Miguel Arpa Robig - 1668833
Andreu Gascón Marzo - 1670919

Project Graphs - Analysis of Music Networks with Spotify

Part 1: Acquisition and Storage of Data

1. **Provide the order and size of the graphs gB and gD.**
   Order gB: 472, Size gB: 2000
   Order gD: 665, Size gD: 1888
   a. Explain why, having explored the same number of nodes, the order of the two graphs (gB and gD) differs.

      ➔ ***Response****:* BFS explores the graph level by level, potentially finding a broader set of nodes quickly, but it might revisit nodes more frequently while DFS explores as far down a branch as possible before backtracking, which can lead to discovering more 'unique' nodes because it follows longer paths first.

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

      ➔ ***Response****:* It's more likely to be the DFS because, as mentioned, BFS is more prone to finding repeated (previously seen) nodes because of the similarity between nodes.

   c. Explain what size the two graphs should have.

      ➔ ***Response:*** BFS explores neighbouring nodes at each level before moving deeper, which allows more edged connecting into neighbouring nodes of the same level with more frequency. That's why is more prone for BFS to have more size (edges) than in DFS

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

   ➔ ***Response:***
   gB -> In-Degree: {'min': 1, 'max': 38, 'median': 2.0}, Out-Degree: {'min': 0, 'max': 20, 'median': 0.0}
   gD -> In-Degree: {'min': 0, 'max': 24, 'median': 2.0}, Out-Degree: {'min': 0, 'max': 20, 'median': 0.0}

   The minimum in-degree and out-degree equal to 0 indicate that some nodes have no incoming or outgoing edges, respectively. The maximum values show the nodes with the highest connectivity talking of incoming and outgoing relationships.

The median indicates the most common connectivity for nodes in each graph. The median out-degree being 0 in both graphs suggests that many nodes do not have outgoing edges, which could indicate that they are terminal nodes in the traversal process.

3. **Indicate the number of songs in the dataset D and the number of different artists and albums that appear in it.**
   The dataset contains 1244 songs, there are 127 unique artists in the dataset, and there are 1164 unique albums in the dataset
   a. Justify why the number of songs you obtained is correct, considering the input graphs.

      ➔ ***Response:*** Assuming the order of the graphs is 472 (BFS) and 665 (DFS) and we are getting the intersection of both graphs (127 artists) and their corresponding top songs, 1244 seems reasonable.

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

      ➔ ***Response:*** 1164 albums seem reasonable because normally top listened tracks come from different albums of a given artist.

Part 2: Data Preprocessing

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.**

   ➔ ***Response:***
   Nº of weakly connected components in gB: 1
   Nº of weakly connected components in gD: 1
   Nº of strongly connected components in gB: 373
   Nº of strongly connected components in gD: 577

   Both graphs start from a single seed and by the nature of the artist's Spotify graph (where some path relates all artists), all of them appear to be connected by an underlying graph, hence, they should form a single weakly connected component.
   To have a unique strongly connected component, the graph requires a direct path between all pair nodes in the whole graph. That could be the case, but is not the most probable thing since in strongly connected components we consider the direct edges between nodes. These direct edges can form clusters with various strongly connected components.

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.**

➔ ***Response:***
Number of connected components in gBp: 1
Number of connected components in gDp: 14

No, it can never be higher because a weakly connected component essentially ignores the direction of the edges assuming underlying bidirected graphs.
It can at most be the same.

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?

➔ ***Response:***
Most similar artists:
Vanessa Hudgens and Zac Efron, score: 0.999

Least similar artists:
Victorious Cast and Jeremy Zucker, score: 0.995

Most Similar Artists: Vanessa Hudgens and Zac Efron share the highest score. We looked for the edge with the highest value.

Least Similar Artists: Victorious Cast and Jeremy Zucker have the lowest score among all pairs. We looked for the edge with the lowest value.

The weight attribute of the edges is critical here as it quantifies the similarity between artists.

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?

➔ ***Response:***
Artist most similar to others:
Hailee Steinfeld with an average score of 0.999
Artist least similar to others:
Jeremy Zucker with an average score of 0.997

Most Similar to Others: Hailee Steinfeld has the highest average similarity score to other artists with 0.99.

Least Similar to Others: Jeremy Zucker has the lowest average similarity score at 0.997.

We calculated the sum of the weights of edges connected to each node.

Miguel Arpa Robig - 1668833
Andreu Gascón Marzo - 1670919

Part 3: Data Analysis

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?

        ➔ ***Response:*** The number of shared nodes is 127. Although both have plenty of shared nodes, due to the nature of the BFS and DFS algorithms, BFS tends to explore level by level whereas DFS explores as far as possible along each branch before backtracking.

    b. How many nodes are shared between gB and g′ B? What information does this tell us about the reciprocity of gB? And about Spotify's artist related algorithm?

        ➔ ***Response:*** There are only 99 shared nodes. This suggests that a huge amount of relationships are unidirectional. Spotify's artist related algorithm doesn't allow much bidirectionality and reciprocity between artists. This may allow users to not get stuck in similar artists and explore more distinct genres and types of music.

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.**

        ➔ ***Response:*** We found 9 nodes in the intersection between top 25 degree centrality and top 25 betweenness centrality. This may indicate that the most influential nodes (with a higher degree of centrality) are not necessarily the ones more influential regarding the spread of information or connectivity across the network.

3. **Find cliques of size greater than or equal to min graphs g′ B and g′ D. The value of the variable min size clique in the 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.**

        ➔ ***Response:***
        Max clique size for gBp that generates at least 2 cliques: 7
        Total number of cliques in gBp: 4
        Total number of different nodes in all cliques in gBp: 18

        Max clique size for gDp that generates at least 2 cliques: 17

Total number of cliques in gDp: 2
Total number of different nodes in all cliques in gDp: 19

Total number of common nodes in all cliques in gBp and gDp: 0

The result shows a high difference between cliques from the gBp and the gDp. In the gDp graph, we find maximal cliques of at least length 17 whereas in the gBp only of length 7. This seems a logical thing taking into account the different search strategies of both DFS and BFS algorithms as we will see in further Gephi plots.
The lack of common nodes between both graphs suggests that artists in the cliques are quite distinct.

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.**

   ➔ ***Response:*** Selecting the first clique of maximum size we can see that artists from the clique are from the Pop branch. Based on the mean statistics we can see that they share in their tracks important aspects of pop music: High energy levels, high loudness, positive words, a moderately high tempo typical from pop music etc.
   These are some the shared stats:

	Artist Name	Energy	Loudness	Valence	Tempo
14	Niall Horan	0.5997	-5.7661	0.4371	100.9902
49	Little Mix	0.7339	-4.5959	0.5967	107.3353
75	Louis Tomlinson	0.7693	-4.5679	0.5438	123.5327
81	5 Seconds of Summer	0.6887	-4.5986	0.4337	131.7788
90	Liam Payne	0.7353	-5.0232	0.5731	111.9054
122	The Vamps	0.8389	-4.9239	0.7018	116.1179
123	Jonas Brothers	0.7686	-5.0639	0.7624	117.4044

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?**

   ➔ ***Response:*** We used the Louvain algorithm since it is more suitable for large graphs. The modularity obtained is 0.82 which is a good indicator of dense connections inside the community and fewer with other communities.

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.

- ➔ ***Response:*** To ensure that a user who listens to music infinitely will hear your ad at some point, you need to place ads on a set of artists such that every strongly connected component (SCC) of the graph is covered. Every node in a SCC has a path to another node of the SCC, which means that after some time, it will listen to all the artists from the SCC.
  The minimum cost for gD will be 100*577 (SCC) and for gB 100*373 (SCC)

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.

- ➔ ***Response:*** Given a limited budget and that we want the maximum possible spread, it could be useful to set the ads on the artists with the most outdegree centrality (artists with higher out links), but the outdegree centrality is the same for all artists in the Soptify system, hence, a good measure could be to find the nodes with higher betweenness centrality. Given a 400 euros budget, we can only set ads for 4 artists. The 4 artists with more betweenness centrality for gD are:

  Artist ID: 4Uc8Dsxct0oMqx0P6i60ea, Artist Name: Redstone Records
  Artist ID: 3WGpXCj9YhhfX11TToZcXP, Artist Name: Heiko Wolz
  Artist ID: 5cj0lLjcoR7YOSnhnX0Po5, Artist Name: Mr Miln
  Artist ID: 1McMsnEElThX1knmY4oliG, Artist Name: Matthias

  And for gB:
  Artist ID: 4Uc8Dsxct0oMqx0P6i60ea, Artist Name: Conan Gray
  Artist ID: 3WGpXCj9YhhfX11TToZcXP, Artist Name: Sabrina Carpenter
  Artist ID: 5cj0lLjcoR7YOSnhnX0Po5, Artist Name: Doja Cat
  Artist ID: 1McMsnEElThX1knmY4oliG, Artist Name: Troye Sivan

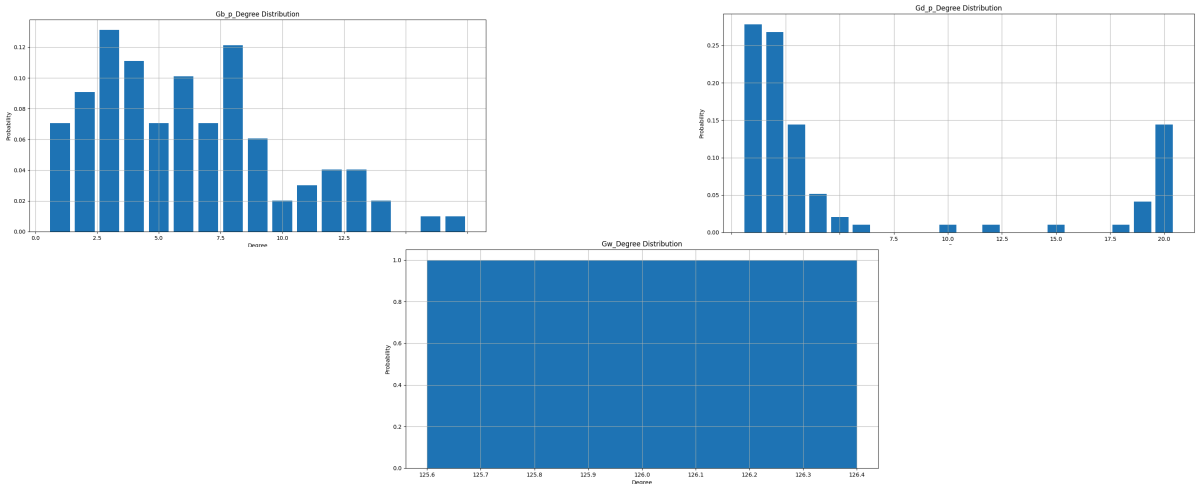
7. **Consider a recommendation model similar to the previous one, in which the player shows the user a set of other artists (defined by the successors of the currently listened artist in the graph), and the user can choose which artist to listen to from that set. Assume that users are familiar with the 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 favourite 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.

➔ ***Response:*** If we want to go from Taylor Swift to THE DRIVER ERA we need to do 3 Hops across the following artists:

1. Taylor Swift
2. Olivia Rodrigo
3. Joshua Bassett
4. 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?



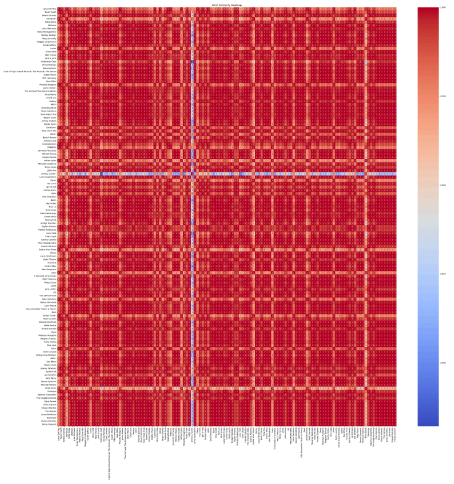
➔ ***Response:***

gB'p: The most common degrees, which are the peaks of the distribution are 3 and 8. There are fewer nodes with very high degrees.

gD'p: This graph displays a bi-modal degree distribution, where there are two prominent peaks. The first set of peaks occurs at low degrees (1 to 3), and a second peak at a degree of 20. There's an absence of nodes with intermediate degrees (between 5 and 18).

gW: All nodes in the graph have a degree of 126. The reason for this number is the number of artists in the data frame, which is 126 artists, all connected by edges that have a weight between them.

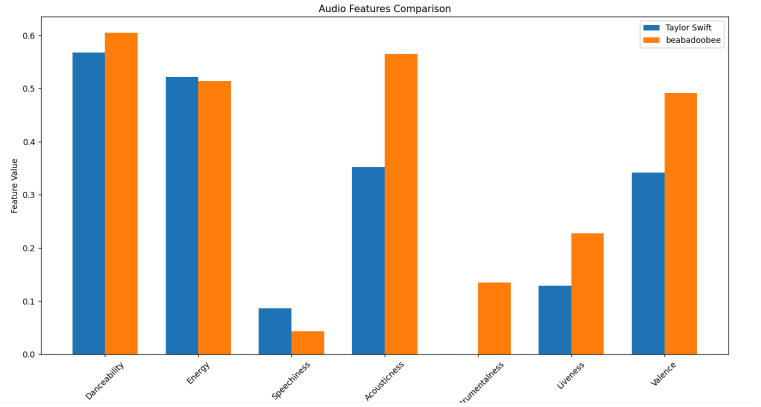
b. What can you infer from the similarity heatmap regarding the algorithm that selects related artists on Spotify?

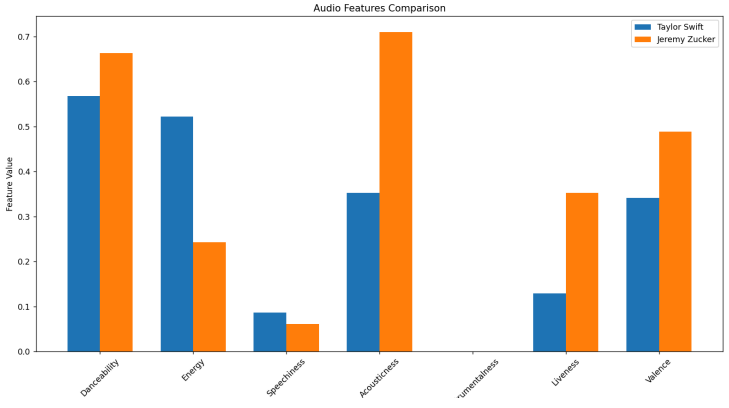


➔ ***Response:*** The similarity heatmap explains that Spotify's algorithm utilizes a variety of audio features to assess similarity. Each featured artist is created from many songs. This ensures that similarity assessment is robust and captures the style of each artist rather than being influenced by single tracks.

The heatmap shows clusters of artists, this means that Spotify groups artists into genres based on their sound profiles. So cool colors between some groups suggest that the algorithm differentiates well.

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.





➔ ***Response:*** Taylor Swift with Beabadoobee and Jeremy Zucker, which represent her most and least similar artists respectively, according to some audio feature metrics.

Taylor Swift and Beabadoobee:
Both artists show closer values in danceability, energy, and valence.
There are notable differences in speechiness and instrumentalness.

Taylor Swift and Jeremy Zucker:
Their energy and valence are somewhat aligned.
Differences appear in danceability, speechiness, and acousticness, pointing to distinct styles where Jeremy Zucker may feature more acoustic and less dance-oriented music compared to Taylor Swift.
There is a relationship between similarity and the distance of the artists in the graph. The artists who show a great correlation in the heat map also have a short distance in the graph, indicating in both cases a relationship between artists. The same thing happens with artists who do not have a relationship, both in the heatmap and in the graph, they will have a low correlation that will be related to the long distance in the graph.

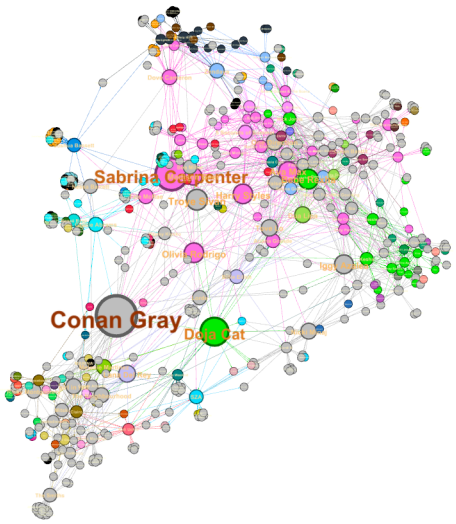
d. At which percentile would you prune the edges of the weighted similarity graph gw to ensure the size of the largest connected component is preserved while minimizing the amount of edges in the graph?

➔ ***Response:*** (empty)

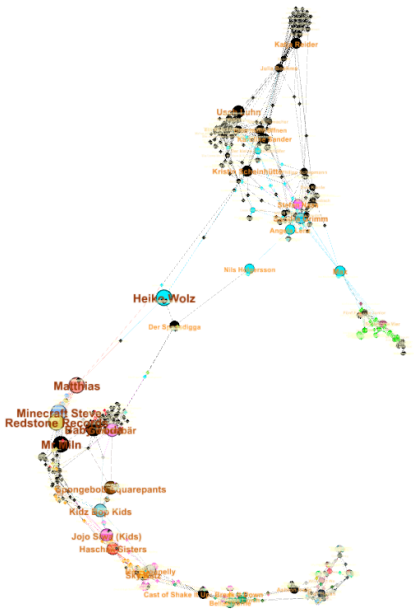
For each of the directed graphs, gD and gB, obtained in the first session of the practice:

2. **Generate a visualization of the graph using Gephi that assigns a color to the nodes based on the community they belong to and sizes the nodes proportionally to their betweenness centrality. Use a layout algorithm that allows for easy identification of the communities. Show the names of the most important artists (highest betweenness centrality) in each community.**

gB graph:



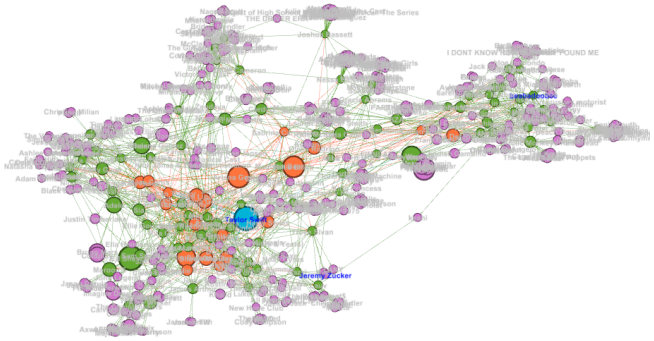
gD graph:



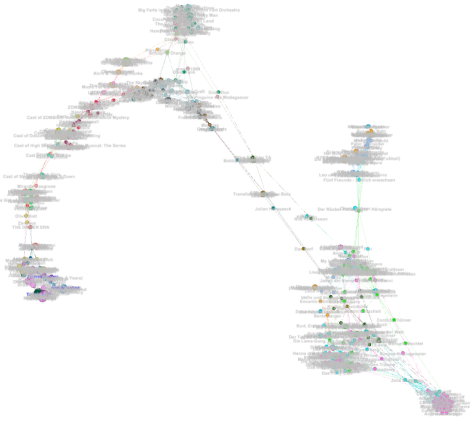
The graph on the left is graph gB and the graph on the right is graph gD. As you can see, the aforementioned configurations have been made. The 'Force Atlas' algorithm has been used for the distribution of nodes.

3. **Generate a visualization of the graph using Gephi while maintaining the same node positioning as the previous exercise's graph, but now assigning node size based on their number of followers and node color based on the distance of each node from the initial node of the crawler (the node representing the artist Taylor Swift). Highlight the two artists selected for the plot audio features comparison (the less and most similar artists to Taylor Swift).**

gB graph:



gD graph:



The graph on the left is graph gB and the graph on the right is graph gD. As you can see, the aforementioned configurations have been made. The 'Force Atlas' algorithm has been used for the distribution of nodes. In both graphs the color represents the distance between a node and the seed node (Taylor Swift), as can be seen in the first graph there are only 3 colors that indicate distance and in the other there are more than 25 different colors. The names of the graph gD do not look very good but they are located at the bottom left.

4. **Comment on the visualizations generated with Gephi.**
    a. Compare graphs gB and gD. What can you say about their properties?
        ➔ ***Response:*** In the first graph, the largest nodes, such as Conan Gray, Sabrina Carpenter, and Doja Cat, appear to be the most central in their respective communities.
        In the second graph, the distribution is more dispersed and less dense than in the graph gB.

        The structure of gB shows a denser network with several connections between communities, this means greater interconnectivity between artists, a more robust and cohesive network.
        The structure of gD is more elongated and dispersed, with less dense connections between artists. This means a more fragile or less cohesive network in terms of interconnectivity.

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

➔ ***Response:*** Artists in the same community in the network graph belong to similar genres or subgenres. The problem is that when we collected information about each artist through Spotify, we collected all the genres of the singer and not the most relevant one (the first). Because of this error, it does not do it correctly since in the genre attribute, there are genres of genres.

We have made the code part in a Github repository (it contains all the parts and graphs), the link is public: https://github.com/Miguel231/ProjectGraphs