CPSC 572/672: Final Project Report
Harmony in Data: Unraveling the Symphonies of Spotify Playlist in Summer 2016

Team - Chords and Code
Gustavo Bravo
Caleb Fedyshen
Kevin Phan

Project summary

The interconnections between human experiences and their music listening habits are arguably one of the most interesting ways of implicitly revealing one's personality. Due to the complexity of this relationship, it is difficult to study with contemporary approaches. However, with network science, an exploration of playlists through the 'Spotify Million Playlist Dataset made public by AiCrowd facilitates this endeavour. This paper seeks to examine the commonalities between songs that are grouped in playlists to understand the psychological factors that contribute to playlist construction and the proliferation of popular songs. A network model was constructed in which songs that occur together on playlists are linked, after which several analyses were conducted to explore communities present across playlists and factors that are correlated with the popularity of a song.

The network structure revealed a highly clustered network, where the communities revealed insights into the underlying organization of the network. The most significant influencers of playlist creation identified in this study were popular genres, eras of music, artists, and cultural and thematic associations. Furthermore, it was found that popular songs tended to be concentrated towards the beginning of playlists, and were of short duration. However, the complexity of these factors in shaping playlists cannot be understated. The study was limited by computing power, and thus the data sample was reduced to a week of playlists. Future work could reveal more about how networks of playlists evolved in recent times, reflecting and shaping the culture of the world.

Research questions

1. How can hierarchical clustering techniques be used to identify and characterize unique communities of songs? Various clustering techniques studied in the course will be attempted to find groups of songs with significant commonalities between them.
2. What insights can communities of songs provide about music genres, mood, music trends, or any other factors present in their similarities? Once the clusters are defined, the characteristics of songs and artists within them are to be studied.

3. Is there an explainable relationship between the position a song has on a playlist (this attribute is known as rank), and its popularity (degree)? Popularity will be defined and compared to the song's rank, then visualized on a scatter plot.
4. Are there any relationships between the track duration and their popularity? Do we find that shorter songs are more popular? Popularity will be defined and compared to the song's duration, then visualized on a scatter plot.

Introduction

Music plays a pivotal role in human culture and society, serving as a powerful medium for expression, communication, and connection across diverse communities.[1] It transcends linguistic and cultural barriers, evoking emotions, and facilitating social interactions. The centrality of Spotify in today's music consumption makes it an informative data source for understanding musical preferences and trends. Particularly, the user-generated playlists on Spotify offer a rich dataset for analyzing how songs are experienced and appreciated. Playlists are a unique way of expressing how an individual categorizes and organizes songs mentally.[2] This data, when viewed through the lens of network analysis, presents a deeper insight into the ways that music is interconnected, and the way people perceive those connections.

Playlists may be curated based on diverse criteria, possibly including similarities in preferences, suitability for specific activities or environments, or simply reflective of what one currently enjoys. This presents a gateway towards gaining a deeper understanding in uncovering how people construct the soundscapes of their lives. Based on the exploration of existing literature, our research aims to identify the gaps in understanding music tastes by examining the relationship between pieces of music across curated collections.

In this study, the structure and characteristics of a co-occurrence network, which is constructed from songs across Spotify playlists, are analyzed to find these gaps in musical connections. The underlying premise is rooted in the understanding that musical preferences and song popularity are not only determined by the attributes of the songs themselves but also by how these songs are associated by listeners. Several network science studies apply concepts to music datasets, for example, researchers have begun to apply network analysis to "study the topology of several music recommendation networks, which arise from relationships between artist(s) or co-occurrence of songs in playlists". Cano et al. (2006)[3] Although this study is quite preliminary and uses data from things like MSN and Amazon, which are not modern leaders in music distribution, it provides an interesting context on the human perception of similarity and grouping in music. Efforts

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3741536/
[2] https://journals.sagepub.com/doi/10.1177/2056305119847514
[3] https://arxiv.org/abs/physics/0512266

have also been made to use community detection to categorize genres in a much more restricted dataset where they selected ”2000 playlists from our dataset, and increased the threshold to only keeping artists that were in at least 10 playlists”[4] The data used was scraped from twitter feeds and only limited discussion of the possible genres was conducted.

This study seeks to aim further into the existing research by taking advantage of trusted data sourced directly from a leading entity that currently dominates the music distribution sector. Diving into greater depth and analyzing a substantially larger dataset compared to prior studies, and thereby facilitating a more comprehensive examination of the contemporary dynamics within the music industry.

Data Collection and Preprocessing

The dataset used within this research is obtained from AI Crowd[5] provided as part of their “Spotify Million Playlist Dataset Challenge”. The dataset contains 1,000,000 playlists, which were created by users on the Spotify platform, dating between January 2010 and October 2017. Details of the data provided for each playlist are the playlist’s name and the track titles. Further details related to the tracks themselves are the track’s position within the playlist, the artist’s name, the track’s name, the duration (in milliseconds), album names, and appropriate spotify URIs. Before constructing the network, the data was filtered to playlists created between August 1st, 2016 to August 7th, 2016. No additional work was needed on preprocessing the data as they are very well organized across all playlists.

Network Construction[6]

The network was constructed as a co-occurrence network, built from parsing the filtered data seen during the first week of August 2016. Nodes in the network represent the individual tracks, where each encapsulates attributes that include the artist’s name, track name, duration, album name, and position within associated playlists. Edges between nodes are created based on co-occurrence within playlists. If two tracks appear together in a playlist, an edge is created between them. This network should facilitate the analysis of track popularity and interconnections, providing insights into the relational dynamics among tracks within the dataset.

There are 7 types of metadata: the edge weights, uri, artist’s name, track name, duration, album name, and position. Edge weight indicates how many times the tracks appear together in different playlists. URI serves as a unique identifier for the music track, often used to locate or reference the track within a database or online server, which in this

[4] https://darehunt.github.io/DSC180B-Project2/
[5] https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge
[6] https://github.com/KyoKii02/CPSC572-Project/blob/main/network.py

case is spotify. Artist's name represents the name of the artist(s) associated with a particular track. Track name represents the title or name of the individual music track. Duration represents the length of the track in milliseconds. Album name refers to the title of the album to which the track belongs. Position indicates the track's position in each playlist it is found in.

Basic Statistics[7]

Referring to the statistics shown in Table 1, the given number of nodes and edges indicate that the network has a considerable number of connections between nodes. With 36 connected components, however, the network exhibits a number of smaller disjointed subnetworks. This is most likely due to people's tendencies to create playlists that are focused around specific genres, moods, atmospheres, etc., resulting in a lack of significantly different types of music being grouped together.

Given the average degree of approximately 42.28, it likely suggests that tracks tend to appear together with a greater number of other tracks, indicating versatility or common occurrences within playlists. This becomes logical when taking into account the multitude of separate networks where tracks circulate around distinct musical themes. When examining the minimum and maximum degrees, there exists a correlation with the popularity of the tracks. Specifically, the higher the degree, the more likely the tracks are to be interconnected and frequently encountered alongside other tracks.

When considering the clustering coefficient, the average given as a result from the graph indicates that tracks within playlists form cohesive thematics or stylistic groups. This suggests the formation of tightly-knitted clusters and communities consistent with earlier analysis. Due to the fragmented network however, the path length can only be calculated through individual subnetworks. This is highly impractical given that the size of individual connected components, found in Table 2, are extremely small, with the exception of the largest component.

Characteristics	**Statistics**
Number of Nodes	5420
Number of Edges	114579
Number of Connected Components	36
Average Degree	42.28007380073801

[7] https://github.com/KyoKii02/CPSC572-Project/blob/main/graph_stats.ipynb

Minimum Degree	1
Maximum Degree	1528
Average Clustering Coefficient	0.6186053620486062
Number of Nodes (LC)	5222
Number of Edges (LC)	114247
Average Clustering Coefficient (LC)	0.624138888343053
Average Path Length (LC)	3.354811032926788

Table 1: Presents statistics specific to the network as a whole, and incorporates data on the large component (LC).

Component	**Size**	**Component**	**Size**	**Component**	**Size**	**Component**	**Size**
1	5222	10	6	19	2	28	2
2	3	11	2	20	6	29	4
3	23	12	9	21	3	30	2
4	2	13	6	22	2	31	2
5	2	14	2	23	5	32	2
6	12	15	2	24	9	33	2
7	2	16	2	25	2	34	18
8	10	17	2	26	13	35	14
9	2	18	11	27	2	36	10

Table 2: Lists the node count of each individual component subnetwork.

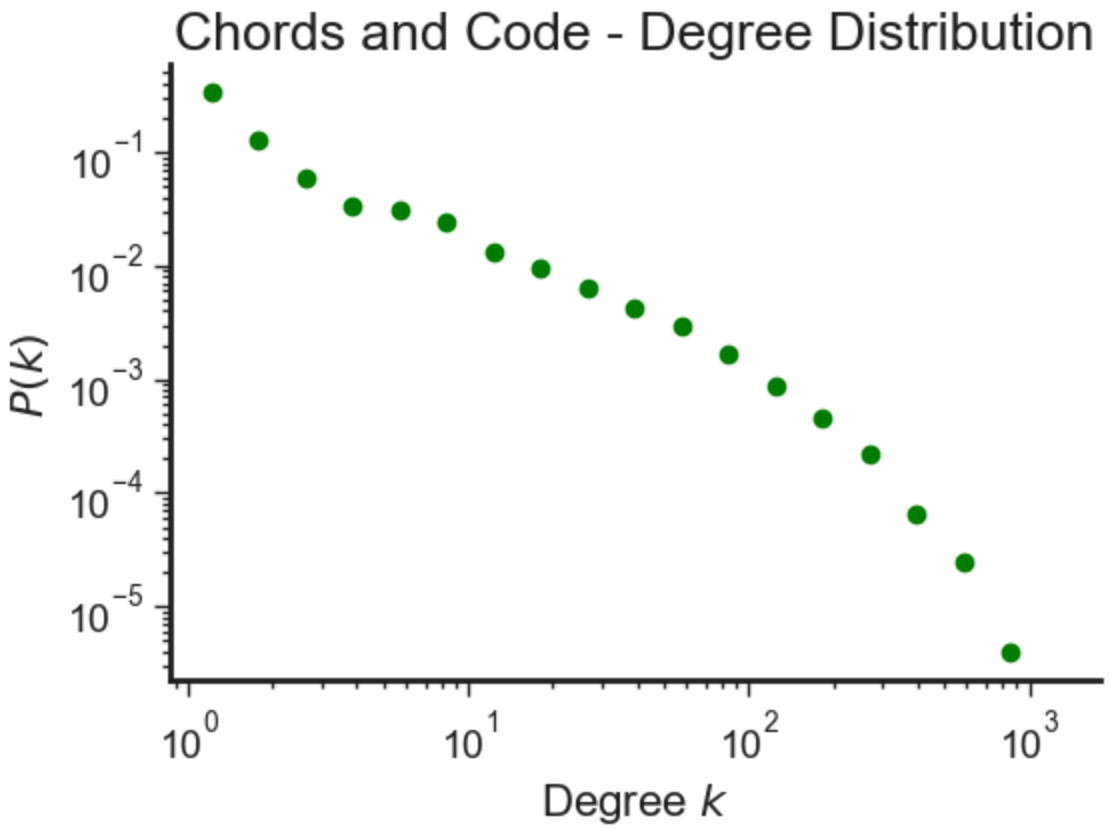

Figure 1: The degree distribution of the network plotted on a log-log scaled graph.

Since the largest component contains the significant majority of nodes and edges of the original network, conducting further analysis on it would be ideal. Now when considering the clustering coefficient, the result compared to the original network is only about a ~0.0055 increase. This indicates the subnetwork is relatively close to the original as it still demonstrates a higher level of connectivity. The average path length is shown to be around ~3.35 which leads to the conclusion that networks are small, since nearly every connected node can be reached within three to four edges. This suggests that tracks are frequently found together in playlists, and that related tracks can easily be discovered and accessed.

The degree distribution from this data, shown in Figure 1, shows a power law distribution given its right skewness of the graph. This indicates that there exists a large number of low-degree nodes and a small number of high-degree nodes. As a result, this likely suggests that high-degree nodes are popular tracks, which most likely acts as hubs that connect different parts of the network and allow for interactions between various playlists.

Characteristics	Erdos-Renyi (ER)	Degree Preservation (DP)
Clustering Coefficient Average	0.007801313114217683	0.17562577036279403
Clustering Coefficient Standard Deviation	0.00006771166540408483	0.002516002750675417
Shortest Path Length Average	2.705746813351138	2.76197115757219
Shortest Path Length Standard Deviation	0.0014058893056229141	0.0026808508205250887

Table 3: Presents statistics related to the null models, specifically for Erdős–Rényi and degree preservation using double-edge swaps.

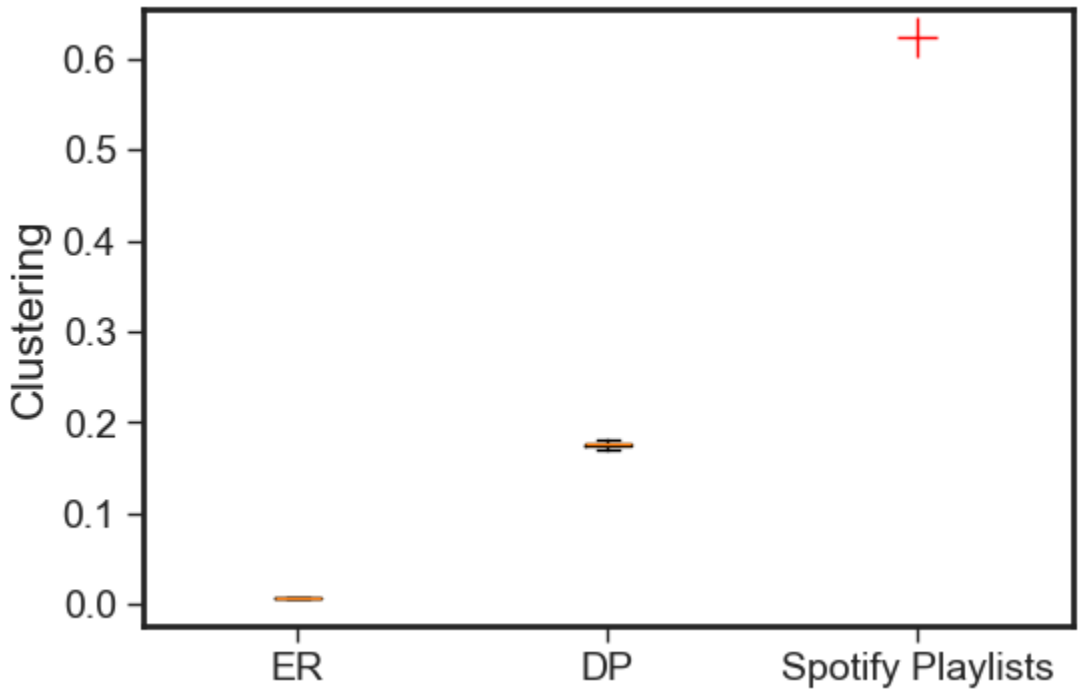

Figure 2: Presents a visual graph of the clustering coefficients, comparing the null models and the original network.

In contrast to the original network, the statistics presented in Table 3 for the null models reveal notable disparities. By focusing on the metrics concerning the largest component, the average clustering coefficient substantially exceeds those of the null models. This discrepancy implies the presence of clustering and interconnectedness among nodes within the original network. As evident in Figure 2, even when preserving the degrees, the null models fall short in replicating the high clustering coefficient observed in

the original network. Moreover, the path length data for the largest component are larger than what was observed in the context of the random graphs. The relatively similar results between the ER and DP graphs suggest that the original network exhibits characteristics indicative of non-random structured topology. This leads to the conclusion that the differences observed for the structural properties of the original network are not solely determined by random or degree-preserving processes. This is most likely due to the influences of other factors such as community structure and preferential attachments.

Network Visualization

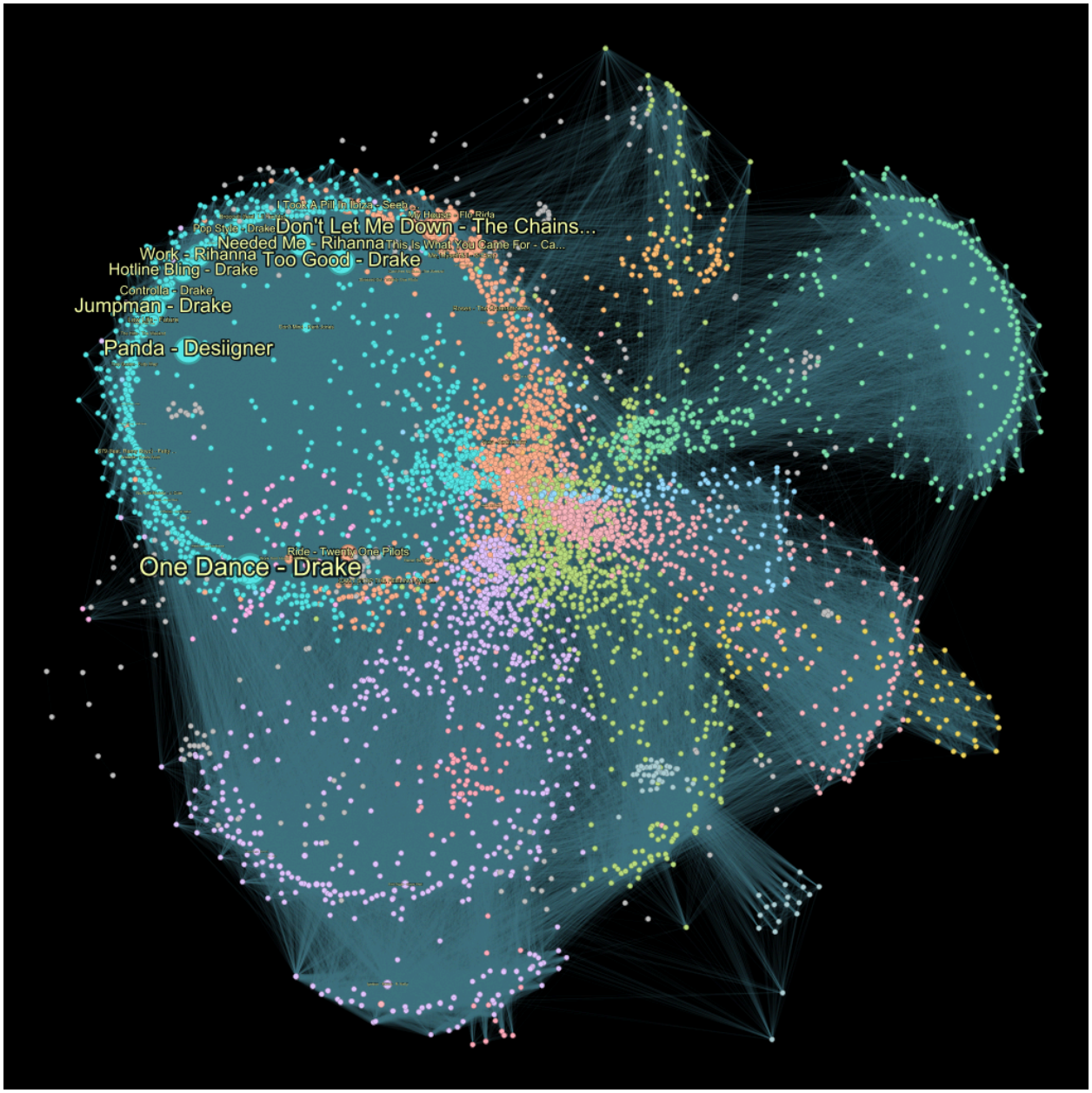

Figure 3: A visualization of the network where node size is correlated to degree, and node color identifies the communities. The top fifty highest degree nodes are labeled with their song title and artist. The edge opacity is set low to highlight the interconnectedness present within communities.

From the visualization in Figure 3, it is apparent that several distinct and densely connected communities are present in the network, more detail about the specifics of these communities can be found in the following section. In order to avoid all the hub nodes being pushed into the center, the attraction distribution feature of the force atlas layout in Gephi was used, so hubs are towards the outside and you can see several smaller (lower degree) nodes gathering in the center of the image. Notably, some communities intermingle with each other, often in quite logical ways. This is most evident in the two largest communities, which contain the most popular songs, assigned light blue and orange, where the two communities almost blend into one. Additionally, this visualization demonstrates how powerfully the popularity of songs can impact the connectivity of the network, even across different communities. This factor is evident by the large conjunction of communities seen in the upper center of the visualization, where popular songs are connected to other popular songs in the network, regardless of their community. Another notable factor from this visualization is the large separation that exists between the less popular communities. This can be hinting at the uniqueness of music taste, as the music becomes more obscure.

Network Analysis[8]: Community Detection Among Songs[9]

As discussed in the null model section, it is evident that this network is highly clustered and communities exist within the network. Therefore, a community detection approach is utilized to determine clusters and partitions based on the modularity of each partition. To ensure that our approach was timely, the greedy algorithm based, Louvain method was used to establish these partitions. Since the Louvain algorithm can have randomness to it, 100 partitions were created and analyzed.

The Louvain partition ensemble resulted with an average modularity score of 0.5651, suggesting a moderately strong community structure, and a standard deviation of 0.0013, which indicates that the various permutations of this algorithm generally arrive at consistent results. A boxplot of this spread is shown on Figure 4. This figure also includes the maximum modularity score at 0.5664, which we selected as the community partition for our analysis.

[8] https://github.com/KyoKii02/CPSC572-Project/blob/main/graph_stats.ipynb
[9] https://github.com/KyoKii02/CPSC572-Project/blob/main/community_detection.ipynb

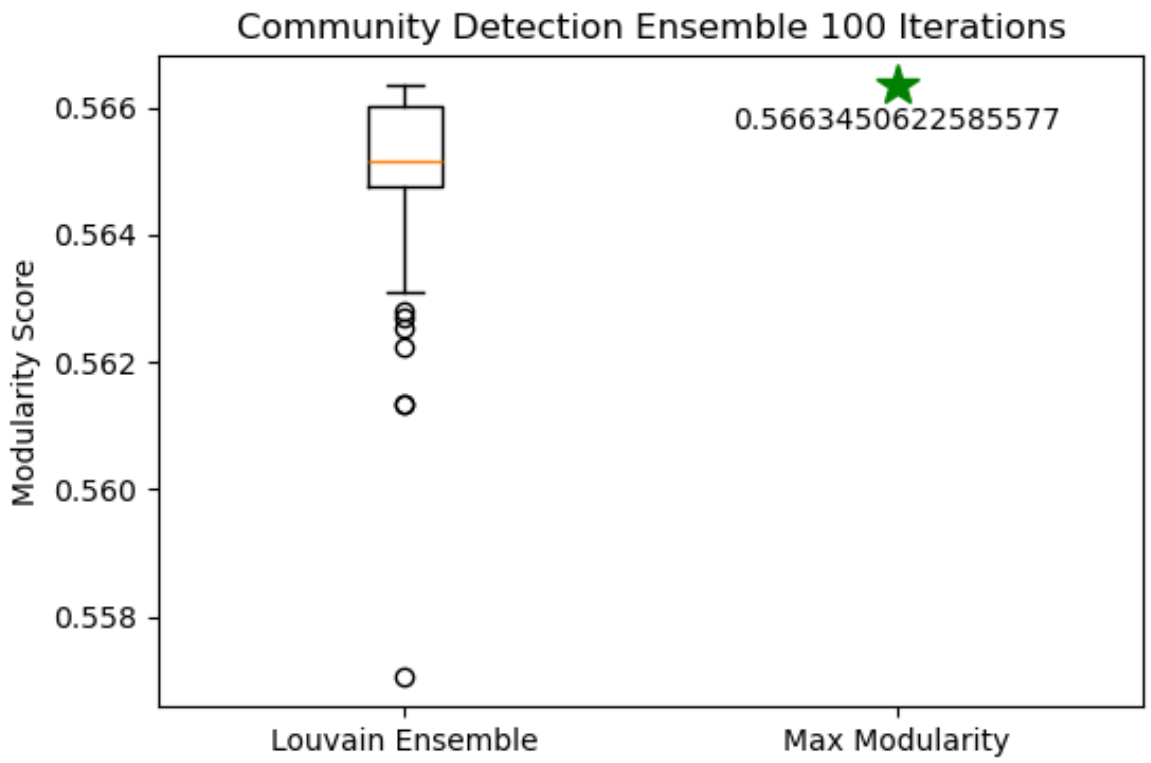

Figure 4: Spread of Louvain ensemble and maximum modularity

This partition of maximum modularity, displayed with the green star on the left, was taken on for further analysis, to find commonalities between these communities.

Community insights on popularity, genre, moods, and trends:

The initial hypothesis was that playlist constructions are highly influenced by popularity, genre, moods, and trends. Below in Figure 5, communities identified by the Louvain algorithm have been grouped and filtered to only include the ones that are of a size larger than 1% of the network, then their top 5 most popular artists are displayed on bar charts. With this, similarities between artists are assessed, and to further conclude what exactly joins these playlists together. To supplement this analysis, the top 20 songs from each community have been taken based on popularity, and included within Appendix A.

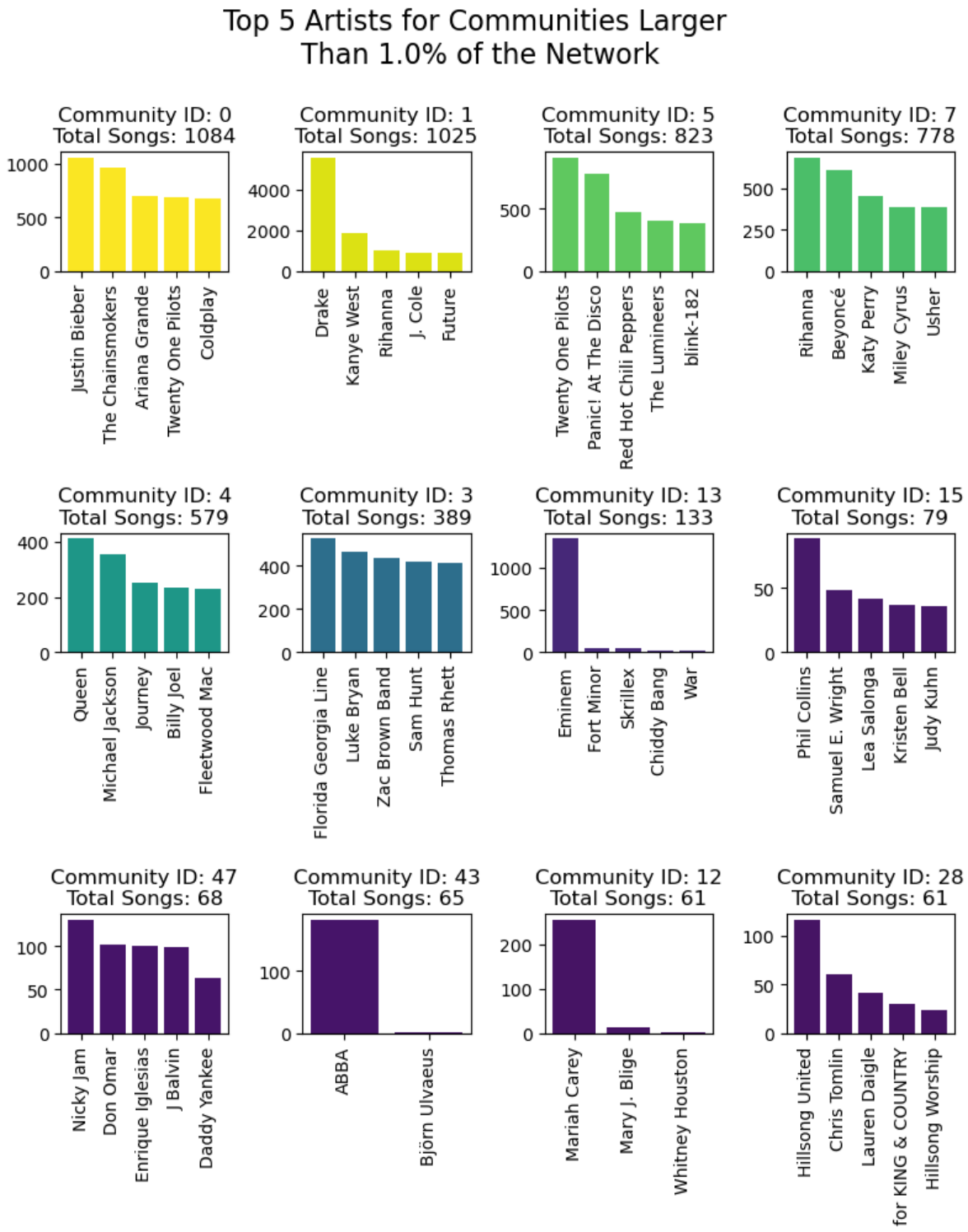

Figure 5: Top 5 artists for each community larger than 1% of the network size, where the x-axis displays the number of occurrences that the artist has in that community and the y-axis displays the top 5 artists, if there are at least 5 in that community.

Analysis for Each Community

1. **Community ID: 0 - (Pop) Top Charts Playlists**

This playlist comprises the most popular artists of 2016. It's a collection that everyone can enjoy, filled with catchy hooks. You will likely know the lyrics to these songs without giving it much thought. This playlist excludes rap songs, as they cater to a specific taste, regardless of rap's popularity in 2016. Here, you'll find all the songs that were played on the most popular radio stations at the time, such as Don't Let Me Down by The Chainsmokers, Can't Stop the Feeling! by Justin Timberlake, and Sorry by Justin Bieber.

2. **Community ID: 1 - Popular Hip-Hop Playlists**

The next community features popular Hip-Hop playlists, with Drake as the most popular artist, followed by other Hip-Hop artists known for their hit songs from that year. What is most intriguing about this community is Drake's dominance. This might be attributed to Drake popularizing the Hip-Hop genre in 2016 by incorporating more dance elements into it. Consequently, individuals who started listening to Hip-Hop because of Drake also explored music by other significant Hip-Hop artists like Kanye, Rihanna, and J. Cole, who were not as popular at the time. In essence, Drake served as a gateway for listeners to other Hip-Hop artists.

3. **Community ID: 5 - Popular Alternative Playlists**

This playlist primarily features alternative artists such as Twenty One Pilots, Panic! At The Disco, Blink-182, and others. Uniquely, this community also includes a significant presence of the Red Hot Chili Peppers, indicating that the community is not strictly unified by genre. Listeners of alternative rock also tend to enjoy a small selection of the most popular classic rock. This provides a small insight into how music tastes begin to overlap across genres.

4. **Community ID: 7 - "Throwback" Playlists (Late 90s - Early 2010's)**

This community showcases playlists featuring songs that were not necessarily new at the time but have remained popular. The younger generation might refer to these as "throwback" playlists, because even though these songs are not old, they were released in the late '90s, 2000s, and early 2010s. Here, you will find songs that were continuously played and are still heard at clubs and parties today.

5. **Community ID: 4 - Classic Hits Playlists**

These playlists feature some of the most timeless artists, including music that was released between the 1960s and the early 1980s. Here, you will find some of the most influential artists of all time such as Queen, Michael Jackson, Journey, Lynyrd Skynyrd,

among many others. This playlist is not limited to one genre, as it incorporates a wide range of music that was popular in the past.

6. **Community ID: 3 - Country Playlists**

This community continues the trend of featuring the most popular artists from a specific genre, focusing exclusively on country music. It is important to note that traditional country fans might not fully connect with this selection. Since our metrics are based on popularity, the more catchy and popular songs tend to stand out. Here, we can find the largest names of pop-country such as Florida Georgia Line, Luke Bryan, and Thomas Rhett.

7. **Community ID: 13 - Eminem Playlists**

Eminem's popularity in 2016 cannot be understated, and this community of playlists showcases just that. This community almost exclusively features songs released by Eminem, with no direct relation to a tour or major releases happening for Eminem that year. Eminem has a similar effect on these playlists as Drake had on the Hip Hop playlist, where most people primarily listened to Eminem but then began to also explore similar music. The most notable finding from this community is simply how influential Eminem has been to the Rap genre.

8. **Community ID: 15 - Disney Soundtrack Playlists**

While looking at the top artists of this playlist one might wonder what the commonalities actually are between them. However, a closer look at the top songs of this community gives us the full story. The top songs here are large hits from popular movies such as, Under the Sea from Little Mermaid, Hakuna Matata from the Lion King, A Whole New World from Aladdin, along with many more. This community shows us correlation that goes further than genre, an association based on the origins of these songs, such as popular movies!

9. **Community ID: 47 - Reggaeton Playlists**

This community continues our trend of focusing on specific genres, featuring playlists that include the most popular songs from the youthful, Latin American genre of reggaeton. Here we find some of the most influential latin artists of that year such as Don Omar, Nicky Jam, Enrique Iglesias, and J Balvin. Surprisingly, this community is more popular than regular Latin playlists, which highlights the rise in popularity of reggaeton among the Latino population in the 2010s.

10. **Community ID: 43 - ABBA**

There is not much more to be said about this community, as it is entirely devoted to ABBA. The other artist featured here, Björn Ulvaeus, was also a member of ABBA. However,

he continued to record music after the band broke up. The key takeaway here is the sheer popularity of ABBA, thanks in part to movies such as Mamma Mia!

11. Community ID: 12 - Mariah Carey

This community is distinctly focused on one artist, Mariah Carey, yet also features other extraordinary talents of Mary J. Blige and Whitney Houston, who share a commonality with Carey in their vocal mastery. All three of these artists have some of the most impressive vocal ranges in the industry. However, Mariah Carey's dominant presence in this community suggests that her popularity has led listeners to explore similar artists, including the aforementioned ones of Blige and Houston.

12. Community ID: 28 - Christian Music Community

This playlist features music celebrated in the Christian church, highlighting artists well-known for their messages of faith, worthship, and devotion. Prominent figures such as Hillsong United, Chris Tomlin, and Lauren Daigle are some of the largest names in the Contemporary Christian Music Community, as showcased in the top 100 worship songs of all time from Praisecharts[10]. This community once again displays the power that collective listening can have on music taste, grouping tracks together not only by genre but also by the shared faith and values of the listeners.

Community Analysis Findings

The results of this community analysis demonstrate the significant roles of popularity and genre in shaping playlist creation. It also shows the influence popular artists can have in inspiring the creation of playlists centered around them, with artists such as Drake, Eminem, ABBA, and Mariah Carey, being the most prominent among those influencers. Additionally, nostalgia can play a pivotal role when constructing playlists, as evident in the throwback and classic hits playlists. Lastly, even cultural and thematic connections of movie soundtracks or Christian communities have been shown to influence the connections between songs, displayed in the Disney soundtrack playlists and the Christian playlists. Ultimately, it can be established that all playlists demonstrate a unique approach to their creation, and there is not a one-size-fits-all approach when constructing playlists.

Position Analysis on Song Popularity

With a cursory analysis of the graph shown in Figure 6, it appears that less popular songs end up on shorter playlists and are more evenly distributed in their position in a playlist. As a song gets more popular it will likely tend toward the beginning of a playlist and is present in longer playlists (think top 100 type lists).

[10] https://www.praisecharts.com/song-lists/top-100-worship-songs-of-all-time

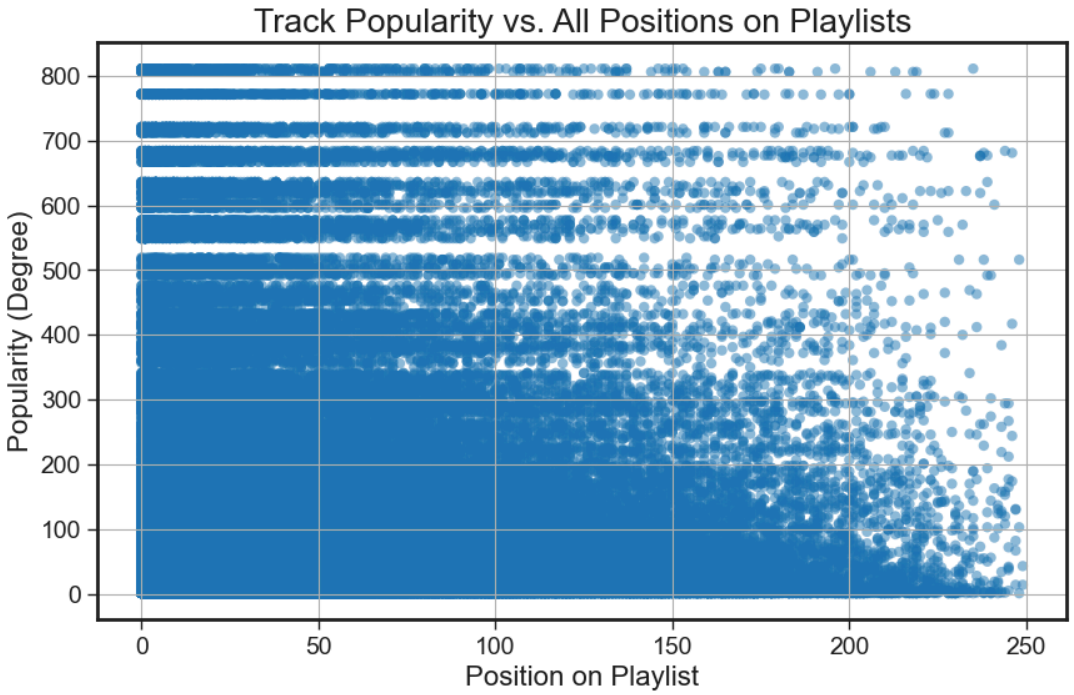

Figure 6: Relationship between the popularity or degree of a song, and its position on a playlist.

Calculating the Spearman's rank correlation coefficient and getting approximately −0.056 indicates a very weak negative correlation between the position on a playlist and the track's popularity as measured by degree. This suggests that, generally, tracks positioned earlier in playlists are slightly more likely to be more popular, but the relationship is not strong. The p-value is extremely low (much less than 0.05) at 7.74e-93, which suggests that the correlation is statistically significant. In other words, the likelihood that this correlation is due to random chance in the sample data is extremely low. Therefore, we can be confident that there is a consistent, though weak, tendency across your dataset for tracks earlier in playlists to be slightly more popular. The implications of this finding could be of interest for further investigation, particularly in understanding playlist dynamics and user behavior. Despite the weak correlation, the result might suggest that playlist curators tend to place more popular tracks earlier, or that being earlier on a playlist slightly increases a track's visibility or likelihood of being played, and thus its popularity.

Position Analysis on Song Popularity

The most popular songs tend to be around the average song length, with a distribution between 2.5 and 5 minutes, whereas less popular songs are more likely to fall outside of that range, (although not by much) falling between 20 seconds and 12 minutes, this is showcased on Figure 7.

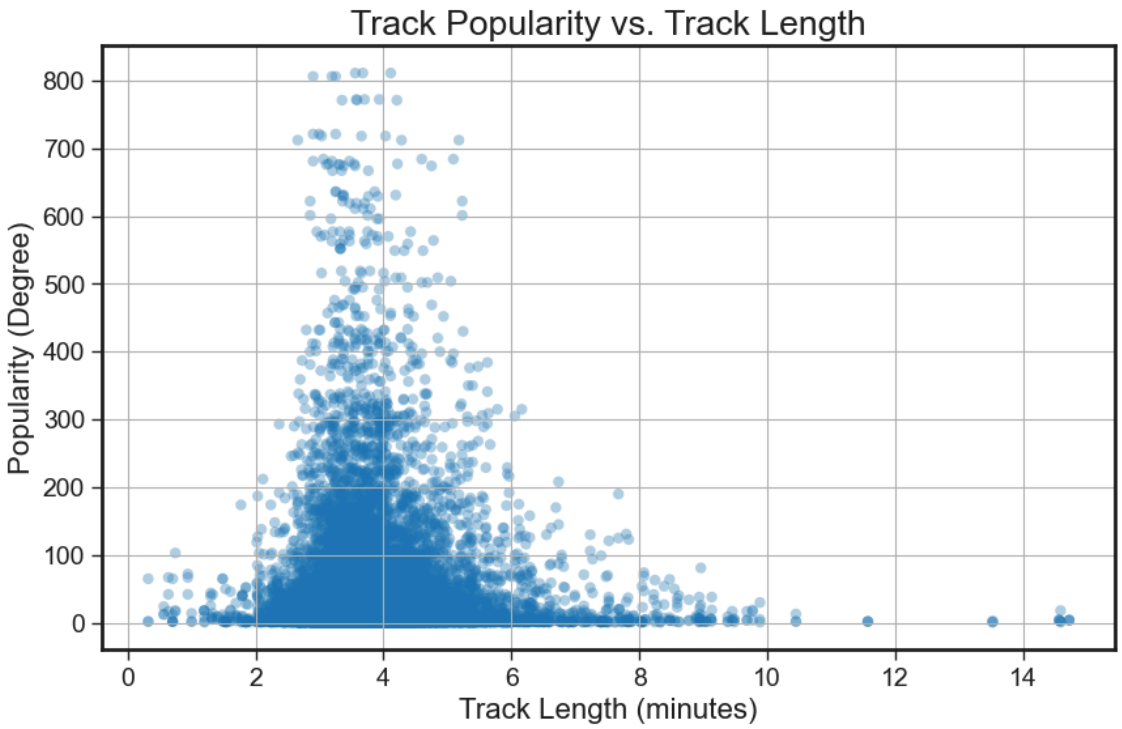

Figure 7: Relationship between a songs degree or popularity and its length

We end up with something resembling a normal distribution around 4 minutes. After doing some statistical analysis, results seem to indicate that there is a slight negative correlation between track length and popularity. The negative curve of the regression indicates that as track length increases, the popularity (degree) tends to decrease slightly. The intercept (38.99) suggests that if a track had a length of 0 minutes (which wouldn't really be much of a song), the model would predict its popularity to be around 39 degrees. It should also be noted that the track length potentially only explains a very small part of the variance in the track's popularity. Therefore, while there may be a tendency for shorter tracks to be more popular, many other factors will be more influential in determining a track's popularity.

Final Results

Overall, the analysis performed has led to the findings of five major forces that people are subject to when selecting playlists: The influence of popularity, genre and mood work as a cohesive force, the importance of nostalgia and cultural themes, visibility and popularity early in the playlist, and the song length. The completed analysis demonstrates the complex interplay between all of these forces, where the overarching factors that were found to be common among all these are, individual preferences, social connections, and cultural or thematic contexts.

Discussion

Our results enable us to discover the gaps in the existing literature on the co-occurrence of songs within playlists. By analyzing the similarities between songs that are grouped within communities, we can infer the factors that influence individuals when creating playlists. However, our study also reveals the complexity of constructions, particularly how there is no single approach that can be generalized to all playlists. Instead, several factors contribute to the addition of other songs when a subset is selected for a playlist.

Playlists are highly interconnected, and unique communities emerge from within them, spanning genres, time periods, and human cultures. The research questions were all addressed, exploring communities, song length, and position on playlist, and how these factors interact with each other in the formation and proliferation of playlists. There is still much more to be understood and learned from the dataset explored, and additional data (perhaps with more recency) would also reveal more depth into the subjects discussed.

The constraints imposed by limited computational resources necessitated a narrowed scope of our exploration, consequently reducing our dataset size to a specific point in time. This resulted in a snapshot view of the dataset as opposed to what could be achieved by exploring all the playlists the dataset has to offer. The data could be greatly complemented by incorporating a deeper set of details about each song, such as genre, recording date, record label, and total number of lifetime streams. Furthermore, supplementing each playlist with demographic information regarding its creator and audience would enable a more nuanced sociological analysis. Future work in this area should consider utilizing more computing resources to process a higher proportion of the data. Another possible avenue of exploration would be to model the evolution of a network like this over time, perhaps letting each playlist in the entire dataset occupy its place in the graph for a set time, in accordance to the playlist creation date.

Methods

Network Construction. The network is constructed using python based on Spotify's playlist data from the first week of August 2016. The dataset used is stored in JSON files which are parsed and extracted using the NetworkX library. The metadata extracted contains the following information: track URI, artist name, track, duration, album name, and position within playlists.

Each track in the playlist is represented by a node in the network using a specific URI (unique identifier) for each track. If a node with a specific URI does not exist, a new node is created with attributed metadata assigned to it. If the node already exists, it instead

appends the new position to the existing list of positions. The edges between nodes are created based on the co-occurrences within playlists. An edge is created between each pair of tracks within a playlist if they haven't been assigned before. Starting at an initial edge weight of 1, this value is incremented each time the same pair is found again within another playlist.

After the creation of nodes and edges, the network is then filtered, removing edges with weights below 5, and isolated nodes without edges afterwards. This reduces the noise present and allows for a better focus on significant relationships within the network.

Null Models. The Erdős-Rényi model serves as a null hypothesis where edges are randomly formed. The properties of the original network are compared to those of the ER random graphs to assess whether the observed properties are significantly different from what would be expected by random chance. This process is iterated a hundred times, creating graphs with similar size and edge density to the original network.

The degree preservation model preserves the degree distribution of the original network while randomizing the edges. This is to establish whether or not the observed properties are influenced by the network's degree distribution alone. Similar to the previously mentioned model, this process is iterated a hundred times as well. Graphs are created using a copy of the original network and applying double-edge swapping, aiming to maintain degree distribution.

Community Detection. Due to the size of our network, the well-known Louvain algorithm to detect networks of songs is used to identify communities of songs tied together by playlists. This algorithm aims to maximize a modularity score for each community, continuously creating communities until the modularity score does not increase. This algorithm is particularly important for our network as it was created with a greedy optimization method that runs in time O(nlogn), with n representing the number of nodes in the network. Due to this, this algorithm can efficiently handle our network size of over 5,000 nodes.

Since the partitions resulting from the Louvain method are different each time, an ensemble consisting of 100 partitions is used, and the partition with the highest modularity score is used.

Software and Code[11]

We made use of the following libraries in our code:

python-louvain[12] (community detection algorithm)
matplotlib[13]
networkx[14]
numpy[15]
pandas[16]

For visualization creation:

Gephi[17]

The following files in our Github repo were included with the dataset distribution and were used for preliminary dataset exploration:
check.py
deeper_stats.py
descriptions.py
print.py
show.py
stats.py

The following files were created by us, and their purposes are listed below:
Community_detection.ipynb - Community detection and analysis.
Graph_stats.ipynb - Computes basic statistics, null models, and deeper analysis of the data.
graph_stats_Poly_regression.ipynb - an experiment with polynomial regression on the previous code.
Vizualize_graph.py - a simple visualization to validate the construction of the network.
network.py - The main network construction code.
spotify_AugWeek1.graphml - the result of network.py
Network_with_communities.graphml - the above graph with community detection applied.

[11] https://github.com/KyoKii02/CPSC572-Project
[12] https://github.com/taynaud/python-louvain
[13] https://github.com/matplotlib/matplotlib
[14] https://github.com/networkx
[15] https://github.com/numpy/numpy
[16] https://pandas.pydata.org/
[17] https://gephi.org/

References

Literature Sources:

1. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3741536/*
2. *https://journals.sagepub.com/doi/10.1177/2056305119847514*
3. *https://arxiv.org/abs/physics/0512266*
4. *https://darehunt.github.io/DSC180B-Project2/*
5. *https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge*
6. *https://www.praisecharts.com/song-lists/top-100-worship-songs-of-all-time*

Technical Sources:

7. *https://github.com/KyoKii02/CPSC572-Project*
8. *https://github.com/taynaud/python-louvain*
9. *https://github.com/matplotlib/matplotlib*
10. *https://github.com/networkx*
11. *https://github.com/numpy/numpy*
12. *https://pandas.pydata.org/*
13. *https://gephi.org/*

Appendix A: Top 20 Songs per Communities Larger than 1% of the Network

Consists of the top 20 songs per community based on the number of playlists each song has occurred in.

Community: 0 | Size of Community: 1084

Name Artist Number_of_Playlists
116 Don't Let Me Down The Chainsmokers 328
117 This Is What You Came For Calvin Harris 317
129 Ride Twenty One Pilots 280
101 I Took A Pill In Ibiza - Seeb Remix Mike Posner 262
283 Cold Water (feat. Justin Bieber & MØ) Major Lazer 262
119 CAN'T STOP THE FEELING! (Original Song from Dr... Justin Timberlake 261
133 Never Be Like You Flume 244
86 Roses The Chainsmokers 239
118 Work from Home Fifth Harmony 232
125 Gold Kiiara 228
110 Middle DJ Snake 227
35 Me, Myself & I G-Eazy 219
88 My House Flo Rida 214
235 Heathens Twenty One Pilots 209
85 Stressed Out Twenty One Pilots 201
284 Closer The Chainsmokers 196
123 Cheap Thrills Sia 196
555 Sorry Justin Bieber 187
134 i hate u, i love u (feat. olivia o'brien) gnash 174
82 Love Yourself Justin Bieber 173

Community: 1 | Size of Community: 1025

Name Artist Number_of_Playlists
114 One Dance Drake 592
120 Too Good Drake 320
111 Panda Desiigner 310
115 Needed Me Rihanna 296
692 Broccoli (feat. Lil Yachty) DRAM 284
602 Controlla Drake 264
43 Work Rihanna 255
538 Jumpman Drake 239
693 Don't Mind Kent Jones 237
959 Pop Style Drake 225
728 Low Life Future 205

467 Hotline Bling Drake 192
1136 No Role Modelz J. Cole 180
260 Famous Kanye West 177
202 The Hills The Weeknd 172
44 2 Phones Kevin Gates 169
1299 oui Jeremih 167
1236 679 (feat. Remy Boyz) Fetty Wap 153
282 No Problem (feat. Lil Wayne & 2 Chainz) Chance The Rapper 152
96 Don't Bryson Tiller 152

Community: 5 | Size of Community: 823
Name Artist Number_of_Playlists
561 Mr. Brightside The Killers 129
384 Ophelia The Lumineers 107
1785 Sugar, We're Goin Down Fall Out Boy 106
448 Santeria Sublime 93
328 All The Small Things blink-182 91
678 Tear In My Heart Twenty One Pilots 91
254 Car Radio Twenty One Pilots 84
58 Semi-Charmed Life Third Eye Blind 82
1782 I Write Sins Not Tragedies Panic! At The Disco 81
357 Stolen Dance Milky Chance 79
381 Midnight City M83 74
259 Home Edward Sharpe & The Magnetic Zeros 74
356 Electric Feel MGMT 74
1439 Island In The Sun Weezer 74
386 Californication Red Hot Chili Peppers 72
462 The Middle Jimmy Eat World 72
2041 Under The Bridge Red Hot Chili Peppers 71
1742 Do I Wanna Know? Arctic Monkeys 71
1903 Walking On A Dream Empire of the Sun 68
2816 Sweater Weather The Neighbourhood 66

Community: 7 | Size of Community: 778
Name Artist Number_of_Playlists
161 Ignition - Remix R. Kelly 192
416 Gold Digger Kanye West 133
756 Hey Ya! - Radio Mix / Club Mix OutKast 123
637 Crazy In Love Beyoncé 121
741 Yeah! Usher 117

256 It Wasn't Me Shaggy 117
160 She Will Be Loved - Radio Mix Maroon 5 117
170 Pumped Up Kicks Foster The People 115
141 Promiscuous Nelly Furtado 114
152 I'm Yours Jason Mraz 113
165 We Can't Stop Miley Cyrus 111
2326 Wonderwall - Remastered Oasis 110
1226 Ride Wit Me Nelly 108
174 Drops of Jupiter Train 104
171 Hey There Delilah Plain White T's 102
143 Paper Planes M.I.A. 102
142 Kiss Me Thru The Phone Soulja Boy 101
179 Chasing Cars Snow Patrol 99
1502 Stronger Kanye West 97
1462 Gives You Hell The All-American Rejects 97

Community: 4 | Size of Community: 579

Name Artist Number_of_Playlists
178 Don't Stop Believin' Journey 146
1684 Bohemian Rhapsody - Remastered 2011 Queen 105
54 Brown Eyed Girl Van Morrison 98
656 September Earth, Wind & Fire 89
915 Piano Man Billy Joel 81
52 Sweet Child O' Mine Guns N' Roses 79
55 Sweet Home Alabama Lynyrd Skynyrd 78
922 Ain't No Mountain High Enough Marvin Gaye 74
1726 Take On Me a-ha 74
2723 Come On Eileen Dexys Midnight Runners 70
2249 Hotel California - Remastered Eagles 70
459 You Make My Dreams - Remastered Daryl Hall & John Oates 68
1223 My Girl The Temptations 68
655 I Wanna Dance with Somebody (Who Loves Me) Whitney Houston 65
610 Livin' On A Prayer Bon Jovi 65
1990 Africa Toto 65
917 Sweet Caroline Neil Diamond 63
1989 Under Pressure - Remastered 2011 Queen 63
277 I Want You Back The Jackson 5 60
458 Carry on Wayward Son Kansas 59

Community: 3 | Size of Community: 389

Name Artist Number_of_Playlists
549 H.O.L.Y. Florida Georgia Line 103
1090 Die A Happy Man Thomas Rhett 97
1099 House Party Sam Hunt 91
1235 T-Shirt Thomas Rhett 85
963 Somewhere On A Beach Dierks Bentley 74
989 Wagon Wheel Darius Rucker 70
1093 Chicken Fried Zac Brown Band 68
51 Cruise Florida Georgia Line 67
978 Knee Deep (feat. Jimmy Buffett) Zac Brown Band 66
980 Take Your Time Sam Hunt 64
2078 Snapback Old Dominion 63
1265 Break Up In A Small Town Sam Hunt 62
553 Crash And Burn Thomas Rhett 61
965 Make You Miss Me Sam Hunt 59
1002 Play It Again Luke Bryan 59
966 Head Over Boots Jon Pardi 58
1081 Leave The Night On Sam Hunt 57
1622 From the Ground Up Dan + Shay 56
1023 Barefoot Blue Jean Night Jake Owen 53
1030 American Kids Kenny Chesney 53

Community: 13 | Size of Community: 133
Name Artist Number_of_Playlists
1493 Lose Yourself - Soundtrack Version Eminem 95
1492 'Till I Collapse Eminem 68
1367 The Real Slim Shady Eminem 60
1363 Rap God Eminem 51
1678 Without Me Eminem 47
1509 Purple Lamborghini (with Rick Ross) Skrillex 45
2984 The Monster Eminem 45
2941 Remember The Name (feat. Styles Of Beyond) Fort Minor 43
1496 Not Afraid Eminem 42
4417 My Name Is Eminem 40
1500 Mockingbird Eminem 40
4030 Shake That Eminem 33
1365 Stan Eminem 29
3808 Berzerk Eminem 26
2070 Opposite Of Adults Chiddy Bang 22
1498 Beautiful Eminem 21

1366 The Way I Am Eminem 20
1679 Like Toy Soldiers Eminem 20
4042 Just Lose It Eminem 20
1680 No Love Eminem 20

Community: 15 | Size of Community: 79

Name Artist Number_of_Playlists
821 Under the Sea - From "The Little Mermaid"/ Sou... Samuel E. Wright 28
823 Hakuna Matata Nathan Lane 26
3698 A Whole New World Lea Salonga 25
3726 I Won't Say (I'm in Love) Lillias White 24
3735 Let It Go - From "Frozen"/Soundtrack Version Idina Menzel 24
1699 You've Got A Friend In Me - From "Toy Story"/ ... Randy Newman 23
820 Part of Your World - From "The Little Mermaid"... Jodi Benson 23
3727 I'll Make a Man Out of You - From "Mulan"/Soun... Donny Osmond 22
3721 Colors Of The Wind Judy Kuhn 21
3717 Circle Of Life - From "The Lion King"/Soundtrack Carmen Twillie 21
822 Kiss the Girl - From "The Little Mermaid"/Soun... Samuel E. Wright 20
3729 Hawaiian Roller Coaster Ride - From "Lilo & St... M. Keali'i Ho'omalu 20
2261 Strangers Like Me Phil Collins 19
2303 You'll Be In My Heart Phil Collins 19
3701 You'll Be In My Heart Phil Collins 18
3732 I See the Light - From "Tangled"/Soundtrack Ve... Mandy Moore 18
3737 Reflection - From "Mulan"/Soundtrack Version Lea Salonga 16
3724 Go the Distance - From "Hercules"/Soundtrack Roger Bart 16
828 Do You Want to Build a Snowman? Kristen Bell 15
3710 The Bare Necessities Bruce Reitherman 15

Community: 47 | Size of Community: 68

Name Artist Number_of_Playlists
2014 Danza Kuduro Don Omar 49
2150 El Perdón Nicky Jam 47
2069 Bailando - Spanish Version Enrique Iglesias 43
2630 DUELE EL CORAZON Enrique Iglesias 40
1868 Hasta el Amanecer Nicky Jam 35
1137 Ginza J Balvin 34
4001 Vivir Mi Vida Marc Anthony 22
4658 Ay Vamos J Balvin 20
2632 6 AM J Balvin 20
2642 La Gozadera Gente De Zona 19

4000 Suavemente Elvis Crespo 19
2152 Travesuras Nicky Jam 19
3027 Limbo Daddy Yankee 17
2643 Mayor Que Yo 3 Luny Tunes 17
830 Cuando Me Enamoro Enrique Iglesias 17
2644 Borro Cassette Maluma 16
2629 Bobo J Balvin 14
3959 Fanática Sensual Plan B 13
2639 Vaivén Daddy Yankee 13
2636 Zumba Don Omar 12

Community: 43 | Size of Community: 65
Name Artist Number_of_Playlists
4527 Take A Chance On Me ABBA 10
4917 Mamma Mia ABBA 10
4485 Waterloo ABBA 8
4697 Fernando ABBA 7
4863 Honey, Honey ABBA 6
5404 Ring, Ring - Swedish Version ABBA 5
4743 The Winner Takes It All ABBA 4
5181 One Of Us ABBA 4
5307 Gimme! Gimme! Gimme! (A Man After Midnight) ABBA 4
4484 Money, Money, Money ABBA 4
3798 I Do, I Do, I Do, I Do, I Do ABBA 3
5369 Under Attack ABBA 3
5411 He Is Your Brother ABBA 3
5407 People Need Love ABBA 3
5389 The Day Before You Came ABBA 3
5381 Love Isn't Easy (But It Sure Is Hard Enough) ABBA 3
5384 The Visitors ABBA 3
5364 Head Over Heels ABBA 3
4862 Chiquitita ABBA 3
5180 Super Trouper ABBA 3

Community: 12 | Size of Community: 61
Name Artist Number_of_Playlists
2171 We Belong Together Mariah Carey 35
3151 Touch My Body Mariah Carey 18
4330 Fantasy Mariah Carey 15
2181 Be Without You Mary J. Blige 13

4316 Honey Mariah Carey 10
615 Emotions Mariah Carey 9
4315 Heartbreaker Mariah Carey 8
4775 Hero Mariah Carey 8
4774 One Sweet Day Mariah Carey 7
5123 My All Mariah Carey 7
4593 It's Like That Mariah Carey 7
3612 Shake It Off Mariah Carey 6
4860 Don't Forget About Us Mariah Carey 5
5118 Without You Mariah Carey 5
5124 Can't Take That Away (Mariah's Theme) Mariah Carey 5
5115 Make It Happen Mariah Carey 4
5117 Dreamlover Mariah Carey 4
5122 Thank God I Found You Mariah Carey 4
5111 Vision of Love Mariah Carey 4
5126 The Roof Mariah Carey 3

Community: 28 | Size of Community: 61
Name Artist Number_of_Playlists
2048 Oceans (Where Feet May Fail) Hillsong United 45
2525 Good Good Father Chris Tomlin 26
4192 Lead Me to the Cross Hillsong United 20
3932 Multiplied NEEDTOBREATHE 20
4191 Touch The Sky Hillsong United 19
4927 Fix My Eyes for KING & COUNTRY 16
4642 How Can It Be Lauren Daigle 15
4232 Trust In You Lauren Daigle 15
4234 This Is Amazing Grace Phil Wickham 15
2526 How He Loves David Crowder Band 15
4892 No Longer Slaves (Live) Melissa Helser 14
4891 Shoulders for KING & COUNTRY 14
4589 God's Not Dead (Like a Lion) Newsboys 13
4587 You Make Me Brave (Live) Amanda Cook 13
3973 Flawless MercyMe 12
4890 It Is Well (Live) Kristene Dimarco 11
4888 Ever Be (Live) Kalley Heiligenthal 11
4961 First Lauren Daigle 11
2524 Just Be Held Casting Crowns 11
5039 Here as in Heaven Elevation Worship 10