Sound vs Song: Exploratory Song Analysis in Neo4J

DS4300 Project Final Report

project by: Alicia Wheeler, Daniel Rossshirt, Tatum Gaudinski Whitehead, Jialin Zhen

Abstract

This project leverages the unique graphing capabilities of Neo4J to perform an exploratory analysis on statistical differences between the categorical and quantitative attributes of songs. The goal is to examine trends and relationships between songs to further understand the music industry and the structure of songs. The examination is performed in a Neo4J network graph. Queries on the graph revealed that songs distinguished by decade and genre still have similar attributes to other genres and decades; there are consistency between-song attributes across the many genres/decades in the industry.

Process

Initially, our group desired to create an engine to recommend songs to users. That is cliche and had already been done so we changed courses. We began our project using MongoDB but quickly decided that Neo4J would better suit our needs. Instead of trying to implement our own version of the *Spotify "Go to Song Radio" feature* as discussed in the project proposal, we pursued an exploration of attribute similarity across genre and decade.

There are vast numbers of unique sounding songs. The purpose of this exploration is to understand features that differentiate songs from one another and explore possible connections between genres and decades of songs.

Design

Data Collection

The data used is a CSV file. Data were obtained from the Spotify API and contains sixteen different attributes for a collection of 41,000 + songs. The dataset is stored in a GitHub repository.

Data Wrangling

The data is imported from the GitHub repository into a Jupyter notebook for preprocessing. Four of the sixteen attributes — tempo, danceability, acousticness, and energy — are selected from the dataset. The song name, genre, and decade are selected from the original data as well. The *Pandas* library is used to store and manipulate data as a DataFrame. The tempo was normalized using a min_max_scaler to the range [0, 1]. One hundred songs are selected at random for further analysis. The Euclidean distance between the four selected quantitative attributes for each song is computed. Each pair of songs will represent a link in the graph. The final export of the data is two CSV files: one which contains the nodes (100 songs with attributes) and one which contains the links (pairs of songs with the euclidean distance between the pair).

Data Import

Both CSV files are imported into MySQL using the MySQL Import Wizard. In MySQL, a schema called *songs* is created. In the schema, the two datasets are stored as nodes and links.

The data is imported into Neo4J using the APOC API for MySQL. In MySQL, a server *song* is created with the newly created *song* schema. Two views are created to import data into Neo4J. One view is for the nodes table and consists of 100 songs. The second view is for the links table. The 20th percentile of links is selected. Thus, the view is queried to import only 1000 links. This is done to reduce the size of the graph to avoid occlusion.

Querying in Neo4J

The focus of the queries is to better understand the music industry via the relationship between a song's derived and categorical attributes. The queries in Neo4J are focused on producing significant visual representations. The data does require intricately designed queries. By writing direct queries we are able to get the most information from the graph.

Disabling the "connect result nodes" function in the graph visualization setup of Neo4J is very helpful in creating straightforward and efficient queries.

Results & Analysis

The analysis compares euclidean distance to identify patterns between songs with low euclidean distances using categorical attributes. Three thresholds for euclidean distance are queried to compare the songs:

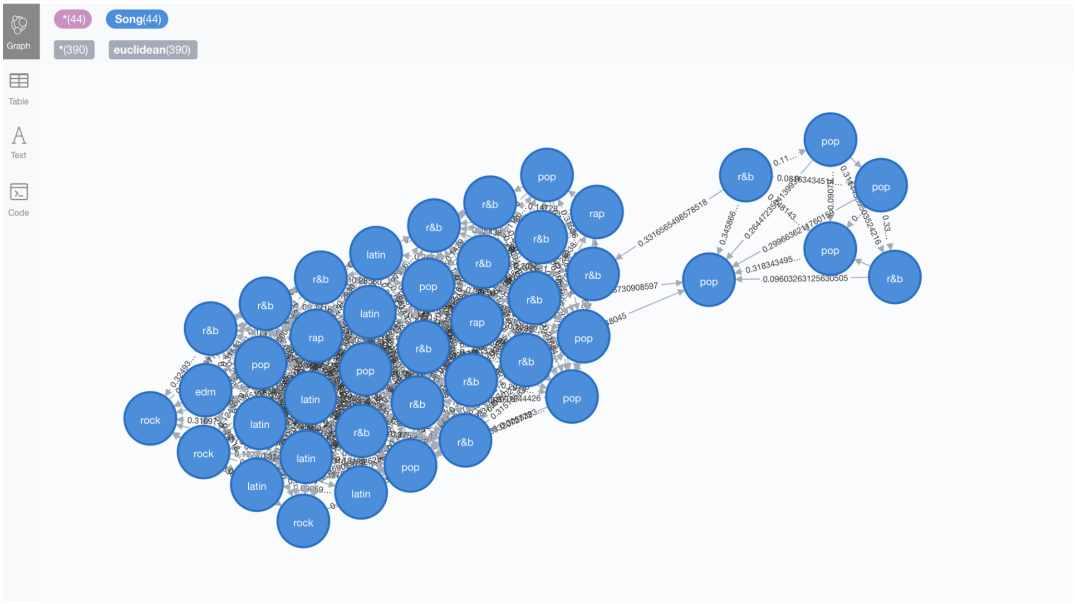
- n.euclidean < 0.121576 *1st percentile*
- n.euclidean < 0.206442 *5th percentile*
- n.euclidean < 0.259790 *10th percentile*

Thresholds are selected from percentile values obtained from a call to the .describe() method on *Pandas* DataFrame in the preprocessing portion of the project. An in-depth report of categorical attributes queried at percentile thresholds can be found here. Queries to the Neo4J graph produced visualizations of songs clustered by links that met the specified euclidean threshold. Links between categorical attributes genre and decade are examined for each threshold.

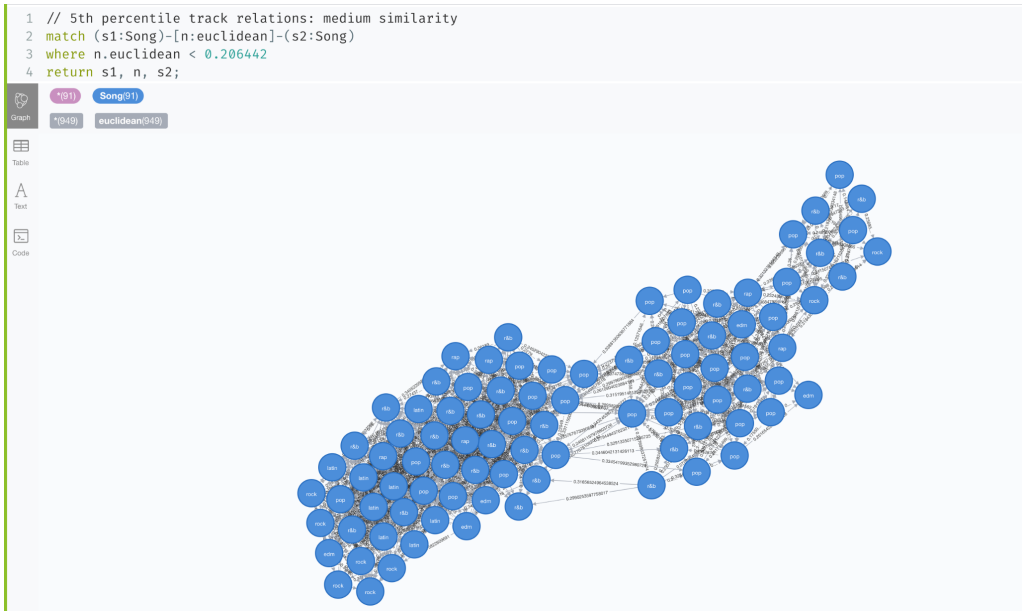
These values are used to see songs with distances below or equal to each threshold value.

Distances in the left tail of the first percentile form clusters of predominantly R&B, Rap, and Pop. Additionally, there are clusters of R&B, Rap, Pop, and Latin and another cluster of EDM, Rock, and R&B.

```
// 1th percentile track relations: highest similarity
match (s1:Song)-[n:euclidean]-(s2:Song)
where n.euclidean < 0.121576
return s1, n, s2;
```



The most prevalent decades are the 80s, 90s, 00s, and 10s. Increasing the distance in the left tail to the fifth percentile, the previously scattered groups merged into two distinct clusters. One cluster consists of R&B and Pop, while the other is R&B, Latin, Pop, and Rap. The 80s, 90s, 00s, and 10s are still the most prevalent decades.

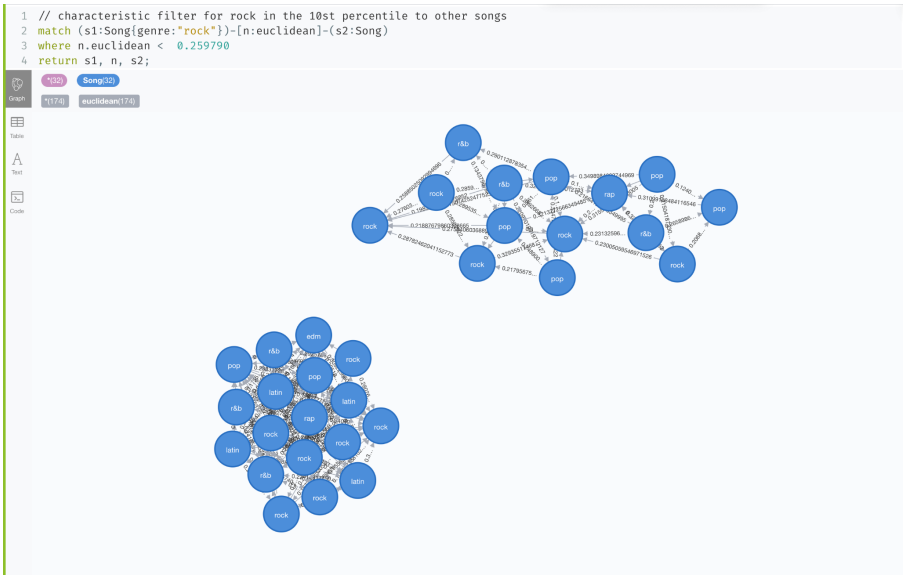


The final distribution showed the previous clusters linking together to show one large network with three distinct groupings. The first section consisted of Pop, R&B, and Rock songs with decades from the 70s to 00s. The second section consisted of Pop and R&B songs with decades from the 60s and 80s. The third section consisted of Latin, R&B, Pop, Rock, and some other genres of songs with decade distribution from the 70s to 10s.

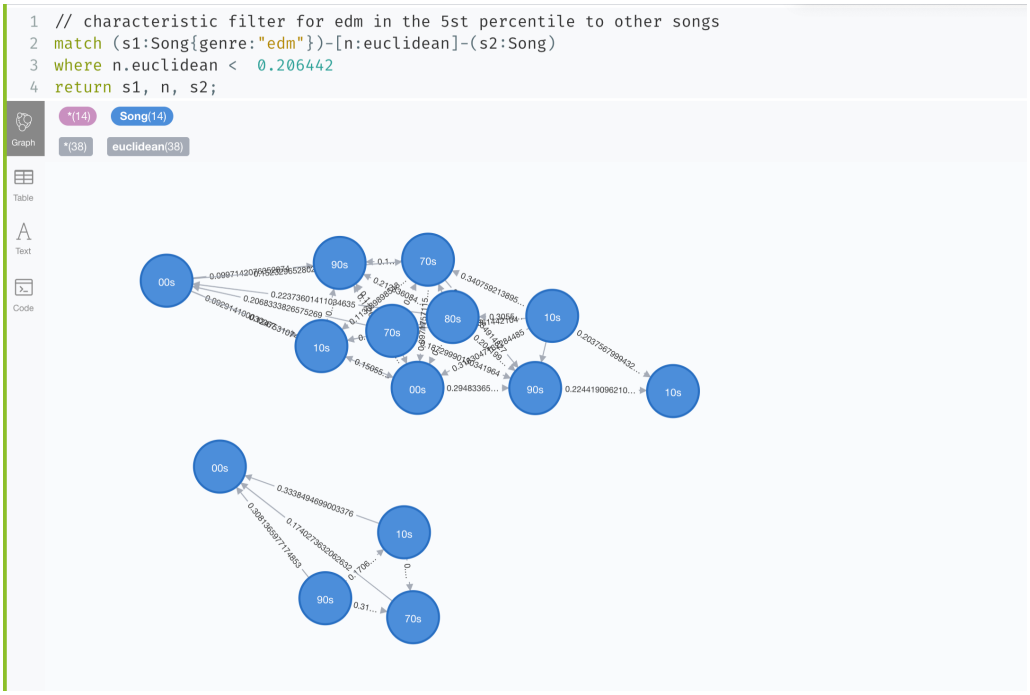
The second section of the analysis filters by song genre in addition to Euclidean distance. Genre-specific song connections are examined for each threshold value to identify a possible similarity between songs. The graph is then modified to display decades. This is to see if specific genres had prevalent connections to other genres and to examine the effect of the decade on the relationship between songs.

The first genre looked at was R&B. Connections between R&B songs and Pop and Latin songs are most prevalent. The majority of songs linked to the R&B genre are from the decades of 80s to the 00s. The high occurrence of these song genres with small Euclidean distances shows an overlap between attribute values in those genres.

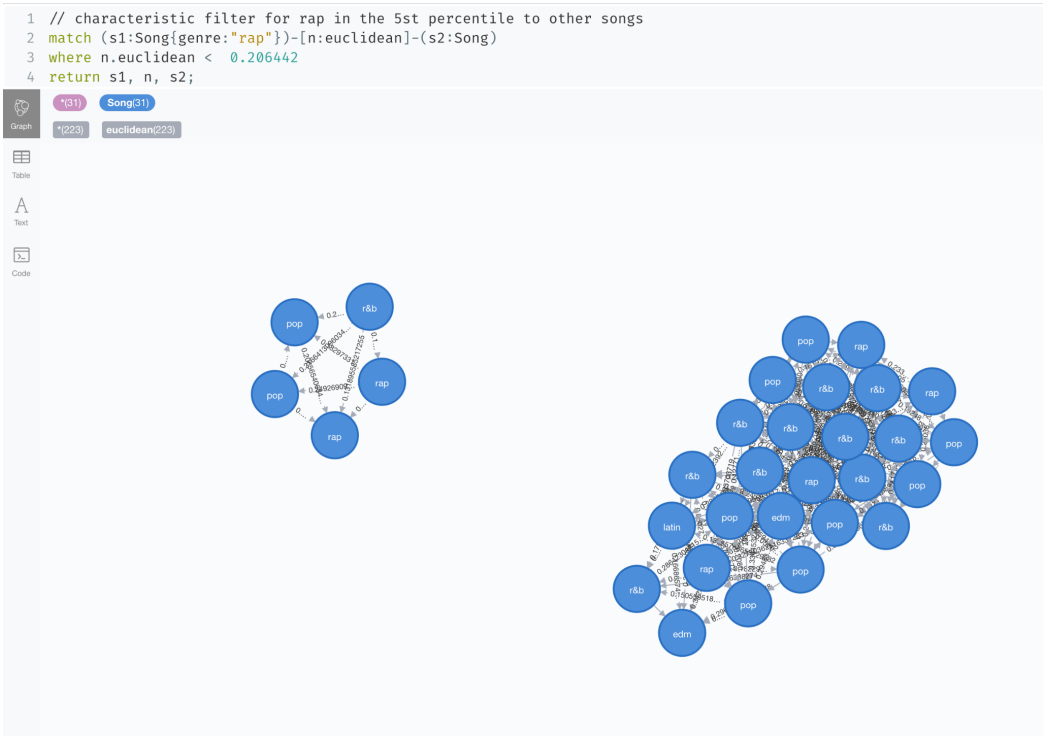
The next genre is Rock. There are not many song genres connected to Rock. There were some Latin and Pop songs. A close relationship between EDM and Rock songs can be observed. The recurring decade for these connections was the 70s, 00s, and 10s showing to be the most influential decades.



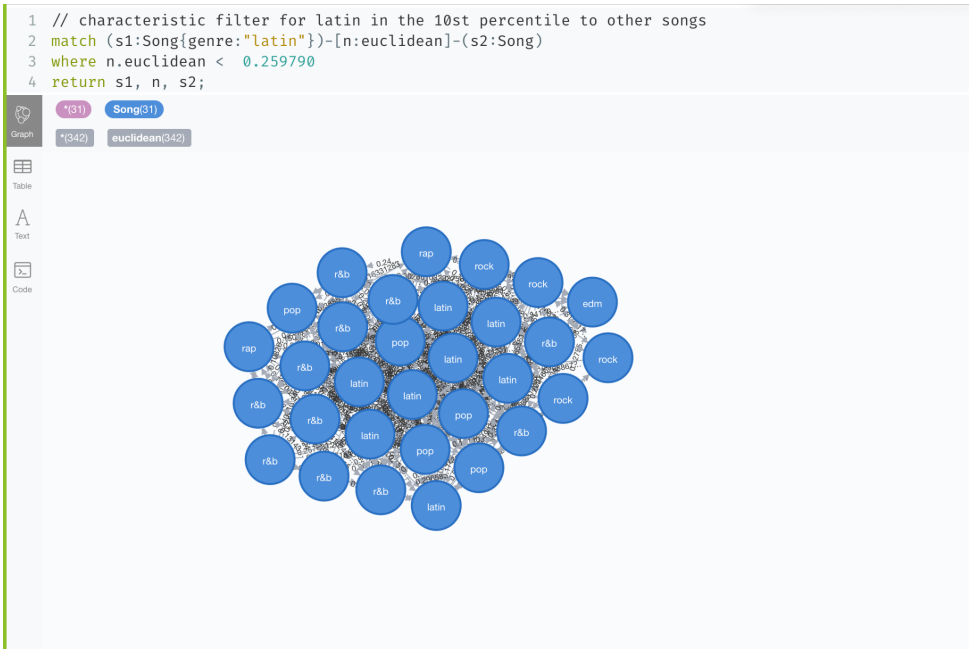
Rap, Pop, Latin, and Rock songs have the highest number of connections to the EDM genre. The 90s decade was very prevalent in this genre network. R&B and Pop songs are highly connected to Rap songs, with some connection to EDM and Latin songs. The prevalent songs were from the 80s, 90s, and 10s decade.



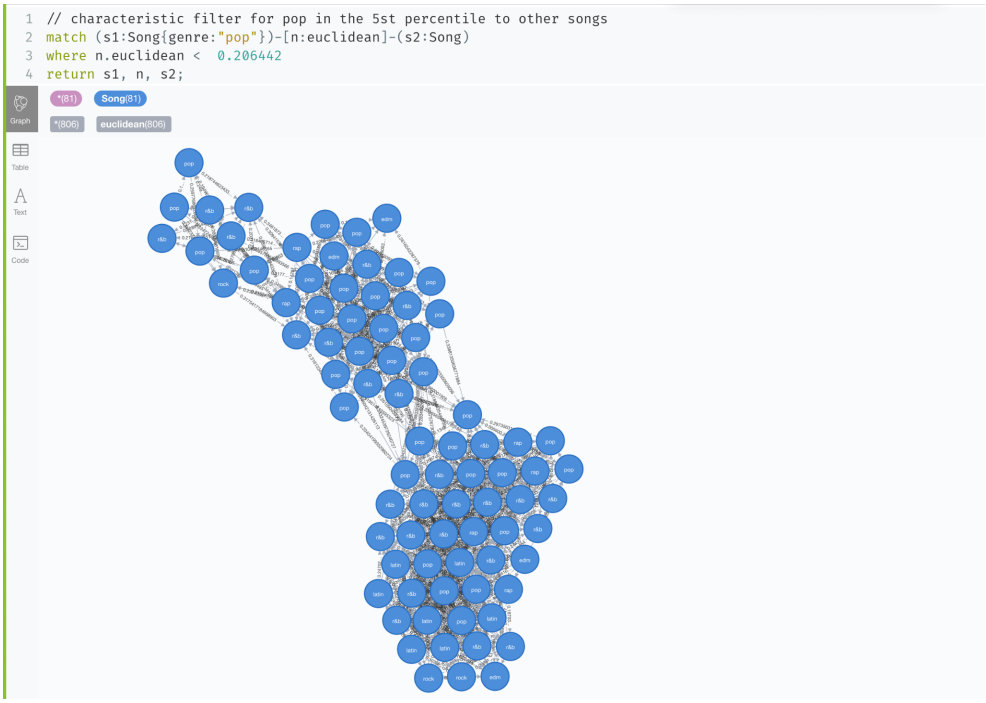
The Rap genre songs were highly linked with R&B and Pop songs from the 80s, 90s, and 10s.



The Latin genre songs were clustered with songs in the genres of R&B and Pop with frequent 70s, 80s, and 00s decade occurrences in the connections.



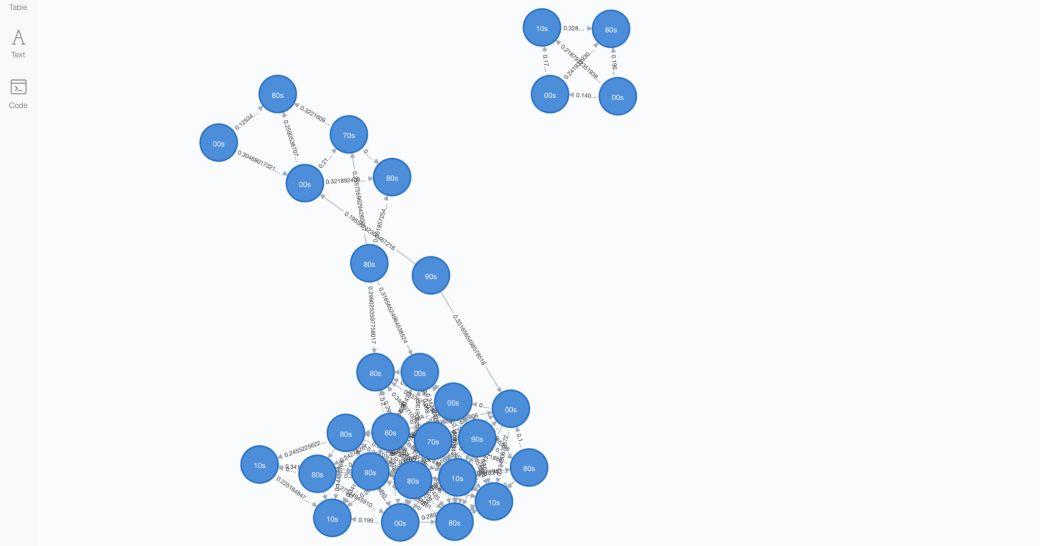
The Pop genre has many connections with R&B songs and fewer connections to Rap, Latin, Rock, and EDM. The decade connection for the Pop genre contains a prevalent grouping of the 70s, 80s, and 00s in one section. This section is then connected to another cluster of 60s, 80s, and 00s songs.



The following analysis compares songs of the same genre for clustering/significant connection between decades of songs in the same genre:

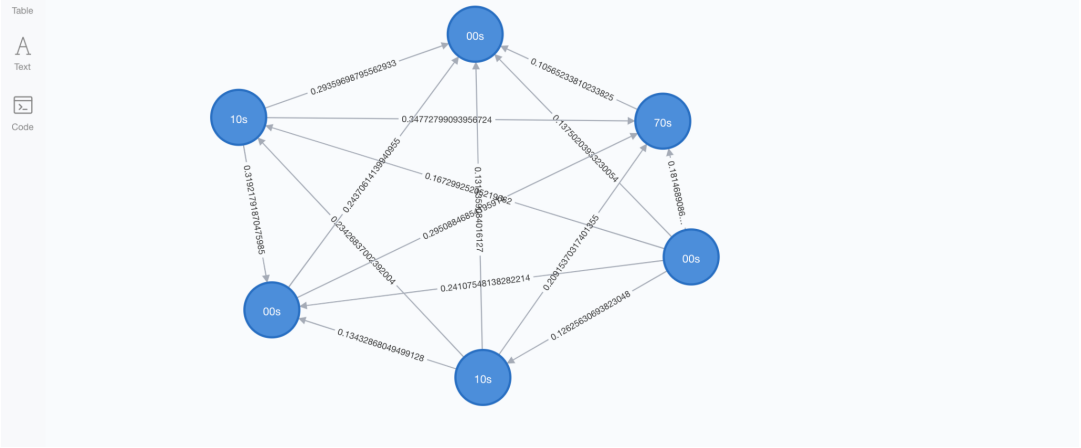
For the R&B genre, the majority of the songs with high connectivity are from the 80s and 00s.

```
// characteristic filter for r&b in the 5st percentile to other r&b songs
match (s1:Song{genre:"r&b"})-[n:euclidean]-(s2:Song{genre:"r&b"})
where n.euclidean <  0.206442
return s1, n, s2;
```



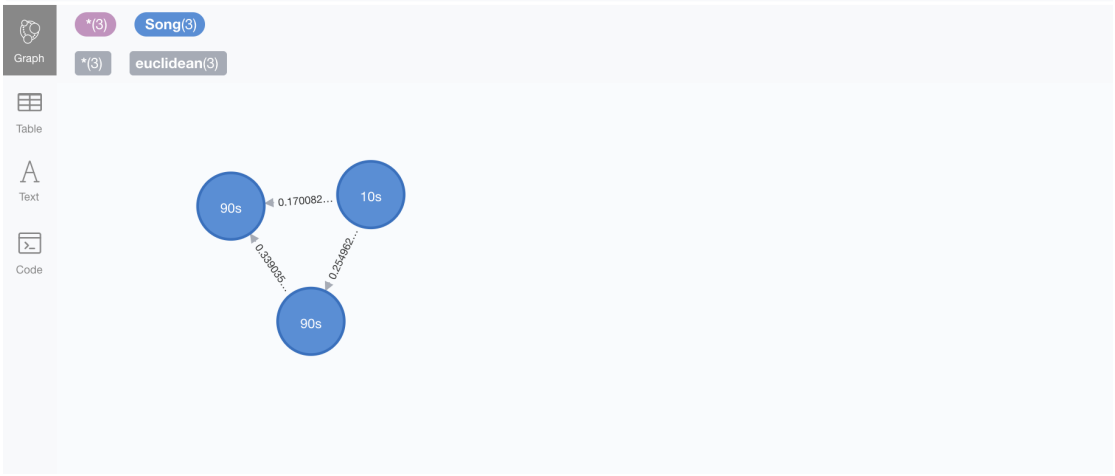
The Rock genre has connections throughout the 70s, 00s, and 10s. Similar to the R&B songs, the high connectivity between decades may indicate influence.

```
// characteristic filter for rock in the 5st percentile to other rock songs
match (s1:Song{genre:"rock"})-[n:euclidean]-(s2:Song{genre:"rock"})
where n.euclidean <  0.206442
return s1, n, s2;
```



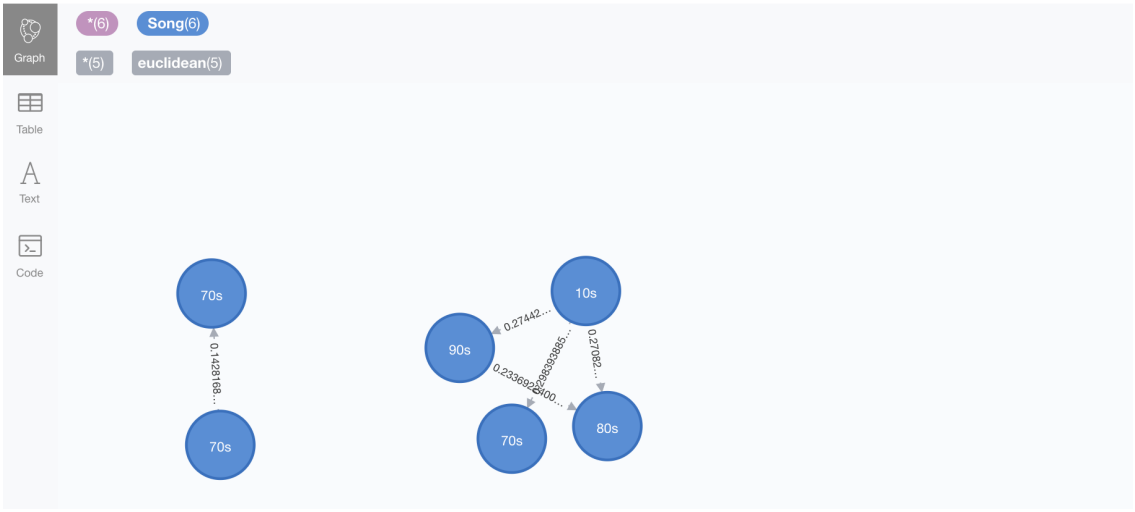
The EDM genre showed that there were not many EDM songs in the dataset. There is one song from the 10s that had connections to two 90s songs.

```
// characteristic filter for edm in the 10st percentile to other edm songs
match (s1:Song{genre:"edm"})-[n:euclidean]-(s2:Song{genre:"edm"})
where n.euclidean <  0.259790
return s1, n, s2;
```



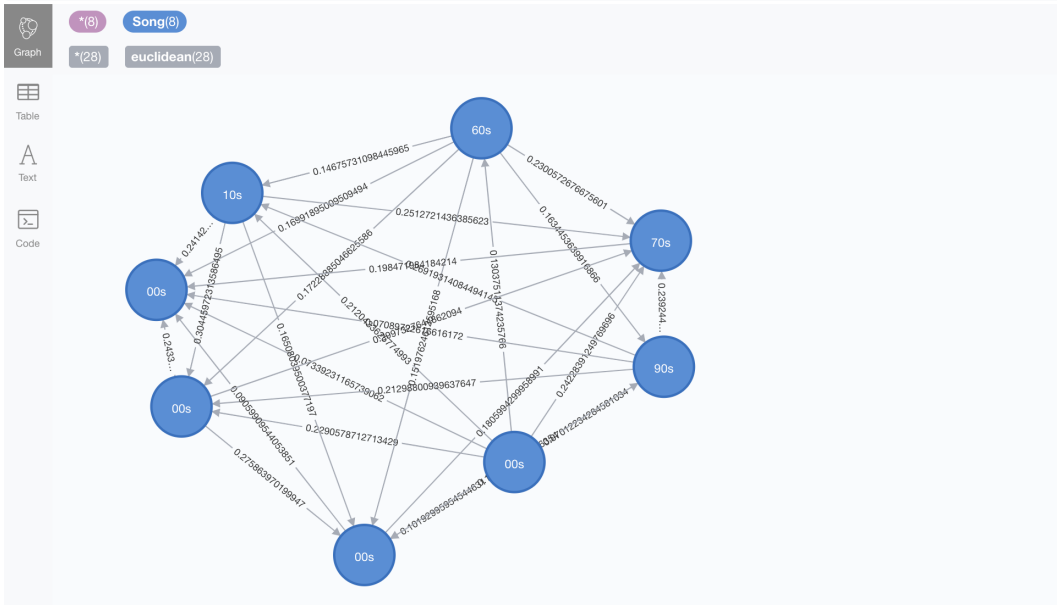
The Rap genre had limited connection between other Rap songs. The Rap songs were from the 70s and 80s. Though there is a limited amount of nodes, the songs are highly linked to each other.

```
// characteristic filter for rap in the 10st percentile to other rap songs
match (s1:Song{genre:"rap"})-[n:euclidean]-(s2:Song{genre:"rap"})
return s1, n, s2;
```



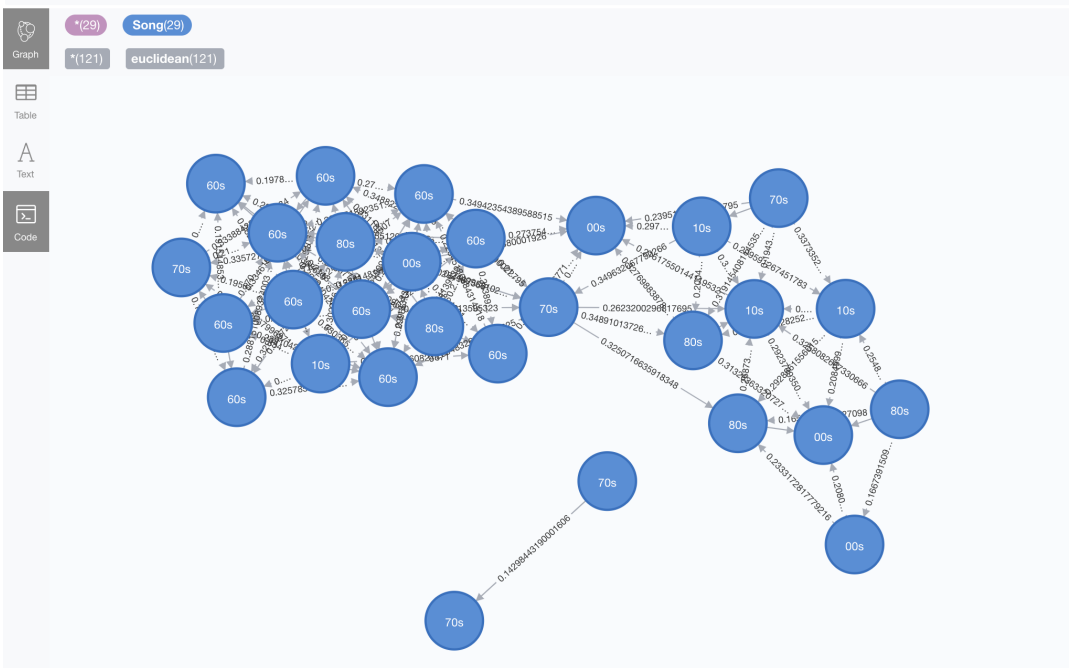
The Latin genre had high connections to Latin songs through the 60s, 90s, and 00s.

```
// characteristic filter for latin in the 10st percentile to other latin songs
match (s1:Song{genre:"latin"})-[n:euclidean]-(s2:Song{genre:"latin"})
where n.euclidean <  0.259790
return s1, n, s2;
```



The Pop genre had the highest number of songs. Songs are connected to each other through the decades of the 60s to 80s. There are also high connections of songs from the 80s to the 00s.

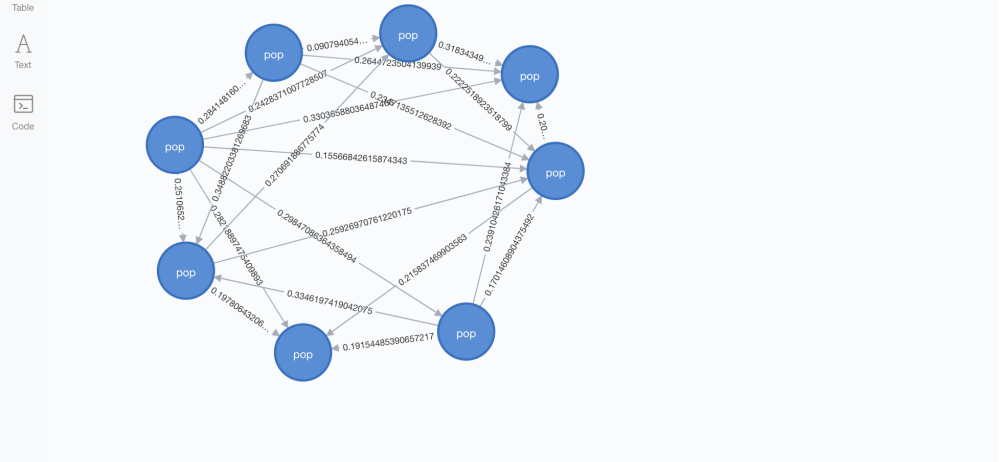
```
// characteristic filter for pop in the 5st percentile to other pop songs
match (s1:Song{genre:"pop"})-[n:euclidean]-(s2:Song{genre:"pop"})
where n.euclidean <  0.206442
return s1, n, s2;

```



Another analysis was conducted focusing on comparing links between songs from the same decade in addition to the Euclidean distance.

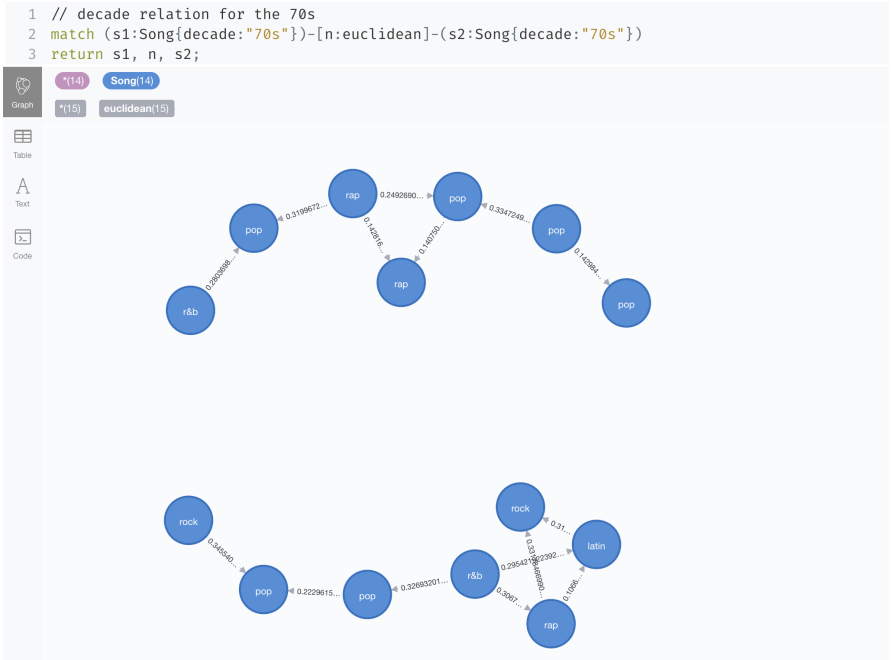
It can be seen that for songs from the 60s, the most similar songs are pop and rock songs. The artists for these songs were bands and solo artists.

```
// 5th percentile decade relation for the 60s
match (s1:Song{decade:"60s"})-[n:euclidean]-(s2:Song{decade:"60s"})
where n.euclidean < 0.206442
return s1, n, s2;
```



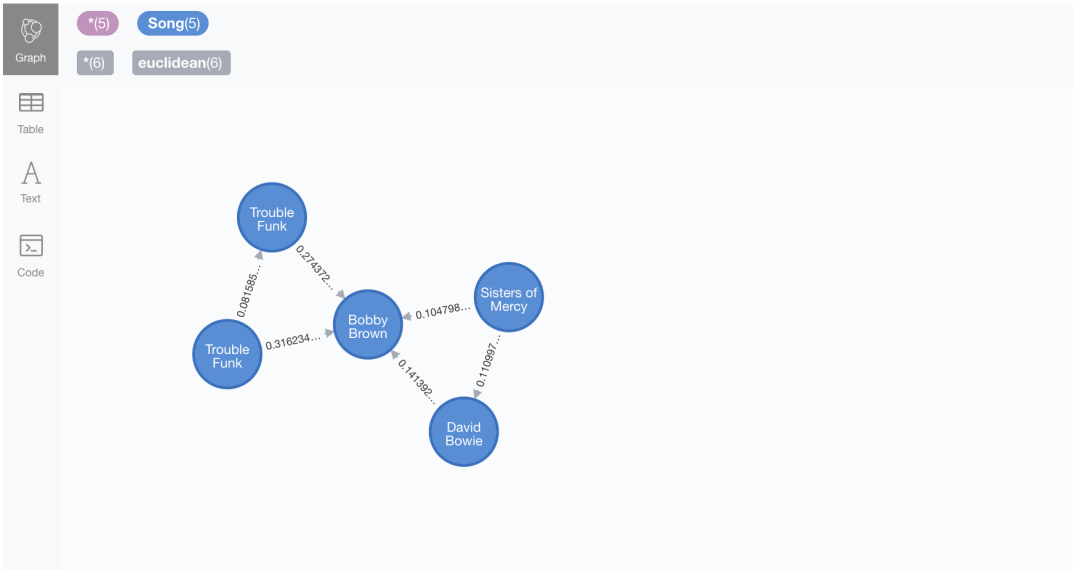
For songs from the 70s, there are two distinct networks. One network consists of Pop and Rap songs. The second network consists of Rock, Pop, R&B, and Latin songs. The majority of the artists are solo artists with a few bands. This shows that during the 70s, Rap and Pop songs were created using similar attribute levels even though the songs are categorized in different genres.

```
// decade relation for the 70s
match (s1:Song{decade:"70s"})-[n:euclidean]-(s2:Song{decade:"70s"})
return s1, n, s2;
```



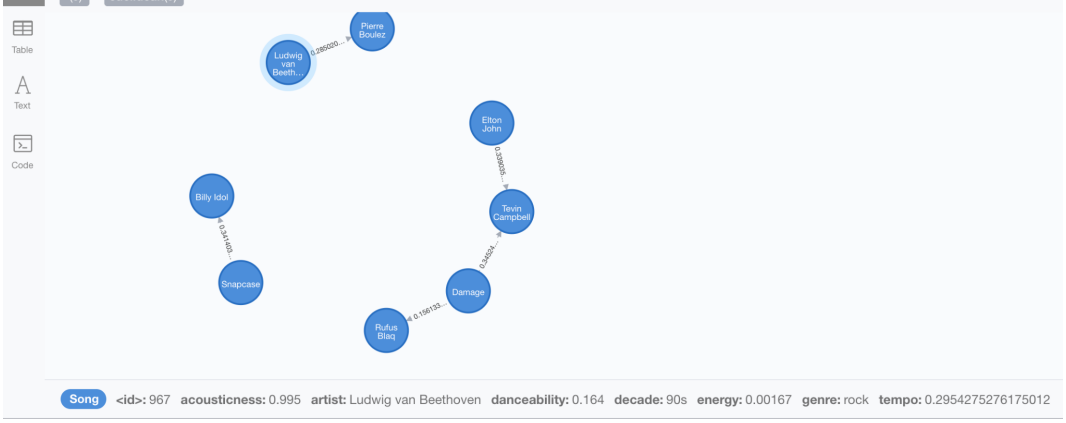
The 80s consisted mainly of R&B and Pop songs. The network has Rap songs connected but Rap songs are not as highly connected as R&B and Pop songs. An interesting insight is that an artist in the 80s created two songs that are categorized in two different genres but their quantitative attributes are very similar.

```
1  // 1th percentile decade relation for the 80s
2  match (s1:Song{decade:"80s"})-[n:euclidean]-(s2:Song{decade:"80s"})
3  where n.euclidean < 0.121576
4  return s1, n, s2;
5
```



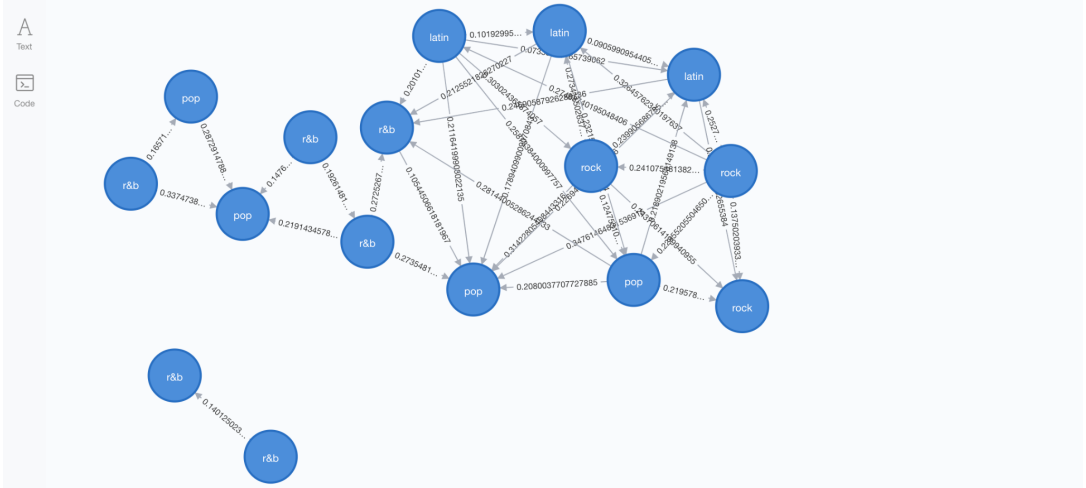
The 90s were not as interconnected as previous decades. The only significant insight came from when there was no Euclidean distance filter. This showed connections between two Rock songs, EDM and Latin, and Rap, R&B, and EDM songs. An interesting insight from this network is that classical songs by artists such as Beethoven are classified as rock songs, which can indicate a flaw in the dataset and how the genre is being categorized in Spotify.

```
1  // decade relation for the 90s
2  match (s1:Song{decade:"90s"})-[n:euclidean]-(s2:Song{decade:"90s"})
3  return s1, n, s2;
4
```



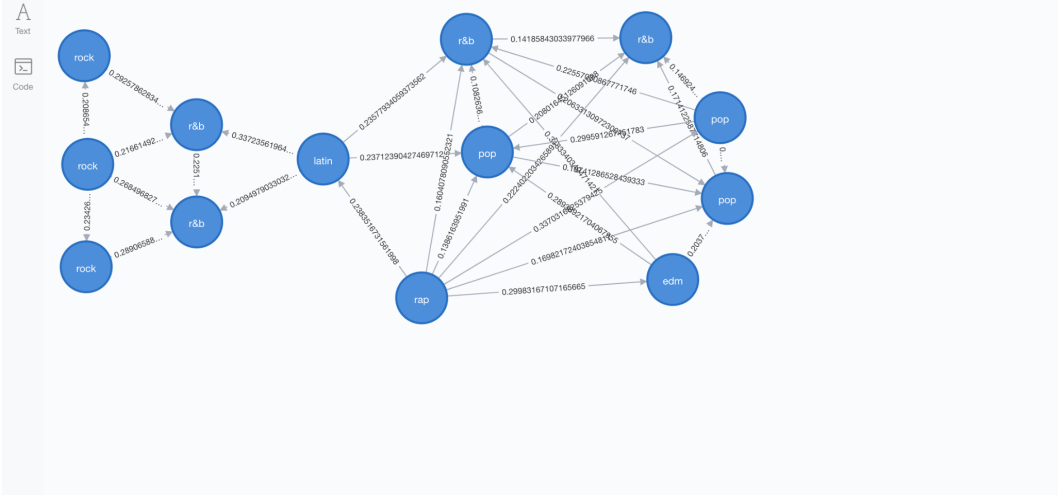
The 00s consist mainly of Latin, Pop, and R&B songs which are all interconnected with each other. The highest connected songs were two Latin songs and two Pop songs with each having nine connections. This network shows that the songs from this decade have similar attributes.

```
// 5th percentile decade relation for the 00s
match (s1:Song{decade:"00s"})-[n:euclidean]-(s2:Song{decade:"00s"})
where n.euclidean < 0.206442
return s1, n, s2;
```

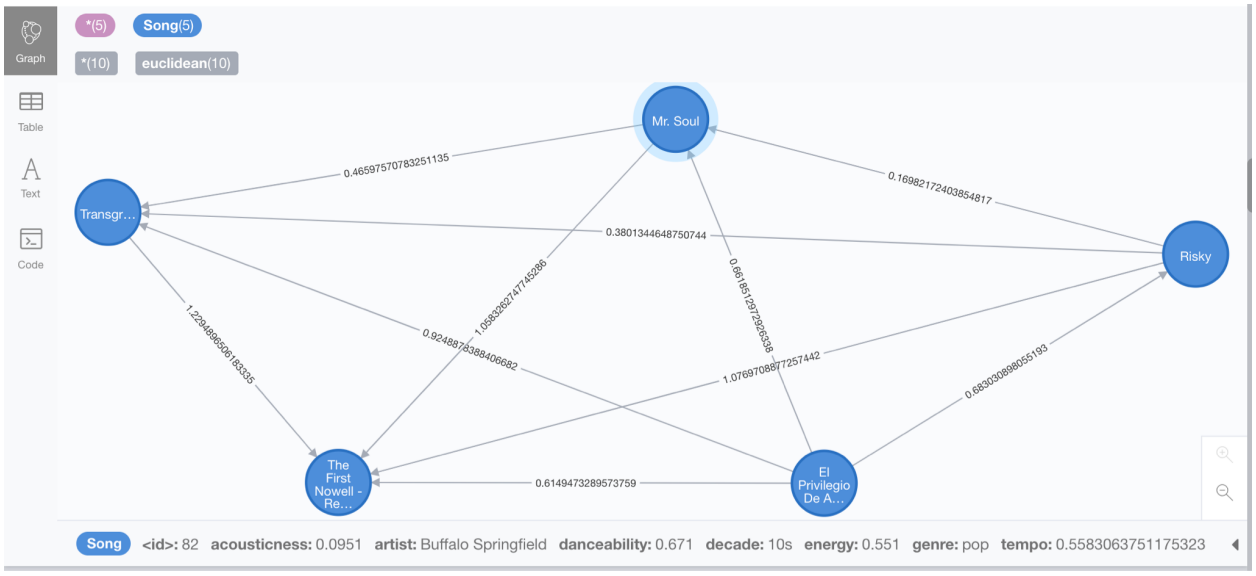


The 10s consists of many Pop and R&B song connections and Rock and R&B song connections. The side of the network with the high Rock and R&B connections consisted of band artists, while the other side consists of mainly solo artists.

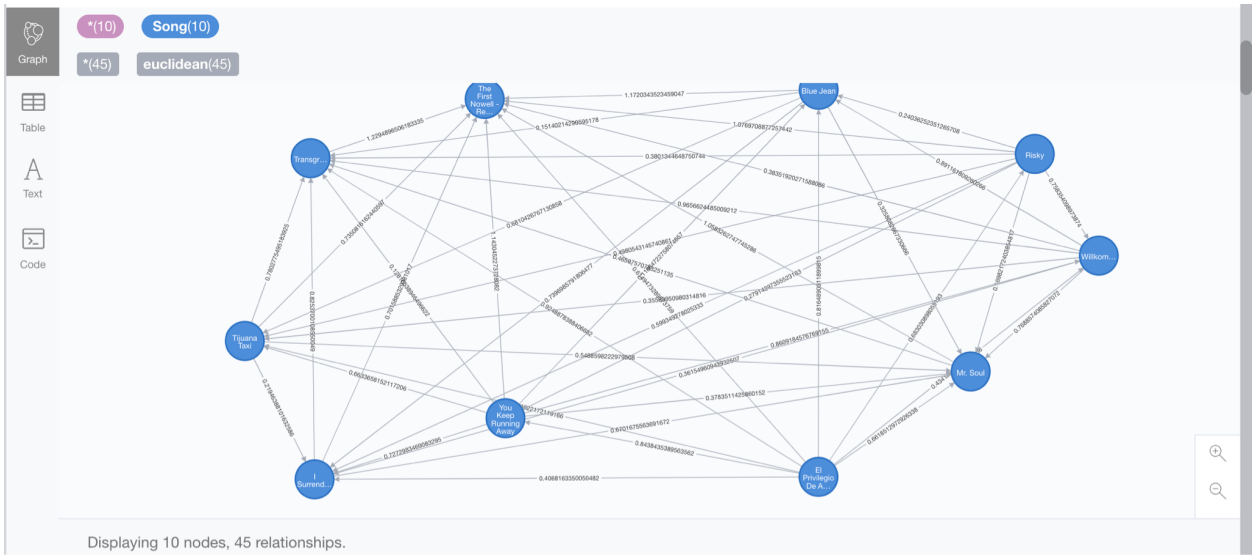
```
// 10th percentile decade relation for the 10s
match (s1:Song{decade:"10s"})-[n:euclidean]-(s2:Song{decade:"10s"})
where n.euclidean < 0.259790
return s1, n, s2;
```



Influential songs are seen to have the greatest influence on songs that also become influential. This suggests that either there are quantitative attributes that can be attributed to a song's success, or that humans just enjoy similar sounds. A sharp decrease in song influence outside of the five most influential songs is also observed.



The 5 songs with the most connections and the 10 connections that they share with each other.



The 10 songs which have the most connections and the 45 connections between each other

Conclusion/Significance

The project reveals that song similarity is not specific to a singular genre or decade. Despite stark differences in sound, the attributes analyzed in the project can be quite similar between songs. Even though songs were created in different decades and belong to different genres, songs at their core can still be very similar to each other. This makes sense. There are features present in all songs. Decade differentiation between genres shows the influence songs from different decades have on each other and highlights the fluidity of music over time. It can be seen that older songs have high relations to contemporary songs. Past songs have set a standard for distinct attribute levels which has carried throughout the different decades. This is a common issue in the music industry. It is not always possible to classify a song as one particular genre; discrepancy often arises. It can be seen in our project that song attributes may be very similar and yet the songs may be classified as different genres. Graph analysis has also revealed commonality between songs based on artists as opposed to genre. The network analysis revealed that an artists' "sound", as determined by attribute similarity, can be consistent without songs belonging to the same genre.

The music industry is always evolving. This exploration reveals that despite nuances in sounds, there remains similarity amongst core attributes of songs. It suggests that there are features that are integral to creating songs and perhaps in differentiating sound from the song.

Further Exploration

- It would be interesting to perform the same queries on Euclidean Distances computed using all available song attributes.
- It would also be interesting to have more data about the music industry such as producers, writers, singers, etc. An examination of connections between these additional attributes may highlight new connections between seemingly unrelated aspects of the industry.
- A similar analysis on songs of non-western origin would also be interesting. The core similarities discovered in Western songs may no longer be present, suggesting a fundamental difference between musical spheres. Conversely, if the same similarities are present in both spheres, it may lend information about what differentiates sound from the song.

Contributions

Alicia Wheeler: Data Wrangling, Data Preprocessing, Querying, Querying Analysis, Project Report
Daniel Rosshirt: Data Processing, Querying, Analysis, Project Report
Tatum Whitehead: Data Acquisition, Data Wrangling/Preprocessing, Slides Presentations, Project Report
Jialin Zhen: