Mini Project 3 – Art Analysis

Hektor Dahlberg (41685)

November 3, 2024

Contents

1 Introduction

The aim of the mini project is to create a network from the given data and do some network analysis on it.

2 Preprocessing and EDA

For preprocessing, I first checked for NaN values. Since this dataset, particularly in the relationship-related data frame, consistently contains NaN values, I didn't focus heavily on them. I then examined data distributions and created bar plots for different categories, including the top 20 movements, top 20 nationalities, and top institutions.

During this process, I also standardized some data entries. For example, I ensured that each category had consistent labels. In cases where country names were mixed with state or city names, I corrected them to show only the country. Additionally, I standardized artist nationalities, converting plurals to singular forms (e.g., "Austrians" to "Austrian" and "Italians" to "Italian").

3 Network creation and analysis

In creating the network, I defined artists, institutions, schools, and movements as nodes. These nodes form a directed graph, with different types of edges representing the various relationships between them. For undirected edges, I simply created two directed edges between the nodes, as NetworkX only supports directed edges in a directed graph.

Early in the process, I noticed that movements posed a unique challenge; there

was no direct way to link artists to movements. To address this, I added movement nodes only when establishing edges where an artist was influenced by, or had an influence on, a specific movement. This approach helped avoid isolated movement nodes in the network.

3.1 Results

For the analytics, the first three questions involved identifying which artist nodes had the highest out-degree centrality for ”influenced on” edges. I then visualized the first layer of connections from the node with the most outgoing edges.

Question 4 was a bit more challenging to represent visually, as it asked, ”Which nationalities concentrate the majority of artists?” Essentially, this question identifies the country with the highest number of artists. To visualize this, I created a network for the country with the most artists, showing edges that represent ”influenced on” and ”friends” relationships.

For question 5, I used the Louvain method for community detection to identify the largest community in the network and then visualized that community.

4 Conclusion

The main challenge with these questions was interpreting ”Which were the most influential artists?” Defining ”influence” can be complex, as it could mean various things—most productive, most influence on other artists, influence on a single prominent artist, or influence on an artist who shaped new movements. To simplify, I chose to measure influence based on the number of ”influenced on” edges each artist, movement, or institution had.

As mentioned, I initially encountered issues with having movements as separate nodes. To address this, I implemented a check during edge creation to ensure that a node is created for each movement whenever there’s an ”influence on” or ”influence by” relationship with it. This approach resolved the issue of movements potentially being isolated in the network.