The Musical Collaboration Graph

ELE/COS 381 Final Project

Akash Levy, Sunny He, Vincent Po, Daniel Wood, Erica Wu

January 16, 2017

1 Introduction

The recent rise in music streaming services such as Spotify presents the opportunity to collect and understand musical data at greater scale than ever before. In the past two decades, pop music has seen a rise in collaboration among top artists, making it interesting to analyze the complex collaborative relationships between them. Using graph centrality measures and ranking algorithms such as PageRank and HITS, it is possible to understand the network of collaborations between musical artists. Analysis enables us to construct new measures of an artist's influence based on their importance in the musical collaboration network. Additionally, we can compare the connection between an artist's degree of collaboration with common metrics of popularity such as downloads or sales.

The ideas presented in this work have significant potential for artists and labels seeking to understand how collaboration can affect popularity and sales. For example, two artists with a large degree of separation in the collaboration graph may find it mutually advantageous to collaborate on an upcoming track. In this way, both artists would be able to increase their relevance in the music industry and possibly even gain exposure to new listeners outside of their genre.

2 Methodology

We used the Spotify API to collect the data required for the musical collaboration graph. Despite having a wide variety of choices for collecting music data, Spotify provided several key advantages:

- Large database of artists and songs
- Accurate artist collaboration info
- Fast API response time
- Almost zero rate-limiting
- No need to login
- Artist profile pictures available through API
- An easy-to-use library available in Python: Spotipy

	Number of Nodes	Number of Edges	Average Degree
Graph 1	12,813	19,991	3.1204
Graph 2	2,302	9,325	8.1017
Graph 3	999	4,999	10.0080
Graph 4	652	4,652	14.2699

Table 1: Basic graph properties.

As a result of limited computational resources, we were able to extract only a very small subset of Spotify's enormous artist graph. As a result, there were many design decisions to make when considering how the subgraph should be extracted and processed. We decided to perform a breadth-first search (BFS) starting from the modern mainstream pop artist, Drake, who is currently the most played artist on Spotify. We performed 1,000 iterations of BFS, expanding outwards from all of Drake's collaborators, and then post-processing the graph in various ways.

Bias in the centrality measures is inevitable when extracting a subgraph using BFS. The centrality measures are biased towards the nodes searched earlier since these nodes, by the nature of BFS, will be more central. We tested four variants of our graph with different levels of post-processing:

1. No post-processing
2. Removal of degree 1 nodes
3. Trimming of the graph to only the 1,000 fully searched nodes
4. Trimming of the graph to only the 1,000 processed nodes followed by removal of degree 1 nodes

Each of the post-processing techniques produced graphs with very different properties, as indicated in Table 1. Graph 1 had over 12,000 nodes, which is above 12x the number of nodes processed during BFS. After pruning degree one nodes, this number was reduced almost six-fold in Graph 2, while the average degree increased to above 8. In Graph 3, where only nodes that were fully searched during BFS were kept, the average degree rose to 10. Finally Graph 4, which is equivalent to Graph 3 but with degree one nodes removed, had a relatively small number of nodes and a high average degree of over 14. A rendering of Graph 1 is given in Figure 1 and a rendering of Graph 4 is given in Figure 2. The images were rendered with Gephi, an open-source graph visualization tool.

3 Directed vs. Undirected Networks

The network of artist collaborations can be represented as either a directed or an undirected graph. In either case, nodes represent artists, and edges represent collaborations. However, in the music world and on Spotify, every track has a main artist and one or more featured artists, whether or not the artists have equal parts (e.g. Drake feat. Rihanna and Lil Wayne). For each secondary artist, we added an edge connecting that artist to the main one. In an undirected graph, the edge does not have a direction, meaning that it does not account for the difference between the main and featured artists. In a directed graph, an edge starts from each of the featured artists and points towards the main artist.

The undirected graph assumes that all artists have equal participation in the track and that all artists experience the same results of the collaboration. It can be thought of as a friend graph, where the main artist is friends with all of the collaborating artists; here, friendship is

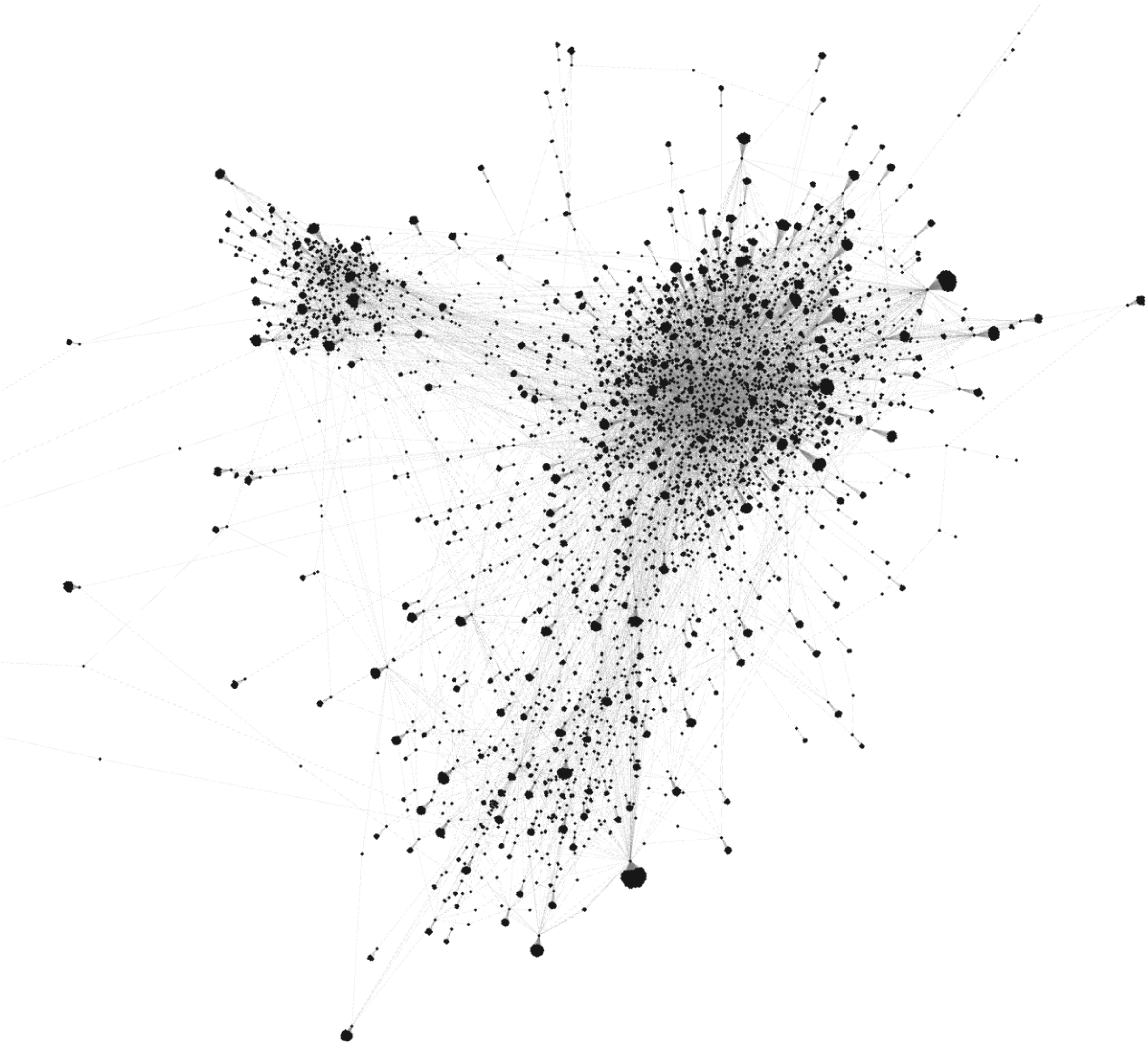

Figure 1: A Gephi rendering of the unprocessed graph, Graph 1.

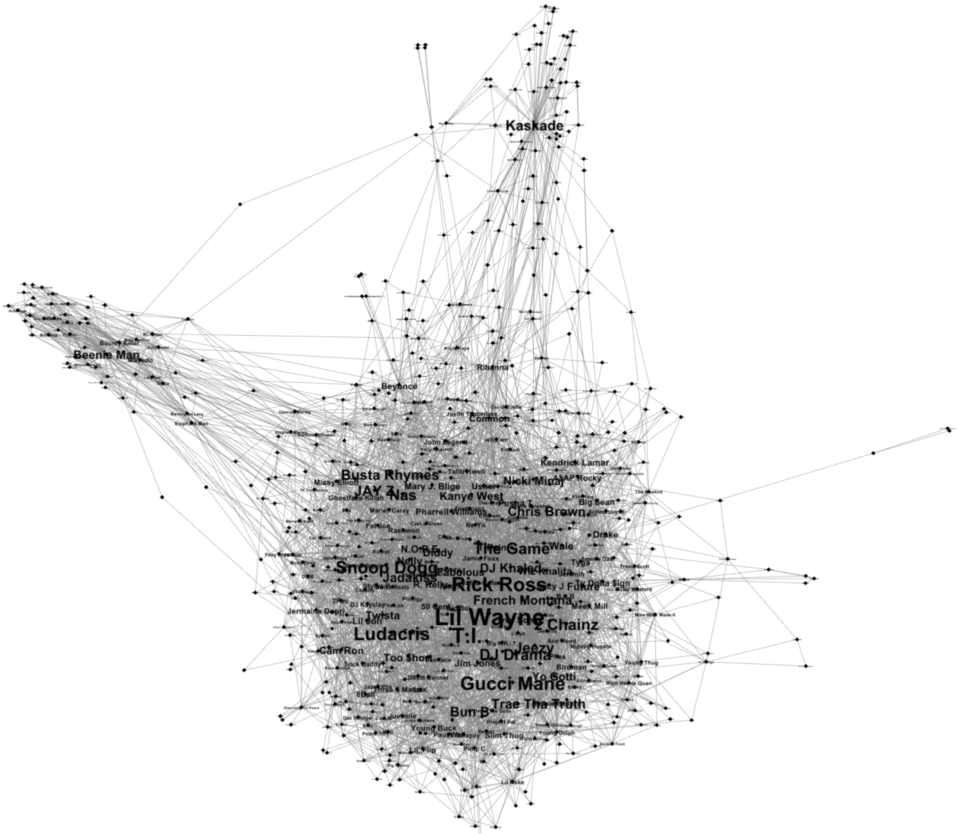

Figure 2: A Gephi rendering of the final pruned graph, Graph 4.

treated as a mutual and balanced relationship. The directed version of the graph assumes a different dynamic, treating each featured artist in a collaboration as only a supporter of the main artist. Since both interpretations are valid, we built and analyzed both an undirected and directed version of the network of collaborations. For the directed graph, we decided to have the edge pointing to the main artist to represent the flow of importance from the supporting artist to the main artist (although a case could be made for importance to flow the other way).

4 Measures of Centrality

We examined the topology of the graph, using different heuristics to determine the importance of different artists. We consider seven measurements of centrality, which together capture a measure of the comparative importance/influence between the network of collaborating artists. Whether or not the graph is directed determines which heuristics are relevant. For the case of the undirected graph, we consider degree centrality, eigenvector centrality, closeness centrality, betweenness centrality, and PageRank. For the case of the directed graph, we additionally consider two elements of Jon Kleinberg's HITS algorithm: Hub and Authority scores.

Table 2 contains the rankings based on the different centrality measures for the undirected graph. Table 3 contains the rankings for the directed graph. We constructed an aggregate score for each artist that equally weights the rankings from each measure. The "Spotify" column represents the popularity score (0-100) reported by Spotify, which is an indicator of how popular a given artist currently is (based on Spotify streams).

The degree centrality of an artist is the most intuitive measure of influence—it is simply the number of collaborators each artist has. It is a good measure of importance because we would expect an artist with many collaborations to be more influential in pop music than an artist with fewer collaborations, assuming that the collaborators are similarly popular. For example, we can see that in Table 2, Lil Wayne had the most number of collaborations (the number was 132). However, there is a problem with just considering degree—in reality, some collaborations are much more valuable than others, so we would like a heuristic that prioritizes artists that collaborate with influential artists, rather than relatively unknown artists.

The eigenvector centrality algorithm addresses the drawback of degree centrality by computing the eigenvector corresponding to the largest eigenvalue of the network adjacency matrix $\boldsymbol{A}$. This measure of centrality can also be computed iteratively by repeatedly multiplying a centrality vector $\vec{x}$ by the adjacency matrix $\boldsymbol{A}$. Intuitively, this spreads the influence across the graph in an successive manner, so that as certain nodes collect importance and influence, the nodes they point to also become gain a larger importance boost than if the source node were less important. As a result, once $\vec{x}$ has converged (or after some specified maximum number of iterations) we again obtain the eigenvector for the largest eigenvalue of the adjacency matrix A. PageRank computes importance in a similar way.

We also consider closeness centrality, which indicates how close an artist is on average to other artists. In order to determine the closeness centrality, we take the reciprocal of the sum of the shortest path distance from a node to all other nodes. This heuristic takes a global view of influence, disqualifying artists that might simply be popular in small sub genres, since it considers all of the shortest paths to every other artist. We could expect a rapper such as T.I. to have a high closeness centrality because he has been featured on multiple high ranking albums ranging a wide span of other artists from The Game to Busta Rhymes.

Another measure of centrality, betweenness centrality, also provides valuable information in the context of a musical collaboration graph. It is often employed when analyzing social networks, such as Facebook and Twitter. Artists with a high betweenness centrality are notable for the fact that they lie on a large number of shortest paths between pop artists, and therefore serve as a "who's who" in the pop music industry. Kaskade topped the betweenness centrality rankings for both the directed and undirected graphs.

For the case of the directed graphs, we also analyzed metrics from Jon Kleinberg's HITS algorithm, since the direction of influence provides a different layer of information about the nature of collaboration. The HITS algorithm is similar to algorithm we used to generate the Musical Collaboration Graph because it is also query dependent, at least in the sense that Graph 4 was compiled by running Breadth First Search starting from Drake. For this reason, the resulting graph would be slightly different if we were to recreate the graph starting from each artist. When we examine the data in Figure 2, it makes sense that an artist like DJ Drama would have the highest HITS authority score. Recall that we represent influence with a directed edge pointing towards the main artist in the song. Thus the artist with the highest authority score would be someone who is the main artist with a large number of collaborations from artists with large hub scores. DJs usually produce many remixes and list their collaborators as the artists they feature, so it makes sense that DJ Drama tops the list.

5 Homophily in Pop Music

Aside from measuring the importance and influence of each individual artist, we also wanted to examine how the musical collaboration graph's topology compared with other networks, such as typical social networks. Exactly how prevalent is collaboration in the pop music industry? Is the industry a "small world", i.e. does the median of shortest paths between artists grow at a rate on the order of the log of the number of artists?

One convenient way to represent the connectedness of the network is to calculate the clustering coefficient. We suspected that collaborating artists, much like actual friends on a network, have many more edges in the graph than a random graph, such as the Poisson graph or Erdös-Renyi

#	Degree	Eigenvector	Closeness	Betweenness	Pagerank	Aggregate	Spotify
1	Lil Wayne	Lil Wayne	T.I.	Kaskade	Kaskade	Lil Wayne	90
2	T.I.	Rick Ross	Lil Wayne	T.I.	Lil Wayne	T.I.	80
3	Rick Ross	Gucci Mane	Rick Ross	Lil Wayne	T.I.	Rick Ross	82
4	Gucci Mane	T.I.	Nicki Minaj	Beenie Man	Rick Ross	Ludacris	77
5	Ludacris	The Game	Ludacris	Beyonce	Ludacris	Gucci Mane	89
6	Snoop Dogg	DJ Drama	Usher	Nas	Gucci Mane	Kaskade	72
7	The Game	Ludacris	Snoop Dogg	Ludacris	Beenie Man	Snoop Dogg	83
8	2 Chainz	DJ Khaled	Busta Rhy.	Usher	Snoop Dogg	The Game	77
9	DJ Drama	Jeezy	Gucci Mane	Rick Ross	Nas	Beenie Man	60
10	Nas	Snoop Dogg	The Game	Nicki Minaj	2 Chainz	Nicki Minaj	91

Table 2: Centrality measure rankings in undirected version of Graph 4

#	Eigenvector	Closeness	Betweenness	Pagerank	HITS (Hub)	HITS (Auth)	Aggregate	Spotify
1	Gucci Mane	Lil Wayne	Kaskade	Kaskade	Lil Wayne	DJ Drama	T.I.	80
2	DJ Drama	Rick Ross	T.I.	T.I.	Rick Ross	Gucci Mane	Rick Ross	82
3	DJ Khaled	Snoop Dogg	Lil Wayne	Gucci Mane	T.I.	DJ Khaled	Lil Wayne	90
4	The Game	Busta Rhymes	Rick Ross	Ludacris	Snoop Dogg	The Game	Gucci Mane	89
5	Ludacris	2 Chainz	Beenie Man	Cam'Ron	2 Chainz	Rick Ross	DJ Drama	67
6	T.I.	T.I.	Snoop Dogg	DJ Drama	Jeezy	Ludacris	Ludacris	77
7	Rick Ross	Nas	Nicki Minaj	JAY Z	Bun B	Trae Tha...	Snoop Dogg	83
8	Trae Tha...	Kanye West	Ludacris	Lil Wayne	Nas	T.I.	Kaskade	72
9	French Mon.	Nicki Minaj	Nas	Beenie Man	Wiz Khalifa	French Mon.	DJ Khaled	79
10	Yo Gotti	Pharrell	JAY Z	Rihanna	Jadakiss	Nelly	The Game	77

Table 3: Centrality measure rankings in directed version of Graph 4

model. In those graphs, the expected clustering coefficient would be:

$$C = c/(n-1)$$

where c is the expected number of collaborations and n is the number of artists in the graph. Taking Graph 4 as our example, the expected (average) number of collaborations is 14.26, so the clustering coefficient would be only 0.021. In reality however, the clustering coefficient is much higher—0.358. This is significantly lower than the clustering coefficient of a regular graph (which in this case would be 0.693). The clustering coefficient value we calculate reveals information about the graph topology, namely the measure to which the nodes in the graph tend to cluster together. In many other real world networks, nodes tend group tightly together—more so than the average probability of randomly assigned edges.

6 Clique Analysis

Another question we might ask is: are there cliques present in the musical collaboration graph? If so, how large do they get? We define a clique as a fully-connected subgraph—in other words, a group of artists who have each all collaborated with every other artist in the group.

We performed a search for maximal cliques on Graph 1 and found many, including several considerably large ones. The maximum size we discovered in our network was 10. Below we list the cliques found:

Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, T.I., Ludacris, Future, Usher
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, T.I., Ludacris, Future, The Game
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, Nicki Minaj, Yo Gotti, French Montana, 2 Chainz
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, Nicki Minaj, Yo Gotti, French Montana, Jadakiss
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, The Game, Yo Gotti, 2 Chainz, Future
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Gucci Mane, The Game, Yo Gotti, 2 Chainz, French Montana
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Nas, Nicki Minaj, Chris Brown, Usher, Ludacris
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Chris Brown, T.I., Ludacris, Future, Usher
Kanye West, Rick Ross, Lil Wayne, DJ Khaled, Jeezy, Chris Brown, T.I., Ludacris, Future, The Game
DJ Khaled, Lil Wayne, Rick Ross, Gucci Mane, Nicki Minaj, French Montana, Yo Gotti, Wale, 2 Chainz, Wiz Khalifa
DJ Khaled, Lil Wayne, Rick Ross, Gucci Mane, Nicki Minaj, French Montana, Tyga, Wiz Khalifa, 2 Chainz, Wale
DJ Khaled, Lil Wayne, Rick Ross, Gucci Mane, The Game, French Montana, Wale, 2 Chainz, Wiz Khalifa, Yo Gotti
DJ Khaled, Lil Wayne, Rick Ross, Gucci Mane, The Game, French Montana, Wale, 2 Chainz, Wiz Khalifa, Tyga
DJ Khaled, Lil Wayne, Rick Ross, Chris Brown, Wale, Tyga, 2 Chainz, Wiz Khalifa, French Montana, Nicki Minaj
DJ Khaled, Lil Wayne, Rick Ross, Chris Brown, Wale, Tyga, 2 Chainz, Wiz Khalifa, French Montana, The Game

Most of the artists appear in multiple size 10 cliques. Some, as in the case of DJ Khaled, Rick Ross, and Lil Wayne, are present in all. Why might this be? For one, these artists are a bit older than many of the other artists, giving them an obvious advantage—they have released more tracks and had more time to work on collaborations with other artists. Additionally, these three have recently moved towards producer roles, where their job is almost *solely* to collaborate with other artists. This position creates a cycle, where artists like DJ Khaled and Rick Ross, who are already influential on their own, become more influential as a result of working with other influential artists.

7 Visualization Techniques

While graph analysis methods can produce useful quantitative measures of importance, graphical visualizations can sometimes provide a better qualitative understanding of the graph's

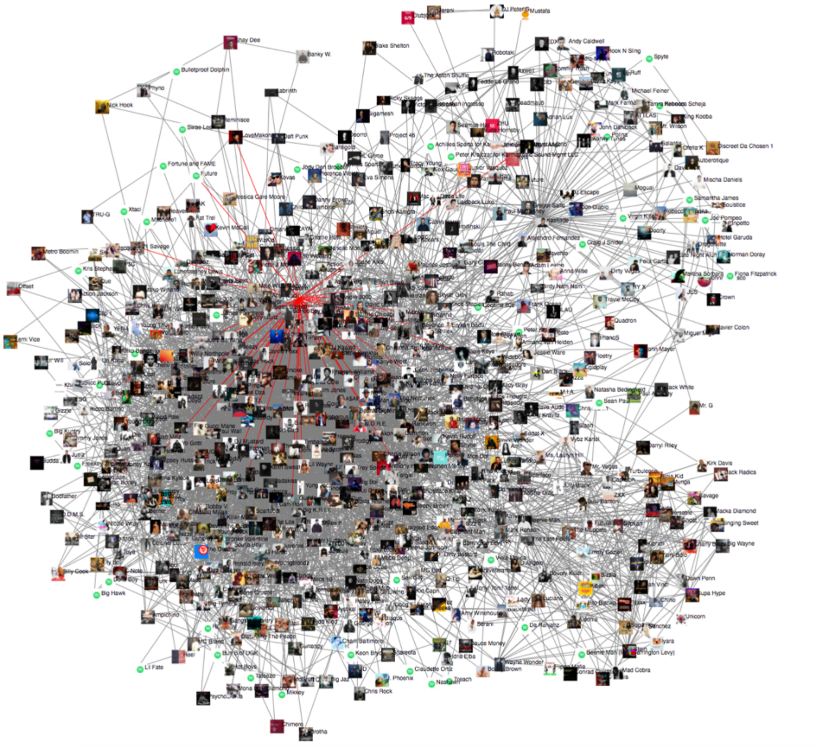

Figure 3: A screenshot of the force simulation of Graph 4.

structure. We created a basic graphical visualization using the d3js Javascript graphical library. Specifically, D3's powerful force simulation library permits a dynamic, interactive visualization of the connections between nodes.

In each step of the visualization, the nodes of the graphs are drawn using the Spotify profile pictures for the corresponding artists and labeled with the artist's name. Graph edges are represented as lines connecting the corresponding artists. In addition, the visualization can detect mouse events and highlight the node the user is mousing over and any outgoing edges by changing the color of the relevant links.

In total, three separate types of forces were used in the dynamic force simulation. The first was attractive force between nodes connected by a graph edge. For a given pair of nodes i and j, this force is modeled with a Hooke's Law relation $F_{ij} = -k_{ij} * d(i, j)$, where $d(i, j)$ represents the distance between the nodes i and j. The stiffness k_{ij} is proportional to the number of collaborations between the artists corresponding to nodes i and j in the Spotify collaboration graph.

In opposition to this attractive force is a general repulsive force between all nodes. This is implemented with Coulomb repulsion, which is an inverse square repulsive force with magnitude $C(i) = k_c^2/d(i, j)^2$ for all $j \neq i$. The constant strength k_c is the same for all nodes and is tuned so that nodes are fairly uniformly spread across the visualization canvas.

The final force is a "gravitational force" that lightly pulls all nodes toward the center of the canvas. This is modeled with an attractive force with magnitude $G(i) = k_g d(i, (x_0, y_0))$ where the center of the canvas is located at coordinates (x_0, y_0). This force has a much smaller magnitude than the other forces, and is only present to ensure that the graph visualization recenters itself as it reaches a stable equilibrium configuration. Without this force, the graph would tend to drift off to one side of the canvas, which is less visually appealing.

A screenshot of the force simulation is given in Figure 3.

8 Future Work

There are a number of new, interesting directions this project could be taken. We could:

1. Increase the number of BFS iterations to get a larger snapshot of the two million artists
2. Create a weighted graph and take into account the number of times artists collaborate with each other
3. Try generating the graph without BFS by using Spotify's list of top 1,000 artists
4. Incorporate song streams/artist popularity to predict upcoming artists
5. Incorporate Spotify user behavior to make personalized recommendations to Spotify users

9 Project Links

The code used to generate the collaboration graphs and all the data used is available on GitHub at: `https://github.com/akashlevy/Musical-Collab-Graph`. The force simulation is available at `https://akashlevy.github.io/Musical-Collab-Graph/index.html`.

10 Reference Links

1. Spotify: `https://www.spotify.com/`
2. PageRank: `http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf`
3. Kleinberg HITS Algorithm: `http://ranger.uta.edu/~chqding/cse6319/classPapers/Kleinberg-web-graph.pdf`
4. NetworkX: `https://networkx.github.io/`
5. Spotipy: `https://spotipy.readthedocs.io/`
6. Gephi: `https://gephi.org/`
7. D3JS: `https://d3js.org/`