Session 2 Questions

1. **Provide the order and size of the four obtained undirected graphs (g ′B, g ′D, g w B, and g w D).**

	Order	Size
g ′B	189	488
g ′D	192	931
g w B	177	435
g w D	186	994

2. **Justify the strategy used to obtain g w B and g w D. 3. (0.5 points)**

The strategy used to obtain a weighted graph (e.g., B) from a given data frame is as follows:

1. Obtain a dataframe from the function get_track_data with input graph B. This function retrieves relevant data about songs from the artist associated with the graph.
2. Once the data frame containing the songs of the artist in graph B is obtained, the function compute_means_audio_features is used. This function calculates the mean audio features for the artist's songs in the graph. It returns a pandas dataframe containing the artist's mean features for the graph.
3. The weighted graph B is obtained using the function create_similarity_graph. This function takes the data frame obtained in the previous step as input and constructs a weighted graph based on similarity measures between songs. The specific details of how the similarity is computed and translated into graph edges may depend on the implementation of the function.
4. After obtaining the weighted graph, the edges that have weights below a specific threshold are pruned. This is done using the prune_low_weight_edges function, where the argument min_weight is set to determine the minimum weight below which edges are removed. For example, in the case of weighted graph B, the min_weight value is set to 0.31.
5. Finally, the resulting weighted graph B contains nearly the same number of edges as the original undirected graph.

The same strategy is applied to obtain the weighted graph D, following the steps outlined above with the corresponding input data frame for graph D and using for the min_weight argument of the prune_low_weight_edges function a value equal to 0.31.

3. **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.**

The presence of multiple weakly connected components in the digraphs gB and gD is not possible. This is due to the fact that the graphs were expanded from a single seed, and each node was connected to its corresponding artist and their friends. Consequently, all nodes in the graph have an edge connecting them back to the seed node.

As a result of this expansion process, all vertices in the underlying graph are interconnected, ensuring that any node can be reached from any other node through a path. Therefore, there exists a path between every pair of vertices, indicating that all vertices are connected by some path. This implies that the weakly connected component encompasses the entire graph, as all nodes are part of the same interconnected component.

In summary, the underlying graphs gB and gD cannot have more than one weakly connected component because the expansion process ensures that all vertices are connected, allowing for a path between any two nodes. Thus, the weakly connected component corresponds to the entire graph in both cases.

The number of strongly connected components in the digraphs gB and gD can vary, and it is possible to have more than one strongly connected component.

A strongly connected component is a subgraph of a directed graph where there is a directed path between every pair of vertices within that component. In other words, for any two vertices u and v in a strongly connected component, there exists a directed path from u to v and from v to u.

In the case of gB and gD, since the graphs were expanded from a single seed node and edges were created based on artists and their friends, there are groups of nodes within the graph that form strongly connected components.

Consider a scenario where there is a group of artists who are mutual friends with each other but have limited connections to artists outside of their group. In this case, the nodes corresponding to these artists and their connections will form a strongly connected component within the graph.

However, it is also possible that the entire graph forms a single strongly connected component. If there is a directed path between any pair of nodes in the graph, regardless of their initial seed node, then the graph is considered a single strongly connected component.

In our scenario we have that for graph B we have 578 connected components and for graph D we have 331 components which is the result of the scenario where there are groups of artists with mutual friends.

4. **Also justify the relationship between the previous results and the number of connected components in the undirected graphs (g ′ B and g ′ D).**

Undirected graphs	Connected Components
g ′B	2
g ′D	5

The relationship between the previous results and the number of connected components in the undirected graphs (g' B and g' D) can be explained as follows:

A connected component in an undirected graph refers to a group of nodes that are all connected to each other, but not connected to any other subgraph in the graph.

In the undirected graphs g' B and g' D, the nodes are obtained from the initial graphs (gB and gD) by considering the bidirectional edges. These bidirectional edges indicate mutual friendships between artists, meaning that they have connections in both directions.

As a result, in the undirected graphs, the artists who have mutual friends with each other are placed in the same connected component. Each connected component represents a group of individuals who have at least one mutual friend within the component.

In the case of g' B, there are 2 connected components, indicating that the graph is divided into two distinct subgraphs. This suggests the presence of two groups of artists who have mutual friends within their respective groups but are not connected to each other.

For g' D, the number of connected components is 5, implying the existence of more distinct groups within the graph. This suggests that there are multiple groups of artists collaborating with each other, forming different connected components in the graph.

In summary, the number of connected components in the undirected graphs (g' B and g' D) reflects the division of the graph into distinct groups of artists who have mutual friends within their respective components. The presence of bidirectional edges in the initial graphs allows for the identification of these connected components in the undirected graphs.

5. **Compute the size of the largest connected component from g ′ B and g ′ D. Which one is bigger? Justify the result.**

Undirected graphs	Size of the maximum connected Component	Order of the maximum connected Component
g ′B	487	187
g ′D	481	90

The connected component explored using BFS (g'B) is expected to have a larger size than the connected component explored using DFS (g'D). This is because BFS ensures that all nodes within a connected component are reached, resulting in a more comprehensive exploration of the component and a larger component size.

Also, the connected component in g' B is larger. In the case of g' D, the size of the largest connected component is 481. It is interesting to note that although the sizes of the two connected components are not significantly different, there is a distinction in their characteristics when considering the order, the density and radius of the graphs.

The density of a connected component refers to the proportion of edges present in the component compared to the total number of possible edges. In the case of the connected component in g' B, its density is 0.0280, indicating that it has a lower density, meaning it is more sparse in terms of edges. On the other hand, the connected component in g' D has a density of 0.1200, which suggests a higher density, indicating that it is more dense in terms of edges.

The radius of a connected component represents the minimum eccentricity among all vertices in the component. It is a measure of how far apart the vertices are from each other. In the case of g' B, the radius is 10, while for g' D, the radius is 5. This implies that the connected component in g' B has a larger radius, indicating that the vertices are more spread out and potentially farther apart compared to the connected component in g' D.

Taking into account the density and radius values, it can be justified that the connected component in g' B is more sparse, with a lower density and larger radius, while the connected component in g' D is more dense, with a higher density and smaller radius. These characteristics contribute to the distinction between the two connected components, despite their similar sizes.