Session 1:

1. (1.5 points) Provide the order and size of the graphs gB and gD.

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

```
Order of gB:  776
Order of gD:  529
```

The slight differentiation in the order of visited nodes between BFS and DFS is a consequence of the cyclic structures or overlapping paths within the graphs. These structural characteristics introduce variations in the exploration patterns of the two algorithms, resulting in different sequences of visited nodes.

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

In a typical related-artist network, artists have multiple related artists. If we crawl each node and add all its related artists to the graph, the number of nodes can grow quickly. However, the exact growth will depend on the search strategy used.

The BFS strategy, as mentioned earlier, explores the network "widely". It visits all immediate neighbors before moving on to their neighbors. Therefore, it's more likely to cover a larger breadth of the network quickly. This may lead to a higher graph order as more unique nodes are encountered early on, although it also increases the likelihood of encountering previously visited nodes.

On the other hand, DFS strategy goes "deep" into the network, following a path of neighbors as far as it can before backtracking. This may result in a lower graph order as DFS might delve deeper into a particular subnetwork before exploring other parts of the network, increasing the chances of revisiting nodes.

Thus, given that each crawled node is expanded with all its related artists, it's possible that the BFS strategy (gB graph) may result in a higher graph order. However, this also greatly depends on the structure of the network and the interconnectedness of artists. If the network is highly interconnected (high

clustering coefficient), the difference in order between the two strategies might be less noticeable.

(c) *Explain what size the two graphs should have.*

```
Size of gB: 4000
Size of gD: 4000
```

Given that an edge is consistently appended during each iteration irrespective of the node's pre-existing status in the graph (potentially leading to bidirectional connections), it follows that both graphs should ultimately exhibit identical sizes.

2. (1 point) Indicate the minimum, maximum, and median of the in-degree and outdegree of the two graphs (gB and gD). Justify the obtained values.

```
--------------

gB in-degree: Min: 0 Max: 34 Median: 3.0
gB out-degree: Min: 0 Max: 20 Median: 0.0

--------------

gD in-degree: Min: 0 Max: 61 Median: 4
gD out-degree: Min 0 Max: 20 Median: 0

--------------
```

- *In the context of graph gB:*

The in-degree measures vary from 0 to 34, with a median value of 3.0. This indicates that while there exist nodes that don't receive any connections, there is a node with as many as 34 incoming edges. With the median value standing at 3.0, it suggests that typically nodes in gB tend to receive connections from three other nodes.

The out-degree ranges from 0 to 20, having a median of 0.0. This implies that there are nodes from which no connections originate, but there's also a node that connects outwards to 20 other nodes. However, with the median out-degree at 0, it indicates that many nodes in gB do not initiate any connections.

- *In relation to graph gD:*

The in-degree statistics range from a minimum of 0 to a maximum of 61, with the median being 4. This suggests that while there are nodes that don't receive any connections, there exists a node with a high number of 61 incoming edges. This node could be considered as a significant hub in terms of inbound connections. The median value of 4 indicates that typically nodes in gD receive connections from four other nodes.

The out-degree measures range from 0 to 20, with the median value standing at 0. This indicates that there are nodes that don't connect to any others, but there's a node from which 20 connections originate. The median of 0 suggests that many nodes in gD do not establish any outbound connections.

In summary, while there are certain nodes in both graphs that serve as significant hubs of connections, a large portion of the nodes only have a small number of connections. These results shed light on the structures and connectivity patterns within the two graphs.

3. (0.5 points) Indicate the number of songs in the dataset D and the number of different artists and albums that appear in it.

```
The number of songs in the dataframe is 3698.
The number of unique artists in the dataframe is 384.
The number of unique albums in the dataframe is 2031.
```

(a) Explain why the number of artists is between 200 and 400, considering the input graphs.

Given the methodological approach employed in creating the dataset, a few salient assumptions can be elucidated:

We would not anticipate a substantial count of unique artists because the inception point for both graphs is identical, suggesting a degree of overlap in the artists they have traced.

However, it's reasonable to anticipate a number of unique artists at least equal to the total nodes explored across both graphs. It's improbable, though, to have more than double this quantity, even considering the distinct search algorithms employed to expand the graphs. This is due to the shared initial node, which should result in some shared artists.

Taking these assumptions into consideration, the resulting count of unique artists falling between the size of the two initial graphs seems plausible. This reflects a balance between the intersection of common artists (due to the shared origin) and the degree of distinctiveness introduced by the different search algorithms.

(b) Justify why the number of songs you obtained is correct, considering the input graphs.

Given the nature of our data retrieval process, the total quantity of songs obtained is expected and can be accounted for by the exponential characteristic of popular song acquisition from two graphs that share the same initial artist.

This process is inherently non-deterministic, rendering a direct predictive equation unfeasible. This is due to the high likelihood of song repetition from artists who are closely interconnected within the network. Consequently, the variability and unpredictability of the total song count are intrinsic attributes of the search methodology employed and the structure of the artist graphs.

(c) Justify why the number of retrieved albums is correct.

The count of distinct albums aligns with our anticipations, primarily due to the justifications previously delineated. Furthermore, it's essential to note that the number of unique albums is typically lower than that of songs. This is fundamentally due to the structure of music distribution - a single album often encompasses multiple songs, hence the larger count of unique songs in our dataset compared to the number of albums.