Marino Oliveros Blanco NIU:1668563
Pere Mayol Carbonell NIU:1669503

Lab AGX 202324 Report Part 2: data preprocessing

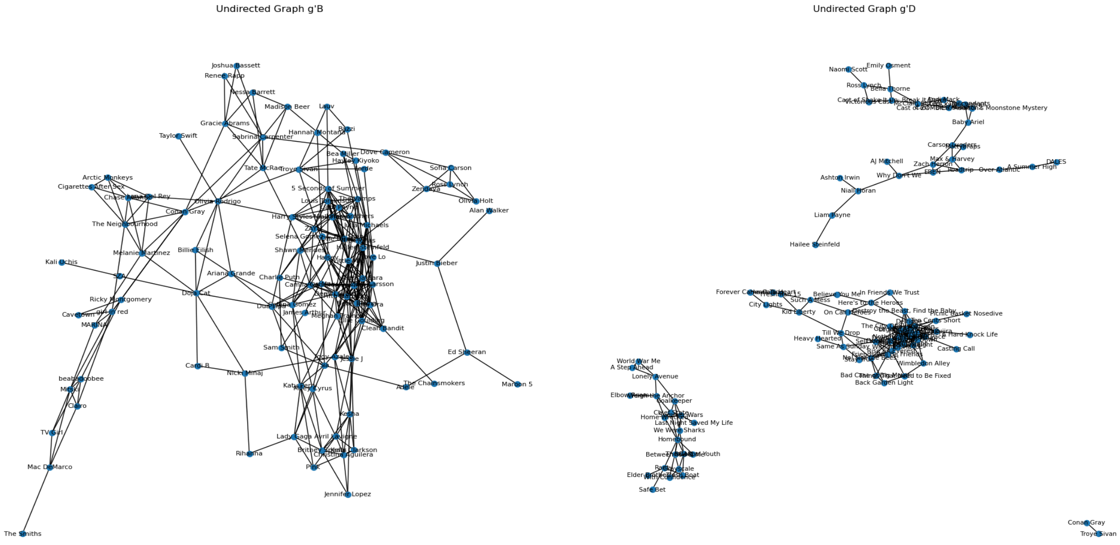
1. **(1.5 points) Justify whether the directed graphs obtained from the initial exploration of the crawler (gB and gD) can have more than one weakly connected component and strongly connected component, and explain why. Indicate the relationship with the selection of a single seed.**

When a single seed is used to start the crawling process, the resulting graph will typically consist of one strongly connected component and potentially multiple weakly connected components. This is because the crawler can only follow outgoing links from the initial seed and the nodes it discovers, but it cannot discover nodes or links that are not reachable from the seed.

The strongly connected component represents the subgraph where all nodes are reachable from each other through directed paths. However, there may exist other subgraphs that are reachable from the strongly connected component but not vice versa, forming weakly connected components.

Therefore, the selection of a single seed can result in multiple weakly connected components in the directed graphs gB and gD, as the crawler may not be able to discover all the nodes and links in the network due to the limitations of the crawling process.

As we can see by the outputs gBp has 1 connected component and gDp has 4.



2. **(0.5 points) Can the number of connected components in the undirected graphs (gB′ and gD′) 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 an undirected graph, two nodes are considered connected if there exists a path between them, regardless of the direction of the edges. On the other hand, in a directed graph, two nodes are weakly connected if there exists a path between them, ignoring the direction of the edges.

Consider the following example:

Directed graph: A -> B B -> C C -> D D -> A

In the directed graph, there is only one weakly connected component, as all nodes are reachable from each other if the direction of the edges is ignored.

Undirected graph: A - B B - C C - D

In the undirected graph, there are two connected components: {A, B} and {C, D}. This is because the edge between D and A in the directed graph has been removed in the undirected graph, breaking the connectivity between the two components.

Therefore, by converting a directed graph to an undirected graph, the number of connected components can increase if certain edges are not bidirectional in the original directed graph. This is the case in the example provided, where the undirected graph has more connected components than the weakly connected components of the corresponding directed graph.

3. **(1 point) Generate a preliminary report from the undirected graph with weights (gw).**

Due to the high number of similarities between artists we would try to implement some sort of normalization to see our results better (this is done in the notebook in our github, not on the skeleton.py file).

(a) Which are the two most (respectively, least) similar artists? What graph attribute allows you to answer this question?

The two most similar artists are found by sorting the edges of the similarity graph by weight in descending order, and taking the first edge, excluding any self-loops. This is done in the find_most_least_similar_artists function the most similar artists are Lana del Rey and Taylor Swift the same with the least similar ones which are Hannah Montana and Mitchel Musso.

The graph attribute that allows answering this is the edge weights in the similarity graph, which represent the similarity between each pair of 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?

The artist most (and least) similar to all other artists is found by calculating the average similarity of each node to its neighbors, and taking the node with the highest (lowest) average similarity. This is done in the find_most_least_similar_to_all function. The graph attribute that allows answering this is the edge weights in the similarity graph, combined with the degree of each node, which allows computing the average similarity of a node to its neighbors.

The most similar artists to all others are Hailee Steinfeld and the least Phillipa Soo.