Analyzing Artistic Influence Networks: A Community and Centrality-Based Exploration of WikiArt Data

Author: Nurul Amin

Introduction:

The purpose of this project is to analyze the structure and characteristics of a network derived from art-related data, with the goal of uncovering insights about influential artists, movements, institutions, and communities within the art world. By representing this data as a network, we can explore relationships and connections between artists, their influences, institutions, and art movements. Based on different network analysis techniques (centrality measures, community analysis, clustering algorithms), find out the specific goal.
The goal is to answer the following questions:

1. Which were the most influential artists?
2. Which were the most influential movements?
3. Which were the most influential institutions?
4. Which nationalities concentrate the majority of artists?
5. Which are the biggest communities in the network?

Methodology:

- **Preprocessing and EDA:**
  The dataset used in this project comprises four CSV files, each generated by scraping data from WikiArt. Each file contains specific information that will contribute to the structure of the network:
  - artists.csv: Contains detailed information about individual artists, including their WikiArt URL, unique ID, image URL, nationality, name, total number of artworks, and the years they were active.
  - relationships.csv: Contains relational information about each artist, including their network of friends, influences, artists they have influenced, the art institutions where they studied, schools they were associated with, and whether the entity is an artist or a collection.
  - institution.csv: Lists the institutions related to the artists, along with their city, country, name, and WikiArt URL.
  - schools.csv: Details on art schools, including the name and WikiArt URL of each school.

  To begin the data analysis, each dataset was loaded and examined to understand its structure, identify any missing values, and check data types.
  From the descriptive statistics for each dataset, here's a breakdown of key observations:

A. **Artists Dataset Summary**:The dataset contains information for 2,963 unique artists, each with specific details, as all values are unique per artist (count = unique = 2963 for artistUrl, id, image, and title). There are 106 unique nationalities, with the most common nationality being "American" (519 entries).The totalWorksTitle column has 294 unique values, with the most common count being "15 artworks" (137 instances). There are 2,181 unique active year intervals, with "born 1937" being the most frequent (24 entries), indicating a clustering of artists from that birth year.
B. **Relationships Dataset Summary**:This dataset is smaller, with 58 entries and 8 columns detailing relationships like friends, influenced_by, influenced_on, institution, and movements.The most frequent "friend" is "Jackson Pollock" (6 instances), and among influencers, "Jan van Eyck" and "Stefan Lochner" are the top influences, each appearing once.There are 39 unique movements, with "Expressionism" as the most common (6 entries). All entries are labeled as "Artists," as indicated by a single unique type with 58 occurrences.
C. **Institutions Dataset Summary**:This dataset has 73 entries across 4 columns, representing art institutions, including the cities and countries they are located in. 52 unique cities and 35 unique countries, with "Paris" and "UK" as the most common city and country, respectively. All entries have unique titles, with "Guild of Saint Luke" being the most frequent title, appearing only once.
D. **Schools Dataset Summary**:The dataset contains 220 entries, each representing a unique art school.The "Flemish School" appears once, highlighting unique school titles, with each school URL being distinct as well.

- **Network analysis and analysis:** Firstly I created a network G within a few nodes and edges, for showing the fresh network in my little colab interface. That's why I choose to create the artists and their friends network. However, for the ease of use I found out the friends name based on primary key "artistUrl" from both artists and relationships datasets. There are 80 nodes and 124 edges in this G network. The network visualization of G is Fig 1. Then I targetted to create the exact network G1 for extracting the desire goal set by the professor Bogdan Iancu. I set the node of G1 by selecting the artist title, institution url, school url, artist's nation, artist's movement. So, the total nodes after addition was 3401. Also, I added the edges node by making tuple with artist's title and other desired elements. After adding the edges there were total 3973 edges in the network. The average degree of the network was 2.08 and density was 0.00054. I defined the isolated node and remove them from the network. The visualization of the network G1 showing in fig 2.

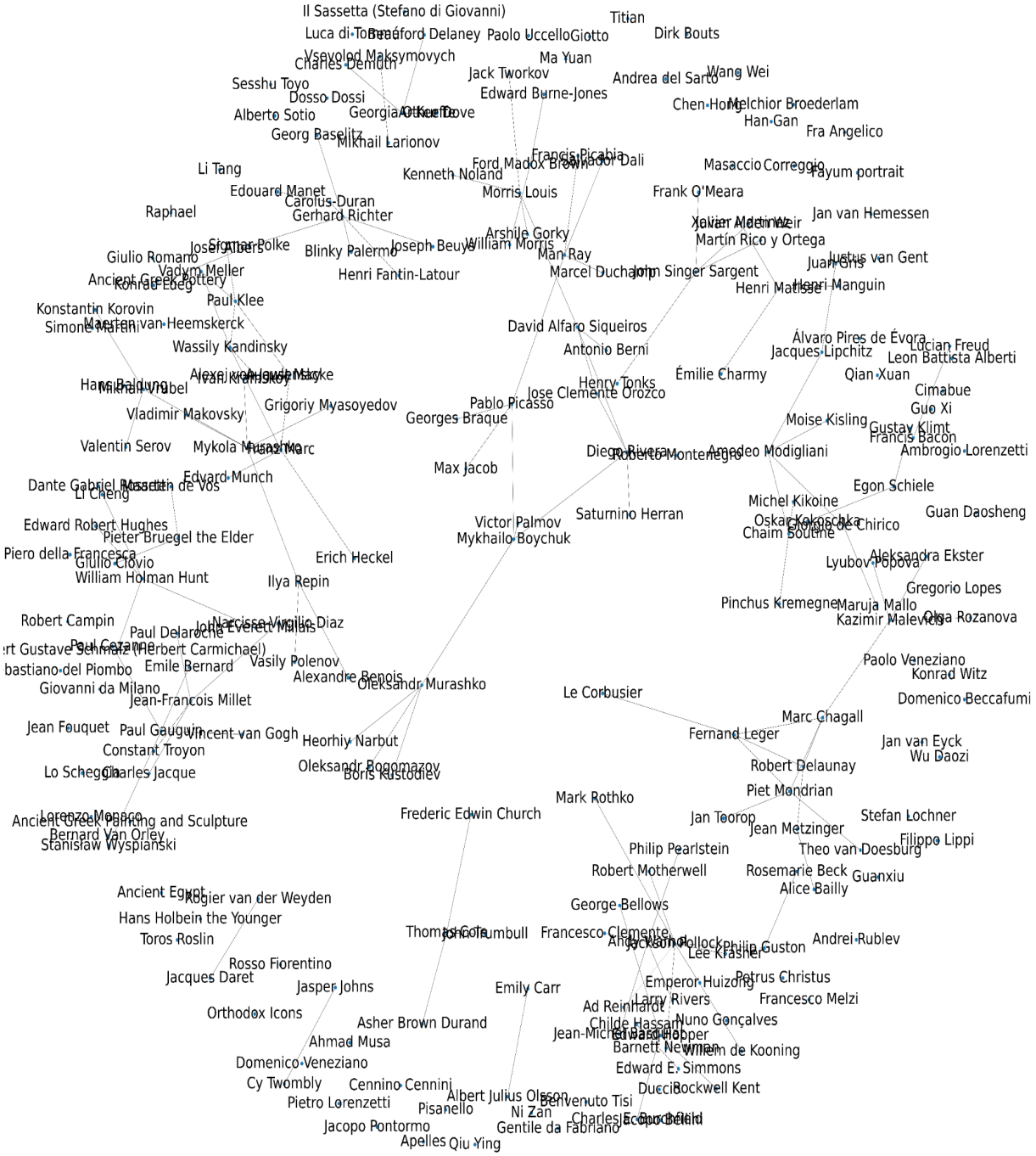

Fig 1: Artists and their Friends Network

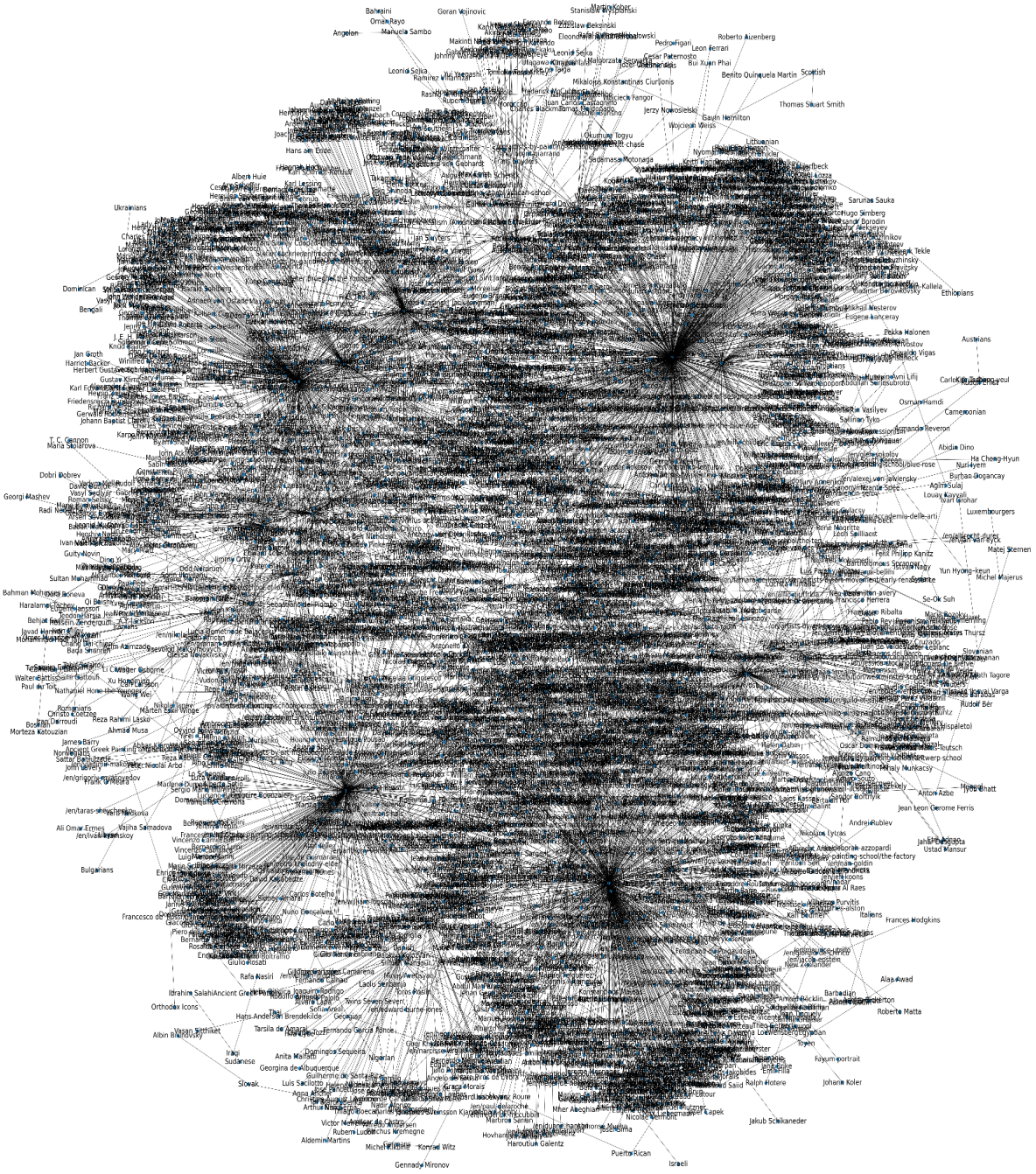

Fig 2: G1 Network

Results:

I have found out the exact results of my goal using three centrality measurement which were Degree Centrality, Betweenness Centrality and Closeness Centrality. Degree centrality measures the number of direct connections each artist (node) has in the network. Artists with higher degree centrality are directly connected to more other artists, indicating popularity or influence. Betweenness centrality measures the frequency at which an artist lies on the shortest path between other pairs of artists. Higher values indicate that the artist plays a critical role in connecting different parts of the network. Closeness centrality reflects how close an artist is to all others in the network, based on the shortest paths. A higher score means the artist can be reached more easily by others, indicating influence and accessibility. However, Communities (or clusters) in the network represent groups of artists who are more closely connected to each other than to those outside the group. This can indicate artistic schools, geographic proximity, or stylistic similarities.These centrality measures and community detection results help us identify key players in the art world network, understanding both individual influence and group connections. The G1 network have 107 community and the Top Community ID is 29 within the 501 Members. There are top findings in table 1.

Title	Degree Centrality	Closeness Centrality	Betweenness Centrality
Top Artist	Pablo Picasso, 0.0154	Paul Gauguin, 0.2052	Amedeo Modigliani, 0.0797
Top Movement	Abstract Expressionism, 0.0031	Realism, 0.1871	Expressionism, 0.0422
Top Institution	Art-students-league, 0.0025	Acad-mie-julia, 0.1807	Akademie-der, 0.0078
Top Nation	American, 0.1453	American, 0.1886	American, 0.1903

Table 1: G1 network's top artist, top movement, top institution, top nation

Conclusion: The initial stages of data loading and preprocessing were straightforward. Loading the datasets, handling basic data cleaning, and exploring the dataset with preliminary statistics provided insights into the structure of the data. Basic exploratory data analysis (EDA), such as checking for missing values and analyzing distributions, was also manageable and revealed key patterns, such as the distribution of artist nationalities and active years. The community detection and network analysis aspects were challenging. Different methods for detecting communities yielded varying results, reflecting the complexity of defining communities in real-world networks. Additionally, interpreting the centrality measures and understanding the implications of various community structures required careful analysis, as different approaches could lead to different insights on the network's influential nodes. In future work, a deeper exploration of community detection methods would be beneficial, potentially involving multiple algorithms for cross-validation to ensure the robustness of the community structure findings.