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

Lab AGX 202324 P3 Report: Data Analysis

1. **(0.75 points) Study the number of common nodes between the obtained graphs. Use the function num common nodes.**

Using the function num_common_nodes we obtain the following.

(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?

Number of common nodes between gB and gD: 138

This indicates that the "Related Artists" algorithm on Spotify has a different focus or criteria for determining artist relationships.

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

Number of common nodes between gB and gBp: 99

The lower number of common nodes (99) between gB and gBp compared to the total number of nodes in gB suggests that the "Related Artists" relationships on Spotify are not completely reciprocal. If the relationships were fully reciprocal, we would expect gBp (the reciprocal version of gB) to have the same set of nodes as gB

The relatively low overlap between gB and gBp also indicates that Spotify's "Related Artists" algorithm may have a different focus or criteria compared to a simple reciprocation of relationships. It suggests that the algorithm likely incorporates additional factors or objectives beyond just mirroring the relationships in both directions.

2. **(0.5 points) Calculate the 25 most central nodes in the graph gB′ 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.**

Number of common nodes between degree and betweenness centrality: 9.
The overlap between the top nodes ranked by degree centrality and betweenness centrality in the graph gB′ highlights the existence of highly influential and strategically positioned nodes that

are crucial for maintaining connectivity, facilitating information diffusion, and ensuring the network's robustness and resilience.

3. **(0.5 points) Find cliques of size greater than or equal to min size clique in the graphs gB′ and gD′ . 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.**

The maximum value that generates at least 2 cliques (in total from gBp and gDp) is 7; this was obtained by trial and from 4-7 until 8 did not provide 2 cliques in total.

Size 7 generates 4 cliques in gBp with a total number of nodes in the cliques of 18 and 1 clique in gDp with a total of 7 nodes in cliques.

4. **(0.5 points) 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.**

The largest clique in gBp is composed by the following: [('64M6ah0SkkRsnPGtGiRAbb', 'Bebe Rexha'), ('1Xylc3o4UrD53lo9CvFvVg', 'Zara Larsson'), ('1zNqDE7qDGCsyzJwohVaoX', 'Anne-Marie'), ('5CCwRZC6euC8Odo6y9X8jr', 'Rita Ora'), ('2wUjUUtkb5lvLKcGKsKqsR', 'Alessia Cara'), ('5p7f24Rk5HkUZsaS3BLG5F', 'Hailee Steinfeld'), ('1l8Fu6IkuTP0U5QetQJ5Xt', 'Fifth Harmony'), ('3e7awlrlDSwF3iM0WBjGMp', 'Little Mix')]

These artists are primarily female pop and R&B which reached their peak of fame in the 2010s to a similar demographic.

5. **(0.5 points) Detects 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?**

The code uses the Louvain method to detect communities in the graph gD. The detect_communities function takes the graph and the method name as input. When the method is set to 'louvain', it converts the directed graph to an undirected graph (if necessary) and then applies the community_louvain.best_partition function from the community module to find the partition of nodes into communities.

The modularity of the obtained partitioning is 0.7218587984529132. A modularity value closer to 1 indicates better community structure in the network. A modularity value of

0.7218587984529132 is considered relatively high, suggesting that the partitioning is reasonably good, with a clear community structure in the gD graph. By our slides >0.3 is considered a sign of community structure.

6. **(1 point) 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.

The minimum cost to ensure that a user who listens to music infinitely will hear an ad at some point is determined by the number of strongly connected components (SCCs) in the graph. Each SCC needs at least one artist to have an ad placed to ensure that a user traversing that component will eventually encounter the ad.

The min_ad_cost function finds all strongly connected components using nx.strongly_connected_components and returns the number of SCCs as the minimum cost.

For graph gB, the minimum cost is 37400 euros, as there are 374 strongly connected components. For graph gD, the minimum cost is 50600 euros, as there are 506 strongly connected components.

   (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.

With a budget of 400 euros, the best strategy is to select the most central artists within each strongly connected component, as they have the highest likelihood of being reached by users traversing the component.

The select_artists_for_budget function implements this strategy:

1. It finds all strongly connected components in the graph.

2. For each SCC, it finds the node with the highest degree centrality, considering the subgraph induced by the SCC.
3. It sorts the central nodes by their centrality in descending order.
4. It selects the top central nodes, subject to the budget constraint (assuming each artist costs 100 euros).

For graph gB, the selected artists within the 400 euros budget are:
[('6oW9KRAZbC1xOlmg2RRyFL', 'McClain Sisters'), ('6G9bygHlCyPgNGxK2l3YdE', 'Vanessa Hudgens'), ('5bmqhxWk9SEFDGIzWpSjVJ', 'THE DRIVER ERA'), ('542yUd4rGzUEOLd1diV94f', 'Rocky')]

For graph gD, the selected artists within the 400 euros budget are:
[('4Uc8Dsxct0oMqx0P6i60ea', 'Conan Gray'), ('1gzqMaNtHI85YIVvZlcZHe', 'Nothing Planned'), ('3dUtEOe21FQJhD834hUkAm', 'Brigades'), ('3XHn51FdwX1GZCGMz6RMYM', 'Pentimento')]

7. **(1 point) 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.

The minimum number of hops is 3. An example hop path would be : [('06HL4z0CvFAxyc27GXpf02', 'Taylor Swift'), ('1McMsnEElThX1knmY4oliG', 'Olivia Rodrigo'), ('4VdV2qRAYBLINR6uU72V1J', 'Joshua Bassett'), ('5bmqhxWk9SEFDGIzWpSjVJ', 'THE DRIVER ERA')] making 3 hops in total.