Improving Music Artist Recommendations Through Analysis of Influences

Matt Grimm
Department of Computer Science
University of Alaska Anchorage
Anchorage, Alaska, USA
matt.t.grimm@gmail.com

Bilal Gonen
Department of Computer Science
University of West Florida
Pensacola, Florida, USA
bgonen@uwf.edu

Abstract—To improve the quality of search results in huge digital music databases, we developed a simple algorithm based on artist influences and complex network theory that produces interesting and novel results. Traditionally, music recommendation engines use audio feature similarity to suggest new music based on a given artist. We propose a search that takes influences into account provides a richer result set than one based on audio features alone. We constructed an artist influence network using the Rovi dataset and studied it using complex network theory. Analysis revealed many complex network phenomena which we used to tune the search algorithm. Finally, we consider the difficulty of qualitatively rating our results and the need for a tool to exercise the algorithm.

Keywords—complex networks; recommender system; artist network

I. INTRODUCTION

With the explosive growth of digital music databases comes an increasing challenge to provide relevant recommendations. Much of the existing research in this area focuses on using audio features to determine similarity between pieces of music. Such features include timbre, harmony, melody, and beats per minute. In this work, we study influence relationships between musical artists using complex networks theory, with the goal of improving precision in music recommendation systems. We show that incorporating influence relationships in search algorithms reveals interesting results, and in combination with a standard similarity search, provides a novel result set.

We constructed a network where the nodes represent artists and the links represent an “influenced by” relationship. The network is directed and cyclic, so two artists may have zero, one or two directed links between them. Mostly, there is at most a single directed link between any two artists, but it is technically possible that two artists can influence one another. To our knowledge, the source dataset does not restrict this possibility. The artist data was extracted from Rovi’s extensive and well-groomed dataset of music artist information. Rovi curates the data that powers AllMusicGuide (AMG), an essential online music database, and it is experts employed by Rovi Corporation that create the influencer relationships within the dataset.

In our network, the indegree of a node indicates how many other nodes are influenced by it, while the outdegree indicates how many artists the node is influenced by (or, how many influences the artist possesses). The degree correlation of the resulting network shows that nodes with a low indegree will prefer to associate with nodes with a higher indegree, possibly because lesser-known artists often emulate well-known ones. The outdegree of an artist in our network is not specifically limited by the Rovi dataset; the API has a hard limit of 100 results for the influencers data, but our queries retained at most 10. Older artists in the dataset tend to have fewer influencers, especially when the influencers are mostly classical composers. We explore the relationship of indegree and betweenness to a node’s importance in search results.

Multiple previous studies considered the “similar to” relationship between artists [1–3], but few specifically addressed the “influenced by” relationship.

II. RELATED WORK

[2] applies complex network theory to music recommendation networks to discover possible optimizations in their design. They use artists as the graph nodes and similarity as the links to other artist nodes. Four separate directed networks are created, each based on a different online music database. They found that all four networks have small world characteristics, and suggest this is a positive finding given that humans navigate small world networks easily. The outdegree of an artist node in their networks is bounded by web page size since the “similar artist” links must fit on the pages from which they mined the dataset. This is an unfortunate limitation, and one that would not be an issue had they used a public API like Rovi’s or an extant dataset like the MSD. The most interesting conclusion is that the structure of the four networks is dependent on social data.

In [1], the authors use two sources of data to build artist similarity networks: expert opinion mined from AllMusicGuide (AMG), and playlist cooccurrence mined from a social music website. If an artist has “similar artists” links on AMG, then an undirected link is created between each pair of related artists. And if two artists coexist on a playlist from the social music website, a link is created between them. Their findings show that the resulting networks are sparse but well-connected due to one large component that connects the majority of the nodes. The networks also have small world properties. The idea of using social playlists is intriguing but the data is almost certainly noisy since a user can add any
A risk of the snowball sampling technique is choosing a poor seed artist to start the network. If there is little or no influencer data available for the artist, the resulting network may terminate at a shallower depth than our limit and therefore may not be representative of the network as a whole. It is also possible to choose a seed that generates a network with a low branching factor, resulting in a small number of children for each node. Such a network would be sparse and relatively unconnected, giving little insight into the relationships between artists. Finally, the bias of the researcher in selecting a seed artist may also skew the results.

To avoid these pitfalls, we constructed multiple networks, choosing each seed artist from a different genre. The genres were deliberately selected to be as broad as possible: Pop, Rock, Rap, Country, Jazz, and Blues. We removed the possibility of researcher bias by selecting our artists from the 2011 year-end Billboard charts. Each genre has a chart for one or both of “<Genre> Albums Artists” and “<Genre> Songs Artists”, and we chose the top artist from one of these charts. The Billboard charts are calculated using proprietary methods that account for total sales as well as streaming and radio play.

The influencers network is constructed in reverse temporal order, so all followers of the selected artist will be excluded. For example, if the seed artist is Radiohead, then its follower The Arcade Fire cannot appear in the network. Technically, this is also a problem with the opposite network, where we start with an ancestor artist and the links indicate an “influenced” relationship. Any ancestor of the seed artist would not appear in the resulting network. However, music genres have relatively definite origins and so a good seed artist would be easier to select in that case. In our influencers network, we simply select a window of time that will constrain our network and ignore any ancestors or followers outside that window.

B. Data Mining Strategy

Rovi’s API is accessible via a RESTful HTTP interface. A RESTful web service is just a simple HTTP service where the service’s resources are identified in the Uniform Resource Identifier (URI) and can be operated upon with a set of verbs. In the Rovi API, the URI defines the search parameters and the desired return values. Responses may be in either XML or JavaScript Object Notation (JSON). Rovi permits free ratelimited access to their API for non-commercial purposes, once one signs up for an API key.

To mine the artist and influencer data from the Rovi API, we developed a custom Perl script. The script contains a function which is called recursively to a specified depth and which saves a specified number of influencer relationships to a flat file in JSON format. A second script converts that data into Simple Interaction File (SIF) format, suitable for importing into Cytoscape, a tool for visualizing and analyzing complex networks [8]. We decided upon a recursive depth limit of 5 and an influencers limit of 10 to generate the artist influence networks. A depth of 5 ensured we would not exhaust the available data in the API by descending too far into the past; eventually, all influencers are classical, and we were not interested in classical music for this research. Similarly, an influencers limit of 10 meant we would have a network with a sufficient branching factor, without having our results diluted.
by the less relevant influencers. We executed the script once for each of the six seed artists, resulting in six separate network files. After importing these networks into Cytoscape, we merged them into a single network using Cytoscape’s Advanced Merge Tool. The final node and edge counts for the separate and merged networks are displayed in Table 1.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Artist</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>Katy Perry</td>
<td>522</td>
<td>2,856</td>
</tr>
<tr>
<td>Rock</td>
<td>Foo Fighters</td>
<td>432</td>
<td>2,334</td>
</tr>
<tr>
<td>Rap</td>
<td>Chris Brown</td>
<td>272</td>
<td>1,298</td>
</tr>
<tr>
<td>Country</td>
<td>Blake Shelton</td>
<td>454</td>
<td>2,525</td>
</tr>
<tr>
<td>Jazz</td>
<td>Michael Buble</td>
<td>269</td>
<td>1,138</td>
</tr>
<tr>
<td>Blues</td>
<td>Joe Bonamassa</td>
<td>224</td>
<td>936</td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>1,173</td>
<td>6,810</td>
</tr>
</tbody>
</table>

Table 1. Node and edge counts for genre networks

IV. RESULTS

The artist influence network shows many properties of a complex network. The characteristic shortest path is 5.157 and the (directed) network diameter is 20, indicating that any artist can be reached from any other artist in relatively few steps. Considering the span of genres and time periods this network represents, a diameter of 20 is still somewhat short. The clustering coefficient $C = 11.2 \times 10^{-2}$ shows that the network is a small-world one, since the equivalent random graph has a significantly lower value of $C = 0.495 \times 10^{-2}$. Given an equivalent random graph, we can calculate the clustering coefficient $c_r$ as:

$$c_r = \frac{k_a}{n}$$

where $k_a$ is the average node degree and $n$ is the number of nodes in the network. The average node degree of a directed graph is given by:

$$k_a = \frac{m}{n}$$

where $m$ is the number of edges in the network. Interestingly, the average indegree and outdegree are both 5.81, while the average number of neighbors is 11.57. So on average each artist has approximately 6 influencers and 6 influencees. From observation, it is unlikely that the API is limiting the results. Many modern artists have 20, 30 or even more than 40 influencers in the Rovi dataset. Indeed, an early version of the API crawler script did not filter out duplicate queries and quickly exhausted our daily limit of requests due to the exponential nature of the unfiltered graph. If duplicate artist queries are allowed, the growth is given by $n = i^d$, where $n$ is the number of nodes, $i$ is the influencers limit, and $d$ is the recursive depth limit.

Finally, the indegree distribution follows distinct power law decay, with a few well-connected hub nodes containing a much higher indegree than the average. This is somewhat intuitive in the world of music artist influences, when one considers that there is often one seminal artist that launches or popularizes a genre and inspires many future generations of artists. The five most influential artists in this network are The Beatles, Bob Dylan, Hank Williams, James Brown, and Chuck Berry. The Beatles can certainly be said to have popularized the Rock genre, and Hank Williams is an essential Country artist. James Brown and Chuck Berry both appear in the majority of Rap and Hip-Hop artist’s influencers’ lists.

V. DISCUSSION

With few exceptions, no artist creates in a vacuum, and must therefore have at least one influence. Even the most original innovative artist will have links back to his roots, probably in some well-established and highly clustered genre. This link, however, will separate the innovator’s ancestors from his followers into separate communities.

A. Search Algorithms

In its most basic form, an influencers search consists of just two steps: 1) select a seed artist, and 2) return the influencers of that artist. This is the idea behind Music Bloodline, a website that traces ancestors and followers of music artists using the Rovi API [9]. There are several drawbacks to the naive approach. First, the results are unlimited, so information overload is a risk. In our network, 13.37% (157) of the artists have more than 10 influences, and several have over 50. Second, simply looking at an artist’s influences is not necessarily the best way to find artists with a similar “sound”. An artist may count among their influencers’ artists from a wide range of genres and time periods, whether or not that influence is actually apparent in the music. While we are not privy to the processes that Rovi’s experts employ to create the influence associations, it is reasonable to assume that an artist may be influenced by another artist based on non-musical grounds. For example, they may share a political philosophy or simply have common members. Whatever the case, if the association is not musically-based, the quality of the search result may suffer.

As an enhancement to the naive influencers search, our search algorithm incorporates the influencers’ network, but uses it to find contemporaries of the seed artist that share common influencers. The algorithm’s steps are as follows: 1) select a seed artist, 2) for each influencer, find artists influenced by this influencer (subject to pruning and limiting), and 3) sort, prune and limit overall results. Each subset of results is pruned and limited, and the overall result set is also pruned and limited. If a subset has fewer than 10 results, we do not prune or limit the results.

Before they can be pruned and limited, the results must be sorted by relevance. We defined relevance three ways with varying return depending on seed artist: betweenness, indegree, and clustering coefficient. Since betweenness increases roughly linearly with indegree in our network, both it and indegree were essentially equivalent for the purposes of sorting search results. The clustering coefficient produced slightly better quality results in terms of “likeness” to the seed artist. Next, pruning refers to the process of removing some percentage of the results using a metric designed to reduce “noise”. In this case, we chose to apply the power law and remove the top 20% of the sorted results. This technique removes the most influential artists that are likely to appear in any search, and are
therefore less valuable when the goal is discovery of new music. We assume our users do not need to “discover” artists with the highest indegree like The Beatles, Bob Dylan, or Chuck Berry because they are already well-known. Finally, the results are limited to allow a broader set of results without flooding the user with recommendations.

B. Evaluating Results

Using our enhanced search algorithm, we found some promising results and discovered an important drawback. We discovered that the world of music artist influences is diverse and well connected, with artists frequently citing influences across genre boundaries. These cross-genre references create two-length paths between artists with very little in common. For example, Public Enemy, considered the first popular political rap group, is connected to The Who, a progressive arena rock band, via James Brown, “The Godfather of Soul.” It is not a stretch to say that a user who seeds a search with Public Enemy would be surprised to see The Who among the results. Another example finds Frank Zappa, a radical 60’s rock artist, connected to Ray Charles, one of the inventors of soul and R&B, by way of Guitar Slim, an early blues performer [9]. It is not that it is necessarily surprising that these artists share the common influence, but that they appear in the same search context. Despite these occasional unrelated results, our algorithm did perform relatively well in finding musically similar results. Using the Public Enemy example again, we found nine influencers in the first step of the algorithm. Since this was less than 10, we kept all nine and did not prune or limit at this stage. In the second stage, all followers of these artists were found. Finally, we sorted the list by clustering coefficient in descending order, selected the top 10 artists and pruned off the two with the highest clustering coefficient (“Above the Law” and “Kool G Rap & DJ Polo”). Table 2 lists the nine influencers and the final search results.

<table>
<thead>
<tr>
<th>Influencers</th>
<th>Search Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funkadelic</td>
<td>Rakim</td>
</tr>
<tr>
<td>Ice-T</td>
<td>OutKast</td>
</tr>
<tr>
<td>James Brown</td>
<td>Scarface</td>
</tr>
<tr>
<td>Kool Moe Dee</td>
<td>Sly &amp; Robbie</td>
</tr>
<tr>
<td>Kurtis Blow</td>
<td>Whodini</td>
</tr>
<tr>
<td>Parliament</td>
<td>Sylvester “Sly Stone” Stewart</td>
</tr>
<tr>
<td>Run-D.M.C.</td>
<td>Eddie Hazel</td>
</tr>
<tr>
<td>Sly &amp; the Family Stone</td>
<td>Naughty By Nature</td>
</tr>
<tr>
<td>The Last Poets</td>
<td></td>
</tr>
</tbody>
</table>

*Results sorted by clustering coefficient. Table 2 Search results for “Public Enemy”

VI. FUTURE WORK AND CONCLUSION

Evaluating the quality of search results cannot be done with algorithms and network analysis alone. We can score the results by comparing against other well-known services that present similarity relationship. Two examples we explored for future research are playlist co-occurrence on the social music website Art of the Mix [10], and co-occurrence in the Amazon API’s CartSimilarities results [11]. If we take a seed artist and one of the search results from our network, and those two artists share a playlist on Art of the Mix that provides some evidence that the two have some commonality.

The limit of our qualitative analysis was to “eyeball” the results of manual searches seeded with familiar artists. Obviously what is needed is a focus group of testers who can use a custom-built tool to perform searches and rate its results according to “quality”.

Searches that link artists by similarity tend to be horizontal, that is, they largely locate artists that are contemporaries of the query artist. If the similarity between artists is established using audio features, then it follows that most of the links are between musical peers because the “sound” of a genre does not change radically from artist to artist. Major shifts in harmony, melody, and timbre within a genre are affected over decades. By using influence relationships, we can also delve back into the network of artists that inspired the query artist. These artists may not share exactly the same audio characteristics, but could instead highlight some musical root from which the query artist evolved. In combination with similarity metrics, we can improve the overall quality of user-guided music searches by incorporating influence relationships.

REFERENCES