 2013). In addition, recent studies (Chen, Wu, and He 2013; Farrahi et al. 2014) suggest that one's personality might have a role in the formation and maintenance of music preferences, and diversity of musical tastes could serve as a proxy of the level of openness of one's personality. These studies show that looking at musical diversity as an indicator of openness can have an impact on the performance of a collaborative filtering recommender system. In social computing, diversity has been considered in studying phenomena such as peer influence and music consuming mechanism. Some of this research confirms that informational influence is the key underlying mechanism of music listening practices (Yang, Wang, and Mourali 2014) and systematic recommendations affect users' choices of music tracks and listening behaviors (Buldú et al. 2007).

Research Questions

We believe associations between musical categories (e.g., genre-to-genre and subgenre-to-subgenre) can be reasonably derived from the perception of crowds by analyzing their musical consumption, and these distances may help design better measures of musical diversity. The existing measures, *volume* or *entropy*, are different from diversity, and thus cannot accurately capture its essence. Volume, which is defined as the number of musical categories one listens to, does not consider whether a person listens with balance. A 99%-1% split between two genres would be treated the same as a 50%-50% split. Entropy, on the other hand, takes the distribution into account, so a more skewed distribution would be considered less balanced. However, entropy does not look at the similarities of the musical categories and implicitly assumes all categories to be equidistant to each other (e.g., listening to three different styles of metal music would be the same as listening to classical music, death metal, and salsa). People, however, do consider certain types of music as similar or dissimilar (Mörchen et al. 2005). To define and to quantify this notion of similarity we use *co-consumption* behavior. For example, if both rap and hip-hop are consumed by many people we assume that these two genres are similar. Having musical consumption data for a large user set can reveal the distance between musical categories.

The challenges and opportunities in studying musical diversity lead us to introduce two research questions that guide the remainder of this paper:

RQ1 *Can a novel diversity measure using variety, balance, and distance between musical categories capture the diversity of musical tastes better than existing methods?*

RQ2 *What variables are associated with diversity in music consumption? Is socioeconomic status a factor or are other factors also associated?*

Method

The literature referenced in the previous section points to three major dimensions of explanatory variables: socioeconomic status, demographic information, and 'openness' (degree of appreciation for novelty and variety of experience). With these dimensions and the additional dimension of 'into-ness' (degree of self-disclosed interest in music) as a guide, we identified 15 variables. We inferred socioeconomic status including income, education level, ethnic diversity of area of residence, and urbanness of area of residence by using geocoded tweets. Into-ness (i.e., degree of music-related topics of interest in Twitter) and openness including number of friends, timezone diversity of friends, and interest diversity was inferred by using tweets, profile descriptions, and friendship information in Twitter. We directly downloaded demographic information (e.g., gender and age) and other types of into-ness (e.g., number of event attendance in the past, number of loved tracks, period after registration, and number of friends in Last.fm) through the Last.fm API.

Initial Data Collection

To identify and obtain a sample of Last.fm users in the U.S. who share gender, age, and Twitter user names in their Last.fm profiles, we used the Google Custom Search API and the Bing Search API. We created a custom query containing parameters that returned only Last.fm user pages which contained this particular information. To augment the sample size, we collected U.S. Twitter users who share their Last.fm accounts in their Twitter profiles by using the 'Search Bio' feature in Followerwonk[2]. This allowed us to obtain 23,294 unique users. Then, we collected all publicly available tweets from that user population. During this process 4,392 unique users were screened out since some of them did not allow public access to their tweets or had removed their accounts in the meantime. This left us with

[2] https://followerwonk.com

18,902 unique users. To infer socioeconomic status by using geocodes in tweets, we limited our remaining sample to those users who posted at least ten tweets with geocodes, which resulted in 3,548 users. Along with Twitter data, we collected Last.fm data including 'Top artists' list (i.e., the 50 musicians a user listened to the most; listening frequency for each artist is included) as well as demographic and some into-ness information directly through the Last.fm API.

Socioeconomic Status

We used home location derived from Twitter as an index to approximate socioeconomic data, and news interests, expressed via Twitter's following network, as another proxy for socioeconomic status.

A user's home location can be a marker of their socioeconomic status. In particular, the socioeconomic status of social media users can be estimated by extracting the users' hometown ZIP codes and matching that to the median ZIP code household income according to the Census Bureau (Lewis, Gonzalez, and Kaufman 2012). In addition, using the inferred home location we can check whether a user lives in an urban or rural area (Hecht and Stephens 2014).

To obtain the home location for a user, we followed a procedure that involved three different methods of identifying a user's possible home ZIP code. We first reverse-geocoded all the latitude and longitude tags for the user into the ZIP codes, using the Nominatim API[3]. We also extracted Federal Information Processing Standard (FIPS) codes, which represent specific regions in counties, using the Coordinates to Political Areas API in Data Science Toolkit[4]. Using the ZIP code data for the user, we inferred a probable home location of a user when we found an intersection between the sets of potential ZIP codes for the user computed by three different methods, the *plurality* and *n-days* methods summarized in (Hecht and Stephens 2014) and the *plurality with time limitation* described in (Castelli et al. 2009).

The plurality approach (Hecht and Stephens 2014) assumes that the single region in which a user was the most active is the user's home location. Using this approach, we find the user's mode ZIP code(s) from which tweets were most frequently posted. The *plurality with time limitation* method is based on the finding in (Castelli et al. 2009), that people are most likely home between 10pm – 6am. Using these parameters, we identify the user's mode ZIP codes(s) from which tweets were most frequently posted during that time period. Since the plurality approaches may not be appropriate for users who travel frequently, the final method we used identified the ZIP code(s) in which a user posted over a period of at least 10 days, considering them 'local' to that area if they did.

We selected a single home ZIP code (and FIPS code) for each user by intersecting the ZIP code sets resulting from the three methods mentioned above. The final set of users with non-empty intersection had 1,306 users (there were 3,451, 3,258, and 1,822 users with non-empty sets for each of plurality, plurality with time constraint, and n-days methods respectively). All other users for which we could not robustly estimate a location were removed from the data.

Finally, to extract socioeconomic data, we used each ZIP code to query the 2010 US Census data to determine income, education level, and ethnic diversity in the area. We matched each FIPS code to NCHS data for urban–city classification of the area which places every U.S. county on a discrete scale from 1 (a large central metro area) to 6 (a sparse rural area). For each user we thus have values for median household income, percentage of bachelor degrees, proportion of white people[5], and urbanness: these are our socioeconomic proxy measures. This process resulted in 1,306 users for whom we have self-declared gender and age, as well as inferred income, education level, and characteristics of the area of residence[6].

In addition to location-derived socioeconomic data, we used news interest as a proxy for socioeconomic variables. According to Pew Research (Pew Research Center 2012), regular news audiences often are more formally educated and have higher household incomes. In particular, readers of The New Yorker and The Economist news media tend to be highly educated and high earners (Pew Research Center 2012). We therefore created a variable that indicates whether each of our users follows The New Yorker (@NewYorker) or The Economist (@TheEconomist) on Twitter.

Genre and Subgenre Information Collection

For each user, we extracted the categories of music they listen to at both genre and subgenre ('style') levels. For each user we retrieved the top 50 artists the user listened to via the Last.fm API. We collected genre and subgenre information for each artist using the API for *Allmusic*[7], a well-known music database (DB). Unlike other music content databases, Allmusic's metadata is professionally edited and thus is likely to be more consistent when assigning genres or subgenres to artists. Many high-profile music sources like iTunes and Spotify currently use Allmusic to handle relevant artist information.

We matched each artist name collected from Last.fm to an artist entry on the Allmusic DB only if the result exactly matched the queried artist name. When multiple musicians with the same name were matched, we used the Allmusic

[3] http://www.nominatim.org

[4] http://www.datasciencetoolkit.org

[5] We tested relation between white ratio and 'racial and ethic diversity' by using the Ethnic/Racial Diversity Index which defines racial and ethnic diversity as $1 - \sum_{r \in G} P(r)^2$ where $P(r)$ is proportion of a race population r and G is represented race groups (in our case: white, black, Native American, Asian, Hispanic, Pacific Islander, two or more races, and other races by following ethnicity distribution in the 2010 Census). A higher index number denotes more diversity. However, there is confusion among the general population about the designation of the Hispanic identity since 'Hispanic' in the census refers to any 'race,' both black and white. So, we decided to use the simple metric, 1 – white ratio, as 'Racial Diversity' since it is clearer. The Pearson correlation between the white ratio and ethnic diversity was 0.667 ($p < 0.001$).

[6] We ignored 97 users due to various ZIP code issues, such as ZIP code that were invalid, not available from the census data, or too small to have socioeconomic statistics.

[7] http://www.allmusic.com/

engine's relevance ranking which is based on usage data and editorial weighting. We manually validated the Allmusic ranking for a random selection of 100 artists that had multiple entries. We examined the Last.fm page for the artist (as linked from the user's Top 50 list, i.e. uniquely identified) and the Allmusic page for the top-ranked artist by the same name as retrieved by the API. We found that the top-ranked artist matches with the Last.fm artist for all cases in this sample of 100.

A single artist could be classified into multiple genres and subgenres, in which case we distributed the artist's 'weight' equally between the respective genres or subgenres. During this data processing, we dropped 292 users who did not have full set of 50 artists that were classified by Allmusic and listened to more than 100 times by the user. As a result, data for 1,014 users were analyzed. There were 8,490 unique artists among the Top 50 artists of 1,014 users, and 987 artists among the unique artists were matched with more than one exact name in Allmusic DB (e.g., Nirvana and Spoon).

Measuring Diversity

We calculated the diversity of music consumption for each user using both genre- and subgenre-level data derived from their Last.fm activity. We previously argued that in order to explore diversity, we need to investigate multiple factors, namely: the number of genres listened to (variety), the distribution of playing frequency among genres (balance), and, crucially, how related these different genres are (measured via some distance or similarity). These assumptions align well with the concept of Rao-Stirling diversity (Stirling 1998; 2007; Porter and Rafols 2009; Leydesdorff and Rafols 2011).

To operationalize the concept of diversity, following Rao-Stirling, we computed the diversity of musical tastes of a user u as $\sum_{i,j \in N} p_{u,i} \times p_{u,j} \times d(i,j)$. In this formulation, $p_{u,i}$ is the fraction of user u's preference for genre i (we performed separate and equivalent calculations for genres and subgenre information; the description here focuses on genre information). To compute $d(i,j)$, we computed the pairwise co-consumption between musical categories as a proxy of closeness. Using an $M \times N$ genre proportion matrix of $p_{u,i}$ values (for each row u, $\sum_i p_{u,i} = 1$), we computed every possible pair of genre-to-genre cosine distances between the matrix columns, representing closeness between genres. The distance $d(i,j)$ is the cosine distance, i.e., 1 – cosine similarity, between the genres. As mentioned above, we repeated the same process with subgenre information. For illustration, the resulting distances for *genres*, embedded in two dimensions using multidimensional scaling (Kruskal and Wish 1978) (MDS), are shown in Figure 1[8].

This approach to computing diversity of music consumption has a number of useful qualities. A user who equally (balance) consumes many types of music (variety) that are pairwise highly dissimilar (distance) will have a large diversity score, whereas a user disproportionally consuming a few pairwise similar types of music will has a low diversity score. We evaluate this approach and its robustness below.

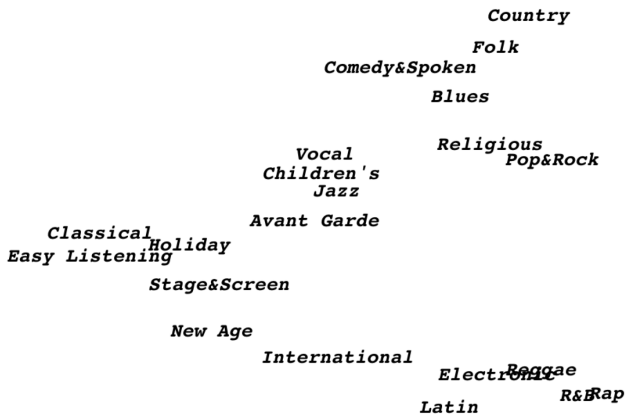

Figure 1: Multidimensional scaling for distance between genres

Into-ness and Openness

For each user, we calculated several variables that capture openness (preference for novelty and variety) and into-ness (degree of interest in music) using Twitter and Last.fm data. To help inferring into-ness and openness regarding each user's interests, we first inferred the user's general interests by using a method proposed in (Bhattacharya et al. 2014). For a given Twitter user *u* (whose interests are to be inferred), the method first checks which other users *u* is following, i.e., users from whom *u* is interested in receiving information. It then identifies the topics of expertise of those users (whom *u* is following) to infer *u*'s interests, i.e., the topics on which *u* is interested in receiving information. Expertise is defined by the users bio or tweets via the Lists feature in Twitter (Ghosh et al. 2012).

Using the interest topics for each user, we computed openness and into-ness measures. As a proxy of openness, we computed the diversity of the user's interests using the same method we calculated music consumption diversity above. In this case, for example, similarity of interests can be derived from the cosine distance between interest in a matrix that captures users' interest breakdown. As other measures of openness, we counted for each user in our dataset the number of people they are following on Twitter and also the number of unique timezone in 100 randomly sampled people from whom they are following. We collected these openness variables inspired by (Schrammel, Köffel, and Tscheligi 2009; Quercia et al. 2011)[9].

As a proxy of music into-ness, we used the proportion of

[8]Interestingly, highbrow and middlebrow genres (e.g., classical, easy listening, and jazz) are close to each other rather than being close to lowbrow genres (e.g., pop&rock, folk, country, rap) even though we used an inductive approach to identify the distance between musical categories rather than assuming that musical tastes are shaped by certain schemes.

[9]We did not consider lexical features of tweets as variables since previous efforts (Golbeck et al. 2011; Qiu et al. 2012; Schwartz et al. 2013) showed a disagreement regarding predicting features for openness.

Socioeconomic Variables	Distribution	Max
Income		192,250
Education		100
Racial Diversity		0.98
High-profile News Reader	High-profile News Media Followers: 12.4% Non-followers: 87.6%	
Urbanness		1–6 (Scale)
Demographic Variables		
Age		52
Gender	Female: 30.7% Male: 69.3%	
Into-ness Variables		
Musical Event Attendance		1,504
# of Loved Tracks		12,619
Days from Registration		4,305
# of Last.fm Friends		2,036
Interest in Music		2,456
Openness Variables		
# of Twitter Friends		10,954
Timezone Diversity of Friends		31
Interest Diversity		0.76
Diversity		
Diversity on Genre		0.67
Diversity on Subgenre		0.80

Table 1: Fifteen variables used to explain the measured musical diversity scores and genre- and subgenre-level of diversity scores. The distributions accompanying each variable begin at zero and end at the adjacent maximum. Many variables are not normally distributed.

music-related interests (any interest topic that included the term 'music') among the entire set of user interests along with other types of into-ness that were directly collected via the Last.fm API: number of event attendance in the past, number of loved tracks, period after Last.fm registration, and number of friends in Last.fm.

Table 1 presents 15 variables we identified and diversity on genre and subgenre along with their distributions.

Data Validation and Preparation

Given that some of our variables were indirectly derived from social media data, we performed validation tests for our key variables.

Reverse Geocoding To validate our geocoding framework, we matched the inferred ZIP code to the self-reported home location of the user on their Twitter profile. Out of 100 randomly sampled users, eight users did not disclose their location on their Twitter profile or did not properly disclose their location like "not in a cornfield but. . . close" and "up in the air." Among the rest of them (92% of users), only eight users' locations did not overlap with the inferred zip code location. In other words, more than 90% of inferred locations were well-matched to the self-reported home locations at town/city/state levels.

Note that it is unusual to have as much as 92% of users with a valid location field (Hecht et al. 2011). Our dataset, though, includes Twitter users who are also heavy users of the geo-tagged tweets feature; it is conceivable that the same group more readily exposes location in their profile data.

Socioeconomic Status Even if we get the user's location right, the derivation of their socioeconomic information may be wrong as the user may not be *representative* of where they live. For example, it is possible that people who use both Twitter and Last.fm have similar socioeconomic status, regardless of what sort of neighborhood they live in. However, if the inferred socioeconomic information are correct, they should correlate with our other proxy for socioeconomic status: following the New Yorker or Economist. We thus validate our socioeconomic measures by examining whether our inferred income and education level are associated with following the New Yorker (@NewYorker) or The Economist (@TheEconomist) Twitter accounts. Indeed, compared to other users, New Yorker and Economist followers had higher status for all inferred income and education values, including adjusted gross income (AGI), household income, and level of post-secondary degree (both bachelor's and graduate). These differences were statistically significant as determined by a one-way ANOVA (New Yorker followers AGI: $p < 0.01$; median household income: $p < 0.05$; bachelor degree: $p < 0.001$; graduate degree: $p < 0.001$; Economist followers AGI: $p < 0.001$; median household income: $p < 0.05$; bachelor degree: $p < 0.001$; graduate degree: $p < 0.001$).

Data Imputation and Standardization In our final dataset, 189 out of 1,014 subjects had missing values in one or more variables. According to (Hair et al. 2006), if the missing data level is under 10% in each variable, any imputation method can be used to augment the missing values. We used multiple imputation methods in our dataset: we applied Bayesian linear regression for continuous variables, and linear discriminant analysis for factor variables. We also standardized all the variables for the final analysis.

Results

Our primary purposes for this study were (i) to design a measure that reasonably captures the notion of 'diversity of musical tastes' and (ii) to explore associations between musical diversity and various individual factors regarding dimensions of socioeconomic status, demographics, and personal traits including openness and into-ness in music.

Diversity Measure

To answer RQ1, we estimated the reliability of our diversity measure. We asked three independent annotators to assign a

diversity level to the musical consumption of 25 randomly chosen users. The annotators ranged in their music knowledge; we had an expert (musicologist), a music fan, and a causal listener. We provided the annotators two sets of tables of genre- and subgenre-based listening proportion of the 25 users. We asked the annotators to carefully examine each user's listening pattern and apply a 6-point diversity Likert scale where '5' meant very diverse musical taste, '1' meant very low diversity, and '0' meant no diversity at all (it is possible that a user listened only to one genre). We did not provide the annotators with any other information or instructions (such as "consider the relationship between genres") as we wanted to know their natural impressions and interpretations of diversity based on their own experiences. Fleiss's Kappa and average pairwise Cohen's Kappa were used to assess the inter-rater reliability for the evaluation. For genre-level the Fleiss Kappa score was 0.411 ($p < 0.001$) indicating moderate agreement, and the Cohen's Kappa score was 0.819 ($p < 0.001$) indicating almost perfect agreement. For subgenre-level, the respective scores were 0.011 ($p > 0.1$) indicating slight agreement and 0.415 ($p < 0.05$) indicating moderate agreement. We averaged the rater responses for each user and used that below as the raters' diversity score.

To evaluate our diversity measure, we calculated the Pearson correlation between the raters' average score and our computed diversity score. For genre-level diversity, the correlation between our measure and the raters' diversity was 0.94 ($p < 0.001$). For the subgenre-level diversity, the average correlation was 0.87 ($p < 0.05$). Interestingly, looking at correlations between individual raters' and our diversity score, the expert annotator had the highest correlation with our diversity score in both settings.

Other commonly used diversity measures were more sensitive to the level of analysis. We correlated the raters diversity scores with the diversity scores computed by Shannon entropy and by the count of musical categories a user listened to ('volume'). In the genre-level analysis, both the entropy and volume methods showed significant correlation with the raters. The Pearson correlation between the raters' average scores and the entropy values was 0.95 ($p < 0.001$). The average correlation between raters and the volume measure was 0.86 ($p < 0.001$). However, in subgenre-level analysis we found more notable differences between the raters' and our diversity scores. The Pearson correlations between the entropy and the rater scores was 0.79 ($p < 0.05$). With volume, the average correlation was 0.46 ($p < 0.05$).

This result initially indicates that our diversity measure is promising as it captures human rater evaluations of diversity more robustly than traditional measures—it is less dependent on changes in categorical hierarchies. The distance between musical categories can be an important factor for understanding musical diversity, especially in highly complex musical classifications.

Correlates of Musical Diversity

To address RQ2, we used multiple regression analyses to examine factors associated with the diversity of musical consumption. We examined socioeconomic status variables as well as demographics, openness, and into-ness measures.

Table 2: Multiple regression coefficients of individual factors on the musical diversity of genre and subgenre

	Dependent variable:	
	Genre	Subgenre
	(1)	(2)
Income	-0.047	0.007
	(0.037)	(0.037)
Education	0.027	-0.020
	(0.039)	(0.039)
Racial Diversity	0.108^{**}	0.089^{*}
	(0.036)	(0.036)
Urbanness	-0.040	0.052
	(0.035)	(0.035)
High-profile News Reader	0.366^{***}	0.301^{**}
	(0.095)	(0.096)
Age	0.121^{***}	0.161^{***}
	(0.033)	(0.033)
Gender (Male)	$0.111^{\cdot}$	0.153^{*}
	(0.067)	(0.067)
Musical Event Attendance	-0.145^{***}	-0.042
	(0.034)	(0.034)
# of Loved Tracks	0.079^{*}	0.089^{**}
	(0.033)	(0.033)
Days from Registration	-0.102^{**}	-0.029
	(0.033)	(0.033)
# of Last.fm Friends	0.023	-0.081^{*}
	(0.036)	(0.036)
Interest in Music	-0.143^{***}	-0.113^{***}
	(0.034)	(0.034)
# of Twitter Friends	0.085^{*}	0.050
	(0.033)	(0.034)
Timezone Diversity of Friends	0.026	0.074^{*}
	(0.032)	(0.032)
Interest Diversity	-0.027	-0.017
	(0.032)	(0.031)
Constant	-0.123^{*}	-0.143^{*}
	(0.057)	(0.057)
Observations	1,014	1,014
R^2	0.101	0.087
Adjusted R^2	0.088	0.073
Residual Std. Error (df = 998)	0.955	0.963
F Statistic (df = 15; 998)	7.487^{***}	6.322^{***}
Note:	$^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001	

Table 2 presents the standardized coefficients of the explanatory variables[10]. The model (1) in Table 2 estimates

[10]All variance inflation factors are below 1.64 ($\mu = 1.28$ and $\sigma = 0.16$); Pearson correlation between genre and subgenre diversities is 0.68 ($p < 0.001$).

the effects of socioeconomic, demographic, and other individual variables on the diversity of musical consumptions on genres. Among the 'socioeconomic status' variables, *High-profile News Reader* variable had a high coefficient due to users who follow The Economist or The New Yorker having higher musical diversity than those who do not (one-way ANOVA confirmed the significance; $p < 0.001$). Even though we exclude this variable to check whether income and education variables are associated with diversity of music consumption, we could not find any change regarding significance level and direction of correlation. The readers of high-profile news reports may have indirect or subtle difference in terms of socioeconomic status.

Racial Diversity positively associates with the diversity of music consumption. This may imply that people in our sample who live in more ethnically diverse area are more likely to have higher musical diversity. By considering the relationship between white ratio and ethnic diversity, this result might be related to the effect of residential segregation. Both of *Age* and *Gender* in the 'demographic' variables have positive effect on diversity: being older or male is more likely to have more diverse musical tastes.

Among variables about 'into-ness,' *Musical Event Attendance* and *Days from Registration* appear to be negatively associated with diversity, whereas *Number of Last.fm Friends* does not show a significant relationship and *Number of Loved Tracks* appears to positively associated with diversity. *Number of Twitter Friends* as a 'openness' variable appears to be positively associated with diversity while *Timezone Diversity of Friends* and *Interest Diversity* shows no effect. On this basis, one could speculate that few variables within the same set of variables correlate with musical diversity in different directions. We discuss these trends below.

Model (2) in Table 2 estimates the effects of the same variables on the diversity of musical consumption of subgenres; it shows very similar trends with model (1). However, *Gender* is more significantly associated with diversity. Among the 'into-ness' variables, *Number of Last.fm Friends* is significantly associated with diversity rather than *Days from Registration*. But, the general trends of 'into-ness' are in common. Among the 'openness' variables *Timezone Diversity of Friends* is significantly associated with diversity rather than *Number of Twitter Friends* while the general trends of the 'openness' are in common.

Discussion

Our results provide initial evidence for the value of our 'music diversity measure' which aims to balance three qualities: variety, balance, and distance. Our diversity measure has shown to be more robust than other conventional measures such as volume and entropy.

Differences between Pearson correlation coefficients at the genre- and subgenre-levels computed by our measure, as well as the average rates assigned by independent coders on a 6-point Likert scale, were not significantly different. For the other measures of diversity, when moving between genre and subgenre levels, the average correlation coefficients dropped more steeply, especially the volume measure. Musical diversity can be computed by simple methods, but it may underestimate or overestimate diversity depending on the complexity of musical categories and the disparity between musical categories that people perceive. Our results show that volume and entropy might not be the best solution for computing the musical diversity of people on a highly complex map of musical categories such as subgenres.

We only considered the genre and subgenre categories, but new methods for music classification may result in categories that are even more complex, making a robust diversity measure even more important. For example, research efforts have developed novel methods for music classification using various data sources such as audio features and song metadata (Henaff et al. 2011; Foucard et al. 2013).

In addition, diversity of music consumption was correlated with interest in high-profile news media; users who follow high-profile news media are much more likely to have a higher level of musical diversity. When we think about whether one consumes high-profile news media, it is not necessarily a variable that is as straightforward as income or education level. To understand news reports, readers need more than a basic grasp of word order and word meaning; a particular 'knowledge of the world' is also necessary. Van Dijk (Van Dijk 1996) explains this when he writes: "Readers of a news report first of all need to understand its words, sentences, or the structural properties. This does not only mean they must know the language and its grammar and lexicon, possibly including rather technical words such as those of modern politics, management, science, or the professions. Users of the media need to know something about the specific organization and functions of news reports in the press, including the functions of headlines, leads, background information, or quotations. Besides such grammatical and textual knowledge, media users need vast amounts of properly organized knowledge of the world." Van Dijk's point alludes to the possibility that if one has access to particular understandings of 'the world,' then they are better equipped to seek out and benefit from high profile news sources. If this is the case, then we can begin to think about level of music diversity as a potential variable vis-à-vis knowledge.

Our results also confirm a number of previous findings about demographic variables associated with the diversity of music consumption. They show that male users are more likely to have diverse musical tastes, which confirms prior research showing that males tend to consider mainstream music as *unhip* while females consider it in another way of saying *popular* music (Christenson and Peterson 1988); such perceptions might affect musical consumption. Males are also more likely to prefer more unique styles of music than females (Rawlings and Ciancarelli 1997). In addition, people who are older in our sample are more likely to have diverse musical tastes. This result closely echos the analyses of (Warde, Wright, and Gayo-Cal 2007): young people may identify strongly with one or only a few genres and styles of music, which reveals the significance of their representational dimensions.

A potentially surprising finding is that people who attended more musical events are less likely to have diverse listening habits. (Warde, Wright, and Gayo-Cal 2007) also

argues that there is a tendency, or an openness, towards unfamiliar musical forms and evidence of relatively diverse tastes in people who are limited in how they can engage in musical activities. The development of a broad palette of musical tastes was not valued by people for whom music is more accessible. We note that urban dwellers may have better access to musical activities, but we could not find a significant association with urbanness in our results.

Users with diverse patterns of music consumption are less likely to follow music-related accounts on Twitter. This finding can be due to a different set of music-related interests between diverse listeners—who care more about the music itself—and more casual music fans who may care more about the celebrity factor. If this were shown to be true, we may refer to it as the Justin Bieber effect (no offense to his fans should they be reading this paper).

Final Remarks

In this paper, we have designed a reasonable measure that quantifies the diversity of musical tastes. In addition, we provide an analysis of diversity as it relates to the cultural omnivore thesis. Based on well-known individual factors which relate diversity of musical preferences across various theoretical work and empirical studies, we identified key factors for designing a diversity measure, and located individual-level variables for exploring correlations of musical diversity.

We acknowledge that the manner in which we inferred the socioeconomic status variable could produce significant inaccuracies. For example, users' home locations were inferred in ZIP code resolution and using geocoded Twitter data. These methods are prone to error. Other methods for collecting more direct or fine-grained location data, or maybe even a direct collection of socioeconomic variables, might give us a better opportunity to study this correlation with music consumption. Second, our user population and the music they listen to are both potentially highly biased. Our population is comprised of users who make an explicit connection between their Twitter and Last.fm accounts, which may indicate search biases on our behalf. In addition, the tracks and artists displayed for each user are based on their public listening behavior, which may or may not be reflective of their overall listening habits. Finally, we could see rating differences among coders due to knowledge differences. The measurement validations can be improved by better systematic investigations using more listening history samples and annotators with different levels of knowledge background. At the same time, it would be interesting to see if the way of rating changes when the music listeners themselves are asked about their diversity.

Future research along this vein will provide a richer and more complex picture of musical preferences. This picture will in turn contribute to a greater understanding of the changing face of the cultural omnivore, as it manifests through analyses of social media data, and also contribute to a empirical recommendation system aiming to provide contents based on tastes and aesthetics preferences.

References

Bhattacharya, P.; Zafar, M. B.; Ganguly, N.; Ghosh, S.; and Gummadi, K. P. 2014. Inferring user interests in the twitter social network. In *ACM Conference on Recommender systems (RecSys)*, 357–360.

Bourdieu, P. 1984. *Distinction: A social critique of the judgement of taste*. Harvard University Press.

Bu, J.; Tan, S.; Chen, C.; Wang, C.; Wu, H.; Zhang, L.; and He, X. 2010. Music recommendation by unified hypergraph: Combining social media information and music content. In *ACM International Conference on Multimedia (ACMMM)*, 391–400.

Buldú, J. M.; Cano, P.; Koppenberger, M.; Almendral, J. A.; and Boccaletti, S. 2007. The complex network of musical tastes. *New Journal of Physics* 9(6):172.

Castelli, G.; Mamei, M.; Rosi, A.; and Zambonelli, F. 2009. Extracting high-level information from location data: The w4 diary example. *Mobile Networks and Applications* 14(1):107–119.

Chen, L.; Wu, W.; and He, L. 2013. How personality influences users' needs for recommendation diversity? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 829–834.

Christenson, P. G., and Peterson, J. B. 1988. Genre and gender in the structure of music preferences. *Communication Research* 15(3):282–301.

Coulangeon, P., and Lemel, Y. 2007. Is 'distinction' really outdated? questioning the meaning of the omnivorization of musical taste in contemporary france. *Poetics* 35(2):93–111.

DeNora, T. 2000. *Music in everyday life*. Cambridge University Press.

Farrahi, K.; Schedl, M.; Vall, A.; Hauger, D.; and Tkalcic, M. 2014. Impact of listening behavior on music recommendation. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Foucard, R.; Essid, S.; Richard, G.; and Lagrange, M. 2013. Exploring new features for music classification. In *IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 1–4.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Cognos: Crowdsourcing search for topic experts in microblogs. In *International ACM SIGIR conference on Research and Development in Information Retrieval*, 575–590.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting personality from twitter. In *IEEE International Conference on Social Computing (SocialCom)*, 149–156.

Goldberg, A. 2011. Mapping shared understandings using relational class analysis: The case of the cultural omnivore reexamined1. *American Journal of Sociology* 116(5):1397–1436.

Hair, J. F.; Tatham, R. L.; Anderson, R. E.; and Black, W. 2006. *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ.

Hecht, B., and Stephens, M. 2014. A tale of cities: Urban biases in volunteered geographic information. In *AAAI conference on Weblogs and Social Media (ICWSM)*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 237–246.

Henaff, M.; Jarrett, K.; Kavukcuoglu, K.; and LeCun, Y. 2011. Unsupervised learning of sparse features for scalable audio classification. In *International Society of Music Information Retrieval (ISMIR)*, 681–686.

Hurley, N., and Zhang, M. 2011. Novelty and diversity in top-n recommendation–analysis and evaluation. *ACM Transactions on Internet Technology (TOIT)* 10(4):14.

Kruskal, J. B., and Wish, M. 1978. *Multidimensional scaling*, volume 11. Sage.

Lewis, K.; Gonzalez, M.; and Kaufman, J. 2012. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences* 109(1):68–72.

Leydesdorff, L., and Rafols, I. 2011. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics* 5(1):87–100.

McNee, S. M.; Riedl, J.; and Konstan, J. A. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 1097–1101.

Mörchen, F.; Ultsch, A.; Nöcker, M.; and Stamm, C. 2005. Databionic visualization of music collections according to perceptual distance. In *International Society of Music Information Retrieval (ISMIR)*, 396–403.

Peterson, R. A. 1992. Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics* 21(4):243–258.

Peterson, R. A. 1997. The rise and fall of highbrow snobbery as a status marker. *Poetics* 25(2):75–92.

Peterson, R. A. 2005. Problems in comparative research: The example of omnivorousness. *poetics* 33(5):257–282.

Pew Research Center. 2012. Demographics and political views of news audiences. `http://tinyurl.com/kbb6n9m`. [Online; accessed 30-December-2014].

Porter, A. L., and Rafols, I. 2009. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics* 81(3):719–745.

Qiu, L.; Lin, H.; Ramsay, J.; and Yang, F. 2012. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality* 46(6):710–718.

Quercia, D.; Kosinski, M.; Stillwell, D.; and Crowcroft, J. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *IEEE International Conference on Social Computing (SocialCom)*, 180–185.

Rawlings, D., and Ciancarelli, V. 1997. Music preference and the five-factor model of the neo personality inventory. *Psychology of Music* 25(2):120–132.

Rentfrow, P. J., and Gosling, S. D. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of personality and social psychology* 84(6):1236.

Rimmer, M. 2012. Beyond omnivores and univores: The promise of a concept of musical habitus. *Cultural Sociology* 6(3):299–318.

Schrammel, J.; Köffel, C.; and Tscheligi, M. 2009. Personality traits, usage patterns and information disclosure in online communities. In *British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, 169–174.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One* 8(9):e73791.

Stirling, A. 1998. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series* 28:1–156.

Stirling, A. 2007. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface* 4(15):707–719.

Turnbull, D. R.; Zupnick, J. A.; Stensland, K. B.; Horwitz, A. R.; Wolf, A. J.; Spirgel, A. E.; Meyerhofer, S. P.; and Joachims, T. 2014. Using personalized radio to enhance local music discovery. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, 2023–2028.

Van Dijk, T. A. 1996. Power and the news media. *Political communication in action* 9–36.

Wallin, N. L.; Merker, B.; and Brown, S. 2001. *The origins of music*. MIT press.

Warde, A.; Wright, D.; and Gayo-Cal, M. 2007. Understanding cultural omnivorousness: Or, the myth of the cultural omnivore. *Cultural sociology* 1(2):143–164.

Wikipedia. 2015. Last.fm. `http://en.wikipedia.org/wiki/Last.fm`. [Online; accessed 19-January-2015].

Yang, Z.; Wang, J.; and Mourali, M. 2014. Effect of peer influence on unauthorized music downloading and sharing: The moderating role of self-construal. *Journal of Business Research*.

Zheleva, E.; Guiver, J.; Mendes Rodrigues, E.; and Milić-Frayling, N. 2010. Statistical models of music-listening sessions in social media. In *International World Wide Web Conference (WWW)*, 1019–1028.