CS224W: Project Report (Group # 52)
Evolution of Mainstream Music Homogeneity over Time

Jess Fisher, Caleb Jordan, and Eric Yurko

Introduction

Comedy group “The Axis of Awesome” summed up a common grievance from internet-users in their 2009 “Four Chord Song” (http://www.youtube.com/watch?v=5pidokakU4I) -- that all popular music tends to sound the same. Irate comments to this effect are prolific on the YouTube channels of popular artists, but is there any weight to those claims? Obviously, music production is an art that has become refined over time, and studios have formed conglomerates that produce multi-platinum albums, but has that caused the artists in those studios to converge, musically, or have they simply diverged and produced their own dynamic sound? From forays into the internet, we’ve seen that the opinion among the masses is certainly that mainstream music is becoming more similar over time, but we’re not totally convinced. We sought a database with enough weight to settle this debate, and came across the Million-Song Dataset, which we will describe in further detail in the data collection section of this paper. Since it allows for and contains a notion of artist similarity, we can use that as our primary metric to determine homogeneity over time. Furthermore, utilizing a variety of heuristics, we can remove artists from the graph over time if they stop producing songs in order to give the graphs some notion of relevance. Ultimately, this will not only allow us to evaluate whether or not the music data in these graphs follows our notion of homogeneity but also allow us to attempt to produce a model that matches the graphs that this method produces, given the variety of graphs that the SNAP.PY library allows us to generate.

Review of Relevant Prior Work

The analysis of modern music is not an unexplored landscape. There have been papers published on the trends in popular western music, including “Measuring the Evolution of Contemporary West Popular Music” by Serra et al. This paper explores the specific metrics of pitch, timbre and loudness and their trends over a span of 50 years. This paper is particularly relevant to this project since the authors utilized the same million song dataset as this project. Combining loudness, timbre, and pitch, they attempt to formulate a notion of the music’s beat which is an important characteristic when discussing modern music. Interestingly, the authors discovered that the frequency of pitch sequences follow a power law distribution--a few sequences occur very often while the majority occur infrequently. This power law behavior was found within most of the years of data that were analyzed; therefore the distribution of pitch sequence frequencies was steady over time. However, this does not mean that the same pitch sequences were used in 1950s and in 2000s. In contrast to analysis on pitch, timbre features followed a power law distribution that varied over time instead of being static. The power law became stronger over time, which suggested that timbral variety was decreasing and the songs were becoming more homogeneous. While timbre and pitch followed a power law distribution, loudness followed a steadily increasing reverse log function otherwise known as the “loudness war”, in which music is getting louder over time in an escalating exchange between modern producers.

Overall, the authors concluded that music was static in terms of the nature of timbre and pitch frequencies while the exact sequences of features in the music changed over time. This led to the hypothesis that old music could be modernized by including the currently frequent pitch sequences and

increasing the loudness of the song without losing the character that identifies the old song's originality. This prior work is promising, since it seems that there are evident patterns in music trends regarding popular modern music in the last 100 years. We hope to expand on these findings and extend the same analytical thinking to the similarity metric we are analyzing.

While the previous paper utilized the same dataset that this paper will analyze, there are other music comparison datasets like WhoSampled.com's dataset. This dataset maintains a large collection of pairings of which artist's song sampled music from another artist's song and was produced from community contributors rather than one group. This crowdsourced approach likely explains its large size of 42,447 user-generated records of sampling song pairs. The relationship of musical sampling is a stronger indication of musical influence than songs appearing to sound alike, even though that is the basis for the artist similarity metric in the Million Song Dataset. However, the fact that the data is user-generated obviously creates concerns about the accuracy or potential biases of the data. Therefore, both of these datasets, WhoSampled.com dataset and the Million Song Dataset, have strengths and weaknesses when used to answer the question of music homogeneity.

Furthermore, the structure of this WhoSampled.com data means that any artist comparisons have to be summarized out of individual song pairings. In the paper "Musical Influence Network Analysis and Rank of Sample-Based Music", the authors Nicholas Bryan and Ge Wang attempt to combine this low-level data into overarching networks at the artist and genre levels. First, they built a acyclic song network using unique songs as nodes and each sampling pair as a directed edge from destination song to source song. An interesting observation from this initial network was that the cumulative in-degree distribution follows a power-law distribution. With this observation, the authors claimed that very popular samples will only continue to increase in popularity. However, this is the degree distribution over individual song influences, not the influences between artists. It will be interesting to see if our data between artist and their similarities also follow a power law distribution and what that will show us as far as the influence of popular artists. Second, after building the song to song network the authors propose a collapse-and-sum approach to transform this song network into an artist to artist network. Basically, for each node which represents a song for a particular artist, combine all of those nodes for one artist together into an artist node and keep all of the edges to other songs / artists nodes. After collapsing all artist nodes, if two nodes have multiple edges between then just collapse the edge into one influence edge. Note that edges in the artist graph are still directed. In a similar fashion, a genre graph can be made from the artist graph.

Finally, after creating a genre graph the authors tried to answer the question of music homogeneity that we are also asking. The authors decided to define a notion of genre entropy in the following way:

$$H_{g_k} = -\sum_{g_j \in \Gamma} P_{g_j|g_k} \log P_{g_j|g_k}$$

where g_k is the k^{th} genre in the set of genres Γ and $P_{gj \mid jk}$ is the probability of source genre g_j given the destination genre g_k. In this definition of genre entropy, if a genre only gets its sources from one other genre then the entropy will be zero. However, if all genres evenly sample from all other genres then the entropies will be maximized. Therefore, with their definition, a homogenous genre will have a smaller entropy. This notion of homogenous genre is based on the author's hypothesis that a homogenous genre will only sample from a small range of other genres while heterogenous (diverse) genres will sample from a variety of other genres. At the categorization level of genres, this definition

makes sense. Extending this definition to the artist network, a homogenous artist would only sample from a few other artists. By this notion of entropy, a homogenous set of artists would only sample from each other so the artist network would look like a series of cliques but this sounds like we are describing the notion of genres not of general musical homogeneity. Therefore, this paper will attempt to look at other possible indicators of musical homogeneity within a graph representing the similarity between artists.

Data Collection

For this project, we are utilizing data from the Million Song Dataset, which was formed in a collaborative project between LabROSA at Columbia University and The Echo Nest, a music data collection company. The data was collected using Web crawling to find online sources of music audio which were then analyzed using data mining and digital signal processing to form the artist and song metadata available in the Million Song Dataset. By not actually including any audio in the dataset, Echo Nest can use freely available online music sources to obtain their data without worrying about copyright issues. The dataset was create with the intent of facilitating research on music data so it is easily accessible from the project's website. However, the dataset is 280 GB, which took about a day to download and extract into a usable form. Once this data and its smaller indices were downloaded, we utilized python scripts and SQLite queries to extract the data we wished to analyze.

In the data, each artist has an ID as well as the artist name. This index was read in using python to parse the table data into a lookup table from artist ID to artist name. Next, we parsed the similarity relationship table in which each entry is a pair of artist IDs where the first artist has been found to have similar music to the second artist. This similarity relationship table is the product of the Echo Nest's similarity algorithm which utilizes audio content and cultural cues to determine the top similar artists to a given artist. Here is an excerpt describing their algorithm:

> *For acoustic similarity, The Echo Nest extracts music features directly from the audio. Unlike other automatic methods of acoustic feature extraction, our models are grounded on perceptual models of music listening. We model the auditory path through a series of psychoacoustic filters that mimic the human hearing mechanism, converting the audio wave form into musically meaningful data streams. These are useful for musical interpretation as well as for determining musical similarity.* - "The Echo Nest Song2Song"

A more detailed description of their similarity algorithm can be found on the quoted site, a link to which can be found in the citations section at the end of this paper.

In addition to the similarity data, song-specific metadata is available from the dataset. We will be utilizing the song year which is the year a song was officially released. In order to get this dataset, we had to parse and extract the song year and artist id from a hdf5 formatted file which was surprisingly difficult and required a specialized python package, pytable. Interestingly, 16,522 out of 44,745 artists did not have any songs with year data in the dataset. We only included artists with year data in our dataset since we are explicitly concerned with music over time.

Difficulties / Roadblocks

The major difficulty that we had in the first half of the project was an initial misinterpretation of the similarity dataset. When writing our proposal, we were under the impression that the similarity data was between individual songs. However, after further investigation while utilizing the dataset, we discovered

that the similarity relationships are between artists over their entire careers. This broadly widened the scope that the similarity relationship covered thus limiting the knowledge that we can gain from the data. After learning this information, we were still able to create the year graphs of similarity between artists using heuristics explained below.

Algorithms Overview:

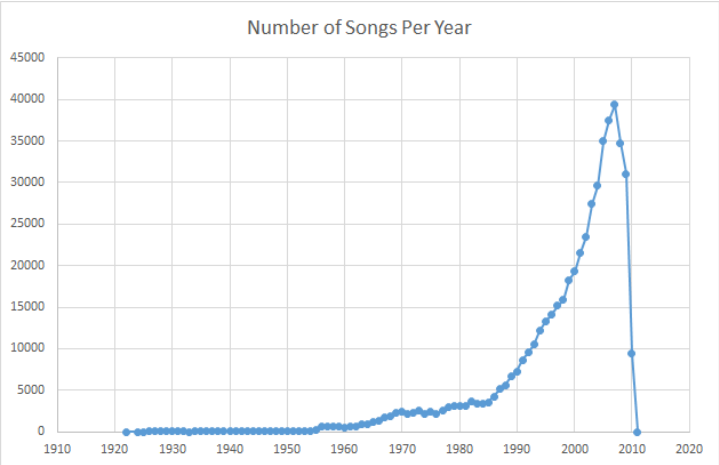
In order to study the evolution of artist similarity over time, we needed to build graphs representing the similarity network for each year in our dataset. We create directional edges representing artist similarity. In contrast to other papers that study evolving communities such as paper citation graphs in which nodes aren't removed once added, it seems natural to remove artists at some point when they are no longer active in the community. Using these principles, we built our set of graphs using the following rules:

1. Nodes represent artists, directional edges from artist a to b means that artist a has been labelled as similar to artist b.
2. An artist exists in the graph from the first year in which he/she publishes a song in our dataset.
3. We add an edge from a to b in the first year that both a and b exist if a is similar to b.
4. Nodes are removed after 5 years of having no songs listed.

After creating our real network, we constructed several other graphs so that we could compare various properties to other network models. We attempted to hold the number of nodes and edges constant across the graphs for each year. The first, pseudo-random graph that we generated followed almost the same mechanism as the real graphs; for every node added in the real network, we add a node in the random, but for every edge in or out of a new node in the real network, we randomized the other endpoint. In this way, the exact same number of in and out edges is created, but their connections are randomized.

In addition to the pseudo-random network, we created a series of small-world graphs and preferential attachment graphs. For a given year, we generate a PA graph of the same number of nodes as the actual graph and the same average node degree; the SW graph has the same number of nodes and average out degree, with a rewire probability of .25.

Summary Statistics



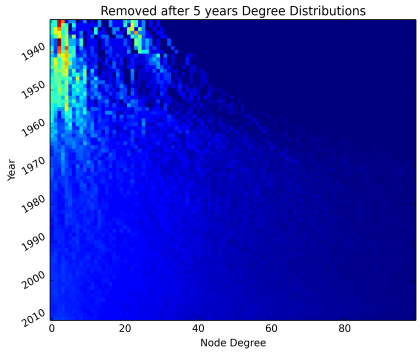
Clearly, most of our data is from after the 90's, however, we still believe we will get interesting data from the data points we have from earlier. Since most of the weight is on the right side of the graph, the average is a bit high (5728 songs per year [out of all songs containing year data, which is a little over 500000]). We lack data from before 1921 (which, while interesting, is not a huge loss) and after 2011, but this should be enough data for our purposes nonetheless. There are also over 44000 unique artists in the dataset, and, counting duplicates, over 2,000,000 similarity relationships between artists.

Results

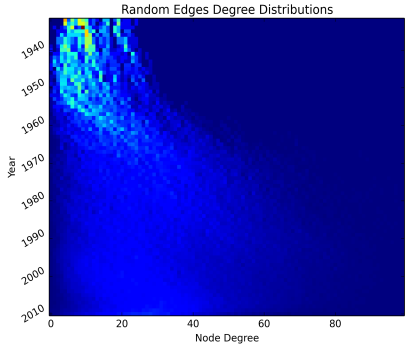
Since this paper is trying to answer the question of whether music is becoming more homogenous over time, we decided to focus on degree distribution and page rank score distributions over time as indicators for homogeneity. Degree distribution shows whether hubs, extremely connected nodes (high degrees), are appearing or whether there are many nodes that aren't connected to anything (extremely low degrees). Page rank scores show how many artists are influential. If a couple of few artists account for a large percentage of the page rank scores then these few artists have a strong influence on new music possibly leading to more homogenous music.

First we will analyze the general degree distribution evolution over time in our real similarity data pulled from the Million Song Dataset (graph a) and the other 3 possible network models which we are comparing it to (graphs b, c, and d).

a)

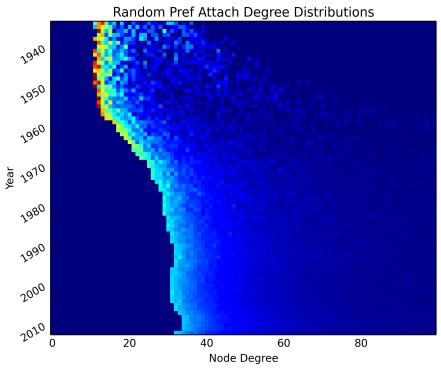


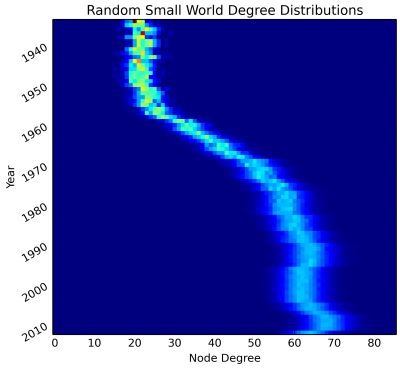
b)



c)

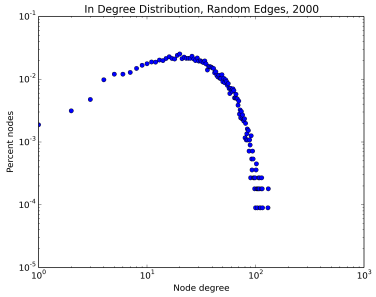
d)

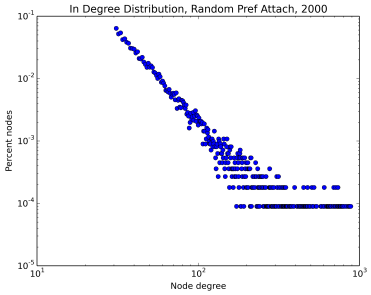


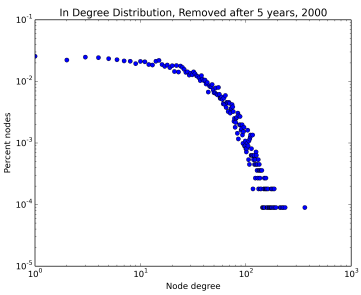


Over time, the mean degree follows same trend in all 3 graphs where initially the average node degree is smaller (around 10 to 20) but when the graphs reach about the 1960's which is when the number of songs starts to increase dramatically, the average node degree increases steadily. It is interesting to note that the preferential attachment graph and the small world graph follow similar patterns in their overall shape and concentration of node degrees within the distribution. Conversely, the real data from our constructed similarity graph results in a degree distribution evolution which is quite similar to the pseudo-random graph (b). At the beginning of the dataset, around 1940s, the degrees of nodes are mostly concentrated in the lower spectrum of 0 to 20 with an extremely similar bimodal structure where there's a non-trivial concentration of nodes around 20 then a slight gap as you decrease the degree then a large grouping of nodes at the lower degrees in both graphs (a and b). This early behavior can be attributed possibly to the lack of electronic media and communications which allowed artists to share their music with a broader audience. This lack of sharing mediums could result in these isolated artists who are not similar to any other artists.

Around the 1970's and 1980's, the degree distribution drastically spreads out in both the real similarity graph and the pseudo-random graph (a and b). More nodes are connected to other artist and there are nodes evenly spread along the degree distribution. This spread of connectivity for artists could be interpreted as more sharing and influence between artists which could lead to more homogeneity in the music landscape. However, the overall degree distribution leads to general trends across time. To learn more about the state of homogeneity in modern music, we took a closer look at the degree distributions for a single year in our constructed similarity graph and some of the models.





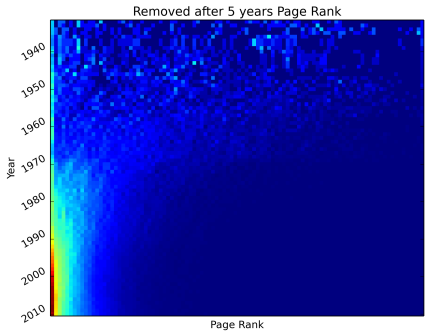


Above, we have the in-degree distributions of three graphs from the year 2000. In the first, when a node is added we randomly add edges until we hit the number of edges it would have had in our original graph. The second graph is a random preferential attachment graph, utilizing the notion that popular artists will get more popular over time. The third one is the graph made from our actual data--in it, we remove an artist if they have not released a song present in our data for 5 years. Intriguingly, it

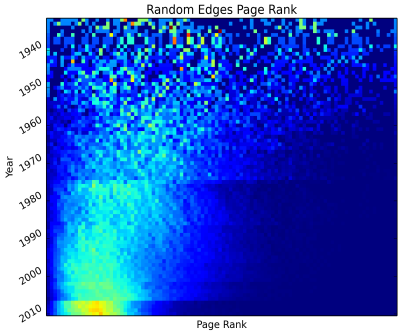
bears a striking similarity to both graphs. While it retains the sort of tail-end diffusion of the preferential attachment graph, the beginning looks a lot more similar to the random graph--there's no clear straight line path like the preferential attachment graph has. Why would this be the case? Well, it's not necessarily true that once an artist becomes popular they will start to develop a critical mass of similar artists following them--they may only achieve popularity inside of their genre. Secondly, achieving success and popularity is somewhat random--it seems like many nodes achieve a fair number of similar artists but never manage to break out into large-scale popularity or adoption, guaranteeing that there will always be a fair number of low-degree nodes. Overall, the graph seems to be a hybrid of the preferential attachment and random graphs, which is interesting--once an artist has sort of crested the hill, they will start following a rich-get-richer model, but before then it's essentially random how many artists will be considered similar. Further work in this area might be to try this inside of genres--especially in genres with sort of notable classics (such as rap or classical), it may be the case that the graphs start to follow more of a preferential attachment model as musicians cite Tupac or Mozart as their inspiration (but likely not both at the same time).

The below four graphs show the distribution of page rank scores for artists over time for the following graphs: a) Similarity data graph where artists are removed after 5 years, b) Pseudo-random graph, c) Preferential attachment graph, and d) Small world graph.

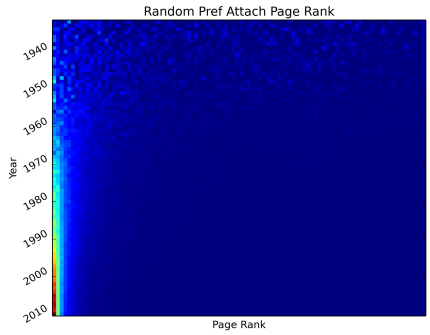
a)



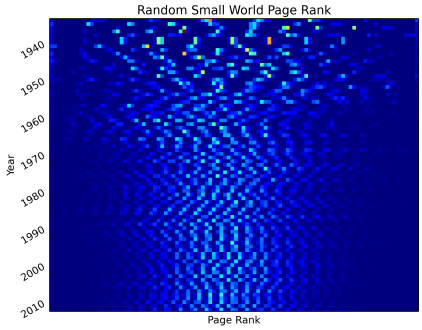
b)



c)



d)



Each of these graphs represents PageRank density over time--each row represents a histogram over the range of PageRank values, but each row does not necessarily represent the same range as each other; in this way we hope to show the spread of node influence, normalizing somewhat for the extreme difference in number of nodes over time.

Out of all of these graphs, the smaller world page rank distribution graph on the bottom right is

the outlier with evenly distributed page ranks for artists with a trend where artist page ranks become more similar with less variance over time. However, the page rank scores are still evenly spread within the smaller range of values. A quick visual check against the other graphs shows this behavior to not match our similarity graph's behavior, and therefore we can strongly rule out the small world model as possible model for our similarity data.

When looking at the other three graphs, the real world similarity graph page rank data seems to be a hybrid of the pseudo-random graph data and the preferential attachment data. At the beginning of our data set which is before the 1960's, the similarity graph has a more even spread of page rank scores than the preferential attachment graph but not as evenly spread as the pseudo-random graph. The page rank scores are still weighted towards the low end in the similarity graph especially as it moves further in time towards what people would consider modern music of the 1990's and 2000's. Therefore, in the first half of the dataset in terms of time, it seems that there are nodes with many different levels of influence. However, as we move into the second half of our data set in terms of time, the similarity graph of page rank distribution in our data set show an increasing number of nodes with a very small page rank as well as a few nodes with large page ranks. Therefore, it seems that as music evolves there are just a few very influential people, for example the Beatles, but most artists do not have any influence on other artists. In fact, the striking similarity between the second half of the similarity graph data and the preferential attachment graph data suggests that pockets of artists are forming where each artists has a small group of other artists that they influence but outside of that group there isn't a lot of similarity. Like the degree distribution analysis before, this could be explained by artist influence being restricted to their genre. Another idea that could explain the super influential nodes in the page rank distributions is the notion of publicity inflation where artist will describe themselves as similar to a currently popular artist in hopes of gaining fans of those other artists and thus publicity. This would artificially increase the page rank score for the attributed artist. With all of these observations, there still doesn't seem to be conclusive evidence to show that homogeneity in modern music is increasing over time.

Findings

Overall, we've found that the similarity graphs we've generated tend to be some sort of hybrid between random and preferential attachment graphs, which does make sense given our original claim--if music is becoming more similar over time, then the "rich-get-richer" model of a preferential attachment graph should be our expected result. As we've found, the fact that our graphs are sort of situated between both random and preferential attachment suggests that there is a lot of variety in non-mainstream artists, but as artists get more and more popular they tend to become a lot more similar to other artists. However, the fundamental question is whether or not that tendency is increasing over time. Just looking at the degree distributions, the answer seems to be no. In fact, it seems that the artist degree distributions tend to favor the random graph over time (and small-world is a completely incorrect model). Looking back at the degree distribution for 2000, it does make sense--it resembles a hybrid of the two graphs, or at least a random graph informed by a preferential attachment model, which is pretty interesting.

So what does this mean for music? So far, it means that based off of our data, everything is currently fine. Music is not all becoming the same--rather, different artists are popping up and going out of vogue, and their cycles are reflected in our heuristics, as we remove an artist after 5 years without production. They likely will have faded from the public eye by then (anyone remember the Jonas

Brothers?), but at that point their removal allows the graph to evolve over time so that we can see the actual trajectory of music. If the artists were becoming more similar, even with removals the degree distribution would start to transition more towards the preferential attachment model than the random model, as the “rich” artists (those who are already pretty similar) would be getting more similar. Furthermore, our graph seems to have even more low-degree nodes than preferential attachment or the random graph, further arguing in favor of heterogeneity. While we’d argue that this is conclusive, we’d also like to see further work that takes on additional artists and notions of similarity to see if these results can be replicated. Even more interesting would be replicating these tests on datasets that support song similarity, as that would result in more granular findings than we were able to produce. Though these results are a compelling argument against homogeneity, we would recommend some additional experiments in this area before breaking out the laptop and trying to take down the YouTube trolls. In the meantime, rest easy--music is safe.

Citations

1. Bryan, Nicholas J., and Ge Wang. "Musical Influence Network Analysis and Rank of Sample-Based Music."
12th International Society for Music Information Retrieval Conference (ISMIR) (2011): 329-34. *CCRMA*. Stanford Unviersity. Web. 09 Dec. 2013.

2. Serra, Joan, Alvaro Corral, Marian Boguna, Martin Haro, and Josep Arcos. "Measuring the Evolution of Contemporary Western Popular Music." *Scientific Reports* 2 (2012): n. pag. *Scientific Reports*. Web. 13 Nov. 2013. <http://www.nature.com/srep/2012/120726/srep00521/full/srep00521.html>

3. "The Echo Nest Song2Song." *The Echo Nest*. N.p., n.d. Web. 13 Nov. 2013. <https://echonest-corp.s3.amazonaws.com/docs/whitepapers/Song2Song-1_0.pdf>.