Daniel Vidal 1634599
Neil de la Fuente 1630223
Jordi Longaron 1630483

Report on Session 3:
Data Analysis

Introduction

During this third session, we will perform some analysis of the data we have preprocessed in the previous sessions, analysis that will be useful to validate our functions from previous sessions.

Questions

1. Indicate the number of nodes shared by the graphs g_B and f_B (seeds Drake and the last crawled artist from the DFS crawl, respectively); and g_B and h_B (seeds Drake and French Montana, respectively). Use the function num common nodes. Compare the number of common nodes with the results obtained from calling the *create similarity graph* function.

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.

3. Find cliques of size greater than or equal to *min size clique* in the graphs g'_B and g'_D. 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.

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.

5. Detect communities in the graph g_D. Explain which algorithm and parameters you used, and what is the modularity of the obtained partitioning. Do you consider the partitioning to be good?

Daniel Vidal 1634599
Neil de la Fuente 1630223
Jordi Longaron 1630483

6. Suppose that Spotify recommends artists based on the graphs obtained by the crawler (g_B or g_D). 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 g_B and g_D, and justify your answer.

(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 g_B and g_D.

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 g_B graph is always used.

(a) If you start by listening to the artist Young Dro and your favorite artist is Travis Porter, 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.