Graphs report:

Part 1 - data acquisition:

Questions:

1. Provide the order and size of the graphs g_B and g_D:

   Graph g_B has an order of 476 and a size of 2000. in contrast, the graph g_D has an order of 466 nodes and a size of 1959 edges.

   a. Explain why, having explored the same number of nodes, the order of the two graphs (g_B and g_D) differs.

      The difference in the order (number of nodes) between the two graphs explored using Breadth-First Search (BFS) and Depth-First Search (DFS) arises from the inherent differences in how these algorithms traverse the graph.

      BFS explores all neighbors of a node before going to the next level. Instead, DFS explores as far as possible in each branch before backtracking. In other words, while BFS uses a queue method, DFS uses a stack.

      This difference in the adding method is then reflected in the node addition. So, in the case of DFS, if a max node is reached, the neighbors of that last explored node will be the last added. these are probably far from the root node. In the case of BFS, it might not reach as far down which means that the last node visited, as well as the majority of the nodes visited, won't be the same as in the case of DFS. This means that the number of nodes visited will be the same but the number of added nodes won't be. This causes the graphs to have different structure, order, and size.

   b. Justify which of the two graphs should have a higher order.

      BFS is designed to explore nodes level by level, meaning it will first explore all nodes directly connected to the starting node, then the nodes connected to those, and so on.

      Due to this exploration, BFS is likely to discover a larger number of nodes within the exploration limit, especially in graphs where nodes are densely connected.

      DFS might follow long chains of nodes, leading to fewer nodes being discovered within the same exploration limit, as it can quickly reach leaf nodes far from the starting node.

      This level-wise exploration ensures that more nodes are covered within the same limit, especially in denser graphs. DFS, on the other hand, is more prone to exploring long paths, which can result in fewer nodes.

   c. Explain what size the two graphs should have

      BFS's level-wise exploration captures more connections between nodes that are close to each other, resulting in a denser network. Overall it captures many local connections which means that in general, the graph explored by BFS should have a greater size.

2. Indicate the minimum, maximum, and median of the in-degree and out-degree of the two graphs (gB and gD). Justify the obtained values

G_B: minimum in-degree = 1 minimum out-degree = 0

maximum in-degree = 38 maximum out-degree = 20

median in-degree = 2 median out-degree = 0

G_D: minimum in-degree = 0 minimum out-degree = 0

maximum in-degree = 28 maximum out-degree = 20

median in-degree = 2 median out-degree = 0

G_B has a higher maximum in-degree and out-degree because BFS tends to discover more highly connected central nodes and nodes with a great number of outgoing connections. G_D results in fewer highly connected central nodes and generally fewer connections. However, both graphs have a similar median in-degree which means that there is a comparable level of node connectivity while the median out-degree remains 0 in both cases which might be a clue towards the prevalence of either leaf nodes or isolated ones.

3. Indicate the number of songs in the dataset D and the number of different artists and albums that appear in it.

   if we print the data frame, we will see there are 1060 songs (num of rows); 107 artists, and 615 albums in total.

   a. Justify why the number of songs you obtained is correct, considering the input graphs.

      The two graphs have Taylor Swift as a starting point and both visit 100 nodes. While one of them has 10 nodes more than the other, we could say the order is similar. Given that there are 1060 songs, it seems pretty valid. The list of graphs contains a unique graph which is the set of nodes that appear both in the BFS and DFS of Taylor Swift. Let's remember that the intersection has 107 artists and each artist has 10 top songs, which would give us around 1070 songs, which is very comparable to the number of songs we have in songs.csv. Minor discrepancies can arise from variations in the number of top songs per artist or specific data inclusion criteria, but 1060 is a reasonable number based on the input graphs.

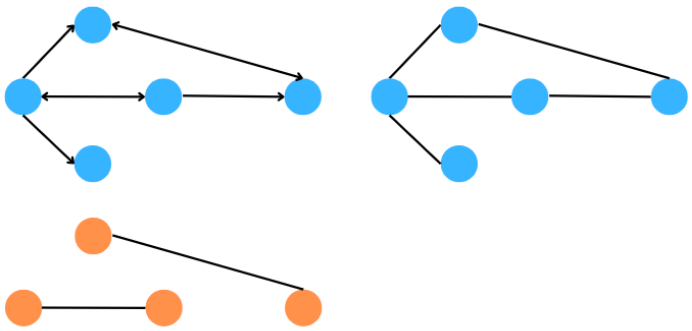
   b. Justify why the number of retrieved albums is correct.

      It should be correct given that no artist is inspected twice. Moreover, when the function retrieves information about the song, the album comes to the data frame from there. Each artist's songs typically come from multiple albums, and this diversity is reflected in the total album count. There shouldn't be any mistake regarding the number of albums.

Part 2 - data preprocessing:

Questions:

1. Justify whether the directed graphs obtained from the initial exploration of the crawler (gB and gD) can have more than one weakly connected component and strongly connected component, and explain why. Indicate the relationship with the selection of a single seed.
   When using a single seed artist, such as Artist A, the crawler explores connections, initially adding outgoing edges from node A. At this stage, all nodes form a weakly connected component, with no bidirectional edges present. However, bidirectional edges may emerge as the crawler visits neighbors. If a neighbor of A is related to A, a bidirectional edge between them is established. The count of weakly connected components typically aligns with the number of neighbors of visited nodes. Conversely, the number of strongly connected components correlates with the total number of visited nodes and the presence of bidirectional connections among them, illustrating the graph's evolving structure during exploration. All nodes that have not been visited, won't be part of a strongly connected component.
2. Can the number of connected components in the undirected graphs (g'B and g'D) be higher than the number of weakly connected components of its respective directed graph (gB and gD)? Provide a minimal example to showcase your answer.
   In the original graphs, there is one weakly connected component. While, when the original graph, only those bidirectional edges are kept. This results into multiple connected components in the undirected graphs.
   In the example below, the original graph is the one with blue nodes and directed edges. The processed graph (which contains only the bidirectional edges) has more components than the directed graph has weakly connected components (graph with blue nodes and undirected edges).



3. Generate a preliminary report from the undirected graph with weights (gw).
   a. Which are the two most (respectively, least) similar artists? What graph attribute allows you to answer this question?
      The undirected graph with weights has been computed with the Euclidean distance so the values are smaller than the ones the graph would have if we used the cosine similarity.
      The least similar artists are "Girl in red" and "Johnny Orlando" with a similarity score of 0.01919350828056299 (euclidean distance). We have found this by looking for the smallest weight in the undirected graph of similarities.

b. Which is the artist most (and least) similar to all the other artists in the network? What graph attribute allows you to answer this question?

The most similar artists to all other artists is JoJo

The least similar artists to all other artists is girl in red

These have been found by summing all the weights of each node and printing the min and max to know the most similar and the least similar.

Part 3 - Data analysis:

Questions:

1. Study the number of common nodes between the obtained graphs. Use the function num common nodes.
    a. How many nodes are shared between gB and gD? What information does this tell us about the importance of the algorithm used by the crawler (i.e. the scheduler) to decide next nodes to crawl? The small number of 122 common nodes out of approximately 2000 between the graphs suggests distinct exploration strategies. Depth-First Search (DFS) prioritizes depth, potentially missing nodes in other areas, while Breadth-First Search (BFS) explores nodes at similar distances more evenly. DFS may spend more time exploring deep branches, whereas BFS covers more ground initially. These differences highlight how exploration strategies shape the crawler's reach and the nodes it prioritizes in the graph.
    b. How many nodes are shared between gB and g ′ B? What information does this tell us about the reciprocity of gB? And about the Spotify's artist related algorithm?
    They have 99 common nodes. Since $g'B$ keeps only bidirectional edges, the presence of 99 common nodes suggests that these nodes participate in mutual relationships with other nodes in gB. This implies a certain level of reciprocity in gB, as these 99 nodes have at least one mutual connection.
    The existence of bidirectional relationships in gB suggests that Spotify's artist-related algorithm considers mutual artist similarities or connections. This could mean that if Artist A is related to Artist B, there is a likelihood that Artist B is also related to Artist A, reflecting a symmetric relationship. The mutual connections might be based on shared listener behaviors, similar genres, collaborations, or other factors that make the relationship bidirectional.
    Spotify's algorithm might emphasize bidirectional relationships when defining artist-relatedness, potentially giving more weight to mutual connections.
2. Calculate the 25 most central nodes in the graph g'B using both degree centrality and betweenness centrality. How many nodes are there in common between the two sets? Explain what information this gives us about the analyzed graph.
    There are 9 common nodes in the sets of the top 25 most central according to betweenness centrality and degree centrality. The disparity in the number of common nodes (9 out of 25) highlights differences in how degree centrality and betweenness centrality assess node importance. Nodes with high degree centrality might not necessarily have high betweenness centrality, and vice versa. This suggests a diverse set of influential nodes in the network, with some nodes being important due to their sheer connectivity, while others are crucial for mediating communication pathways. Many of the influential nodes have high connectivity and low communication pathways, and vice versa.
3. Find cliques of size greater than or equal to min size clique in the graphs g ′ B and g ′ D. The value of the variable min size clique will depend on the graph. Choose the maximum value that generates at least 2 cliques. Indicate the value you chose for min size clique and the total number of cliques you found for each size. Calculate and indicate the total number of different nodes that are part of all these cliques and compare the results from the two graphs.
    We chose 7 as the value 8 gave 0 cliques. we found 4 cliques in g'B and 5 cliques in g'D.

For graph gB there are 18 different nodes that are part of cliques while for graph gD there are 17 different nodes that are part of the found cliques. Given that both results are pretty similar, we could infer that they are similar in structure. Both have more or less the same amount of cliques that have 7 or more nodes.

4. Choose one of the cliques with the maximum size and analyze the artists that are part of it. Try to find some characteristic that defines these artists and explain it.

   Steinfeld, Little Mix, and Fifth Harmony are all prominent figures in the pop genre, known for their catchy melodies and broad appeal. These female or female-fronted acts resonate strongly with younger audiences, addressing themes of love, empowerment, and self-discovery. They have achieved significant commercial success, often topping charts like the Billboard Hot 100, and maintain robust social media presences to engage with fans. Frequently collaborating with each other and other artists, they are recognized for their dynamic stage performances and have received numerous awards and nominations. Additionally, they advocate for social issues such as gender equality and mental health, promoting messages of empowerment and self-confidence through their music.

5. Detect communities in the graph gD. Explain which algorithm and parameters you used, and what is the modularity of the obtained partitioning. Do you consider the partitioning to be good?

   We have used 'louvain' and got a modularity of 0.5405473372781066 indicates a moderate level of community structure in the network. It is a pretty good partitioning taking into account that the range of modularity is [-1, 1].

6. Suppose that Spotify recommends artists based on the graphs obtained by the crawler (gB or gD). While a user is listening to a song by an artist, the player will randomly select a recommended artist (from the successors of the currently listened artist in the graph) and add a song by that artist to the playback queue.

   a. Suppose you want to launch an advertising campaign through Spotify. Spotify allows playing advertisements when listening to music by a specific artist. To do this, you have to pay 100 euros for each artist to which you want to add ads. What is the minimum cost you have to pay to ensure that a user who listens to music infinitely will hear your ad at some point? The user can start listening to music by any artist (belonging to the obtained graphs). Provide the costs for the graphs gB and gD, and justify your answer.

      We need the total number of strongly connected components which don´t have an exit to other nodes. The easiest examples are leaf nodes and loops. Once the user gets there, it cannot get out.

      For graph gB, there are 374 strongly connected components (373 are leaves). This would result in 37400€ in total.

      In the case of graph gD, there are 520 strongly connected components (507 of which are leaves), this would bring the total to 52000€

   b. Suppose you only have 400 euros for advertising. Which selection of artists ensures a better spread of your ad? Indicate the selected artists and explain the reason for the selection for the graphs gB and gD.

      The artists that ensure a better spreading are the ones that have a higher indegree. Given that they have the highest probabilities of a user ending up listening a song of this artist and thus the add we want to place.

Given that the 4 artists with highest indegree of gB are Bebe Rexha, Hailee Steinfeld, Demi Lovato, and Fifth Harmony, it would be intelligent to pay them. The same goes for the top artists with highest indegree of gD: Ethynol, adrenachrome, Ld-50 (metal), Aisling.

7. Consider a recommendation model similar to the previous one, in which the player shows the user a set of other artists (defined by the successors of the currently listened artist in the graph), and the user can choose which artist to listen to from that set. Assume that users are familiar with the recommendation graph, and in this case, the gB graph is always used.
    a. If you start by listening to the artist Taylor Swift and your favorite artist is THE DRIVER ERA, how many hops will you need at minimum to reach it? Give an example of the artists you would have to listen to in order to reach it.

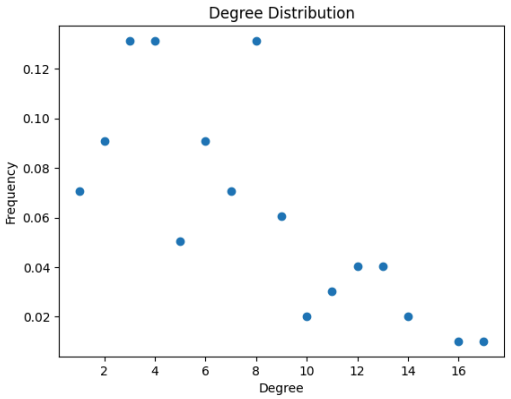
    It takes 3 hopes to go from Taylor Swift to THE DRIVER ERA. One possible path is:
    Taylor Swift -> Olivia Rodrigo -> Joshua Bassett -> THE DRIVER ERA

Part 4 - Data visualization:

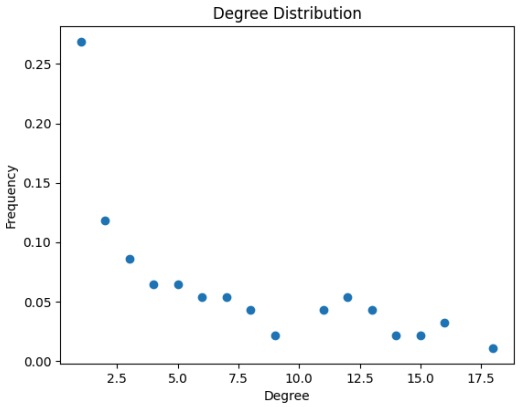
1. Comment on the results obtained in Exercise 4 (it is compulsory to include the results obtained in Exercise 4 in the report):
    a. What are the degree distributions of the three obtained undirected graphs like?

    All three distributions have been plotted using normalisation and with the normal scale (without using the log-log scale).

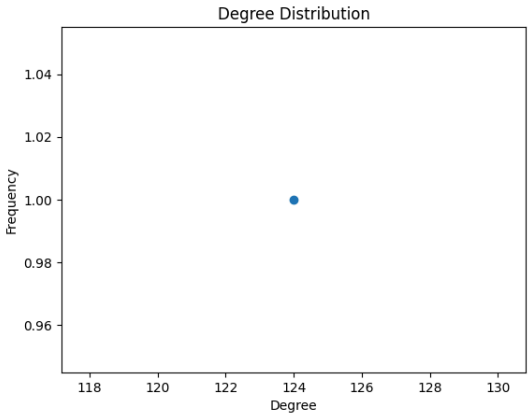
    The degree distribution of "Undirected gB" is the following: {1: 7, 8: 13, 9: 6, 11: 3, 13: 4, 5: 5, 6: 9, 3: 13, 7: 7, 4: 13, 12: 4, 14: 2, 2: 9, 17: 1, 10: 2, 16: 1}



    The degree distribution of "Undirected gD" is the following: {1: 25, 3: 8, 2: 11, 5: 6, 4: 6, 7: 5, 6: 5, 8: 4, 11: 4, 14: 2, 16: 3, 13: 4, 15: 2, 12: 5, 9: 2, 18: 1}



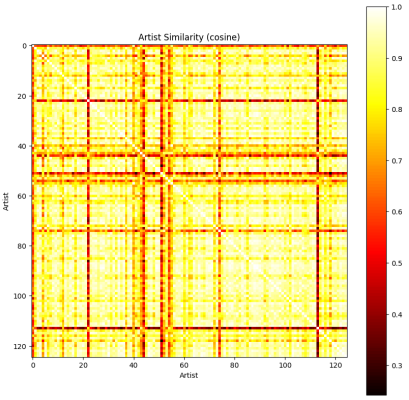
    The degree distribution of "Undirected gW" is the following: {124: 125}



b. What can you infer from the similarity heatmap regarding the algorithm that selects related artists on Spotify?

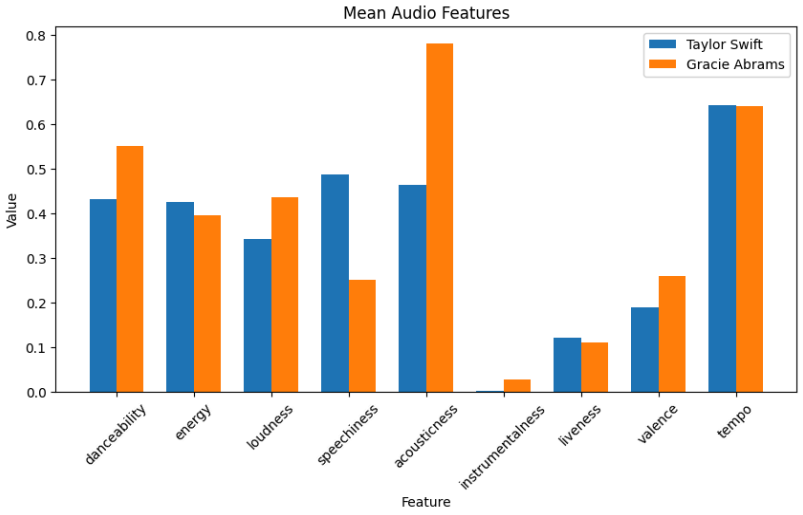
At first glance, what is clear is that the white diagonal (when the similarity is calculated with cosine) represents similarity with one's self.

Groups of artists with high similarity in the heat map represent artists with similar features in their songs (pitchiness, danceability, tempo, speechiness, valence…).

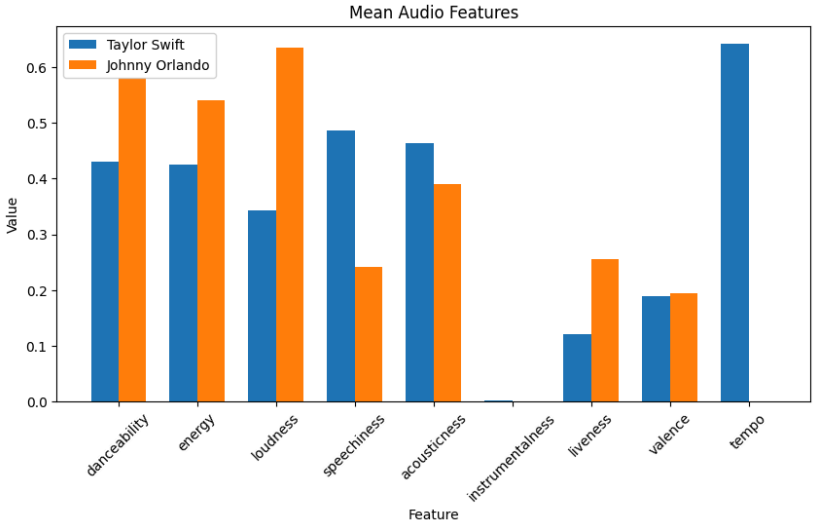


c. Is there any relationship between the similarity of artists obtained from their audio features and the distances of the artists in the directed graphs? For instance, consider Taylor Swift and her most and least similar artists as determined in exercises 4.b and 4.c.

The most similar artist with Taylor Swift is Gracie Abrams:



While the least similar is Johnny Orlando:



The number of neighbors between Taylor Swift and Gracie Abrams is 2 in both gD and gB, and it's 3 and 4 between Taylor Swift and Johnny Orlando. Since the features for each artist come from a graph containing only the common artists between gB and gD, the longest path starting at Taylor Swift has a length of 7.

Since the neighbors depend on factors like collaborations, user interactions… connected nodes do not need to be similar.

In this case, the most similar artist to Taylor Swift is closer to her than the least similar artist.

d. At which percentile would you prune the edges of the weighted similarity graph g w to ensure the size of the largest connected component is preserved while minimizing the amount of edges in the graph?

When pruning with a min percentile of 69.6, the biggest connected component still has all 125 nodes. However, then increasing this value by 0.1 (that is, when min_precentile = 69.7), the number of nodes in the biggest connected component decreases to 124.

2. Comment on the visualizations generated with Gephi.

a. Compare graphs gB and gD. What can you say about their properties?

The BFS graph (gB) tends to exhibit a more evenly distributed structure, exploring nodes at similar depths simultaneously and likely resulting in denser connections among neighboring artists. In contrast, the DFS graph (gD) may display longer paths between Taylor Swift and related artists, as DFS explores one path deeply before backtracking, possibly leading to more isolated clusters of artists.

b. Can you identify common characteristics among artists belonging to the same community? Could you label the different communities?

After exporting the node information in a CSV, we first grouped by modularity class (as it can be seen in the notebook), and tried identifying common genres and centrality measures.

DFS measures:

	modularity_class	count	common_genres	mean_closeness_centrality	mean_betweenness_centrality	mean_harmonic_closeness_centrality
0	0	99	[pop]	0.005049	109.656566	0.010880
1	1	86	[post-teen pop]	0.008010	414.802326	0.015525
2	2	34	[idol]	0.005689	422.382353	0.010225
3	3	40	[hollywood]	0.008695	556.650000	0.014753
4	4	49	[gymcore]	0.009466	321.227891	0.014535
5	5	50	[gymcore]	0.019472	292.849899	0.030514
6	6	109	[gymcore, hard alternative]	0.043148	421.938884	0.060607
7	7	30	[alternative metalcore, rap metalcore]	0.021659	461.200000	0.028417
8	8	110	[rap metalcore]	0.152964	184.090909	0.194540

BFS measures:

	modularity_class	count	common_genres	mean_closeness_centrality	mean_betweenness_centrality	mean_harmonic_closeness_centrality
0	0	68	[pop]	0.150404	559.099188	0.175221
1	1	129	[pov: indie]	0.050420	324.942881	0.058882
2	2	46	[post-teen pop]	0.043198	284.594775	0.050644
3	3	64	[dance pop, pop]	0.067648	209.055855	0.081527
4	4	15	[alternative rock, art pop, dream pop, etherea...	0.010584	95.009511	0.013841
5	5	67	[alt z]	0.028241	134.796225	0.033280
6	6	31	[pop]	0.037533	214.239373	0.043556
7	7	53	[movie tunes]	0.021492	107.788281	0.024831

Community 0:

Common Genres: Pop

Size: 99 artists

Centrality Measures:

Mean Closeness Centrality: 0.005

Mean Betweenness Centrality: 109.657

Mean Harmonic Closeness Centrality: 0.011

Community 0 has artists that primarily belong to pop genre. Some examples of artists in this community are Taylor swift, Selena Gomez, Olivia Rodrigo, Katy Perry, Arian Grande, Harry Styles… These are all big names that surely, hover over other genres, but their main genre is pop. Even if the artists are highly popular, the mean centrality measures of the community are very low (the lowest among the communities). That is probably due to the fact that these values come from the DFS graph, so the nodes are visited in an order that affect this. If we look at the centrality measures with the BFS graph, we will se how this community has the highest.

Community 1:

Common Genres: Post-teen pop

Size: 86 artists

Centrality Measures:

Mean Closeness Centrality: 0.008

Mean Betweenness Centrality: 414.802

Mean Harmonic Closeness Centrality: 0.016

Community 1 artists are centered around the post-teen pop genre, typically appealing to younger audiences. Their genres include “movie tunes”, “pop”, “dance pop”... but most of them are “post teen pop”.

Some of the artists are: Olivia Holt, Zendaya, Ross Lynch, Austin Moon, Rocky, Debby Ryan, China Anne McClain, Jordan Fisher… who happen to be disney actors from Tv shows that involved actors singing some of the songs. This could be for example the main theme song. A clear example are the artists Ross Lynch and Austin Moon which appear as singers of this community. Austin Moon is a fictional character from a disney TV show played by Ross Lynch and he sings most of the songs in the show.

Community 2:

Common Genres: Idol

Size: 34 artists

Centrality Measures:

Mean Closeness Centrality: 0.005

Mean Betweenness Centrality: 422.382

Mean Harmonic Closeness Centrality: 0.010

Common Characteristics: Focused on the idol genre. Some of the artists are David Cook, Kris Allen, Jessica Sierra… Their genres are “idol”, “neo mellow”, “talent show”, “candy pop”, “canadian country” and some others. But mainly, all of them have “Idol” as their primary genre.

Community 3:

Common Genres: Hollywood

Size: 40 artists

Centrality Measures:

Mean Closeness Centrality: 0.008

Mean Betweenness Centrality: 556.650

Mean Harmonic Closeness Centrality: 0.015

Community 3: This community is associated with Hollywood,, with centrality measures similar to the idol community.

Some of the artists are: Paul Carella, Jason Manns, Stewart Mac… These happen to be less popular artists, especially compared to the prior communities. In the list, we can also see names like Scarlett Johanson who is on the podium of popularity of this community. She is a clear example of this community which contains people who might not primarily be a singer but rather an actor of Hollywood. Another example is Robert Downey Jr. or Johnny Depp. These are celebrities who appear in some Hollywood track and still get credit and appear as artists in Spotify

Community 4:

Common Genres: Gymcore

Size: 49 artists

Centrality Measures:

Mean Closeness Centrality: 0.009

Mean Betweenness Centrality: 321.228

Mean Harmonic Closeness Centrality: 0.015

Community 4 : Artists in this community belong to the gymcore genre, indicating a focus on high-energy music suitable for workouts, with centrality measures comparable to the Hollywood community.

Some of the artists are: Saliva, Pop Evil, From Ashes To New, Onlap… Since we are not particularly familiar with any of the artists, we can't really give a qualitative analysis as to why they might be a community or why they are different from other communities, but all of them seem to have this same genre and a medium to low popularity [290 to 47500 followers].

Community 5:

Common Genres: Gymcore

Size: 50 artists

Centrality Measures:

Mean Closeness Centrality: 0.019

Mean Betweenness Centrality: 292.850

Mean Harmonic Closeness Centrality: 0.031

Community 5: Similar to Community 4, with a focus on gymcore music, suggesting a possible sub-division within this genre, with very similar centrality measures. Some of the artists are: Downplay, Fivefold, Leader…

Community 6:

Common Genres: Gymcore, Hardcore
Size: 109 artists
Centrality Measures:
Mean Closeness Centrality: 0.043
Mean Betweenness Centrality: 421.939
Mean Harmonic Closeness Centrality: 0.061

Community 6: This community includes both gymcore and hardcore genres, showing a mix of high-energy and intense music, with slightly higher centrality measures indicating better connectivity within the network. They are also mixed with genres like “alternative”, “hard alternative”, “heavy alternative”... Some of the artists are: Earshot, Ra, Through Fire…

Community 7:

Common Genres: Alternative Metal
Size: 30 artists
Centrality Measures:
Mean Closeness Centrality: 0.022
Mean Betweenness Centrality: 461.2
Mean Harmonic Closeness Centrality: 0.028

Common Characteristics: Focused on alternative metal, this community has a distinct genre that sets it apart from the others. Some of the artists are: Switched, From Zero, Lifer, Pressure 4-5

Community 8:

Common Genres: Rap Metalcore
Size: 110 artists
Centrality Measures:
Mean Closeness Centrality: 0.153
Mean Betweenness Centrality: 184.091
Mean Harmonic Closeness Centrality: 0.195

Community 8: This community combines rap and metalcore, showing a blend of aggressive and rhythmic music styles, with the highest centrality measures among the communities, suggesting that the artists of this community are situated in the middle of the graph. Some of the artists are C-engine, 9th Corner, Kongcrete, Calm Chaos… Their popularity is low [41 to 2912 followers].

Distinguishing Factors

The genre is the primary differentiator among the communities; each community is centred around specific genres of sub-genres of music. The number of artists in each community can affect its influence and characteristics.

Looking into the main artists of each community provides a more contextual understanding of the differences between the communities and the similarities within those.

Given that the number of communities depends on the resolution parameter, adjusting this parameter allows for the merging of similar communities or the creation of more communities. This process would enhance the distinction between nodes belonging to different communities while increasing the similarity among nodes within the same community.