Mini Project- 3

Abhinandan Roul

Introduction

The objective of this project is to construct a network analysis graph and generate insightful visualizations based on data sourced from WikiArt, an extensive online repository dedicated to art exploration and analysis. Using this dataset, the project aims to analyze relationships and patterns through network analysis. The goal is to find the most influential artists, movements, institutions, nations having most artists, and largest communities. The approach is using the graph to facilitate deeper understanding of various interconnected elements within the dataset, and ultimately leading towards the goal.

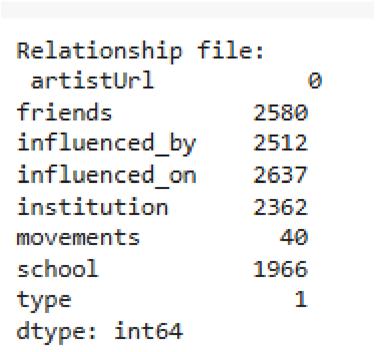
Dataset

The dataset used is sourced from WikiArt with four csv files namely –

1. **artists.csv** – Url of artist, id, image url, nation, name, total artwoks, interval of active years
2. **relationships.csv** -URL of artist, list of friends, list of artist that they influenced by, list of artist that they influenced, list of art institutions that the artist studied, list of schools that was part of, type.
3. **institution.csv** -city, country, name, url of institution
4. **schools.csv**- name, school url, wikiart

Preprocessing and EDA –

In data preprocessing, it is found that for artist file, the artist id and image url is not a part of our use case, so it was dropped. Secondly, many columns in the file relationships had multiple entries in comma separated values. So, a new row/entry was created for each of the values in the column. The columns having multiple values in single entry were, “influenced_on”, “influenced_by”, movements and institutions. Some missing values were found, but were not dropped at this stage.



```
Relationship file:
 artistUrl          0
friends          2580
influenced_by    2512
influenced_on    2637
institution      2362
movements          40
school           1966
type                1
dtype: int64
```

Fig.1. Missing values in relationship file

Exploratory Data Analysis

Some interesting visuals that were plotted are distribution of artist’s nationalities, artists and their total artworks, and frequency distribution of most popular movements.

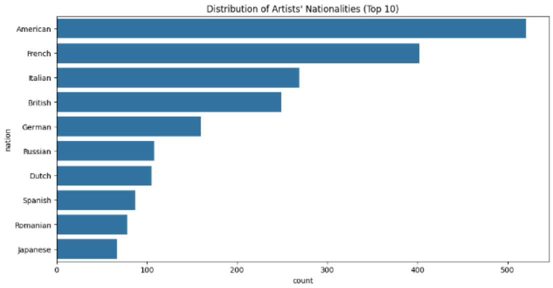


Fig. 2. Distribution of artists’ nationalities

In fig.2. it can be inferred that American, French, Italian, British and German artists are the most frequently occurring in the dataset. That means more artists belong to these countries.

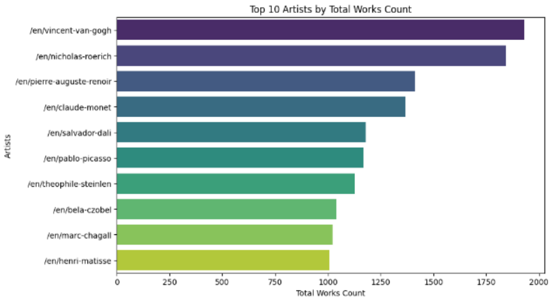


Fig.3. Distribution of artists by total artworks

In fig 3. It is clear that Vincent Van Gogh has the most artworks, followed by Nicholas Roerich. They both have more than 1750 artworks each. Pierre Auguste Renoir is at the third place.

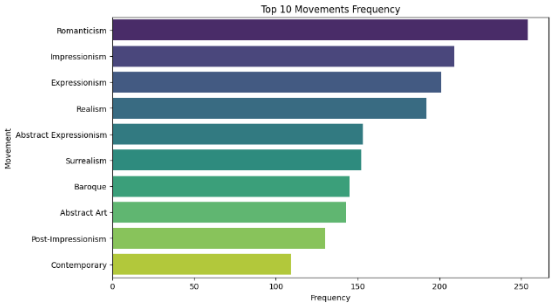


Fig.4. Distribution of top 10 movements frequency

From the figure Romanticism, Impressionism and Expressionism are the most popular movements as per the data.

Network Creation and Analysis

The network graph was made using the NetworkX library in Python, with nodes and edges representing the primary components of our network. Each node represents an entity in the network (such as an artist or institution), and each edge represents a type of relationship between two entities

The goal of finding the most influential artists, movements, institutions, artists' nationalities was kept in mind while creating the structure of the graph. These were the following nodes created, based on requirements.

Artist Nodes: These represent individual artists, with edges connecting them to movements, institutions, or other artists (in cases of "influenced by" and "influenced on" relationships). This captures influence networks and connections to schools of thought.
These nodes were given the attributes of nation, title, and total works. These properties are important for identifying artists name, volume of work and the country they belong to.

Movement Nodes: These represent movements (e.g., Renaissance, Baroque) with directed edges from artists to movements, showing artists' affiliations with or contributions to movements.

Institution Nodes: Represent educational or artistic institutions (e.g., Royal Academy of Fine Arts), with edges connecting them to the artists they have trained or influenced. Attributed created were title, city, country, url.

Nationality Nodes: Indicate the country of artists, with edges connecting each artist to their nationality, allowing for analysis of the geographic distribution of artists.

Table 1. Shows the summary of relationship definition.

Node1	Relationship	Node2
ArtistUrl	Influenced by	ArtistUrl
ArtistUrl	Influenced on	ArtistUrl
Movement	Influenced By	ArtistUrl
Movement	Influenced on	ArtistUrl
Institution	Influenced by	ArtistUrl
Institution	Influenced on	ArtistUrl
ArtistUrl	(related)	Nation

Artist-to-Artist Edge: Edges connecting artists who influence or are influenced by each other represent shared ideas. These relationships are crucial in identifying clusters of artists with a with similar artistic styles.

Movement to Artist Edge: Edges between artists and movements reveal artists' ties with specific institutions. These edges help to construct clusters around movements, demonstrating which movements were popular.

Institution to Artist Edge: Connections between artists and institutions (or schools) indicate where artists studied or associated. This type of relationship illustrates institutional influence, showing where artists may have been formally educated or collaborated.

Artist to Nation Edge: These edges associate artists with their respective countries, helping to visualize geographic clustering and the spread of artists by country.

The relationship was defined as per the table mentioned above. An undirected graph was created, and relationship is established. For each of the node a type is attributed while adding nodes to the graph. For example while adding movement as nodes, the type is declared as movement. This is done for all the declared nodes.

A graph was constructed with 4002 nodes, and 7958 edges.

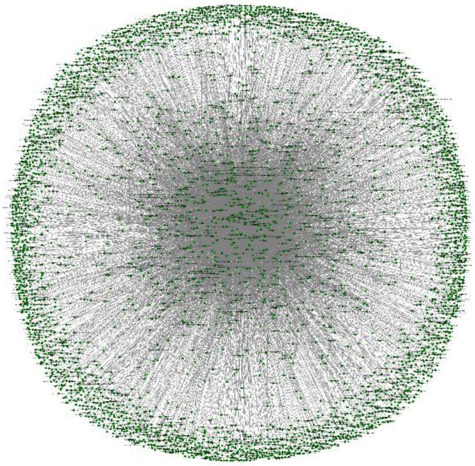


Fig. 5. Visualization of the graph network

From network analysis it is found the average degree to be 3.97. This suggests that on average each node is connected to around 4 other nodes.

The network density is 0.0009. Such a low density suggests that the network is quite sparse, with only a few of the possible relationships between entities. This sparsity could imply that the art world represented here is segmented, with certain clusters or communities forming around specific relationships rather than a highly interconnected network.

Visualization of Graphs and certain subgraphs

In this section, the various nodes are plotted to indicate their relationship. Multiple subgraphs concerning specific relationships are visualized as well.

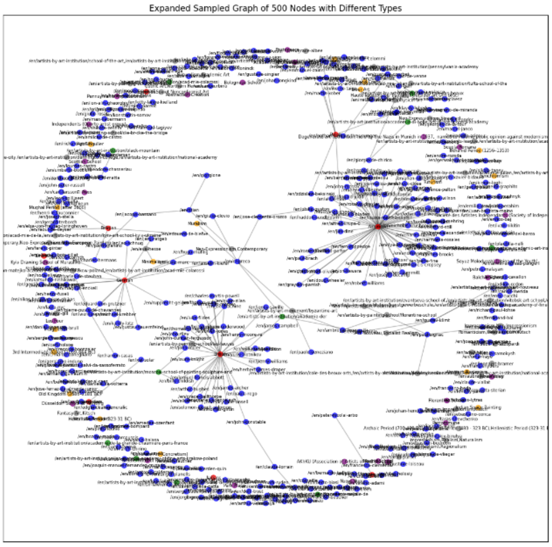


Fig.6. Plot of the graph network considering 500 nodes.

This graph plots the various nodes in the network and shows the relationship between them. **Artist nodes** are depicted as blue, **institution** as green, **movement** as orange, and **nation** as red. A random sample of 500 nodes are taken to plot this visualization as plotting all the nodes take a long time, and might cause a system crash due to insufficient memory. From this graph, we infer that a lot of artists are either French or American. Other nodes which are not connected indicate there are isolated nodes in the graph as well. For example, there might be an institution which isn't connected to any artist in the dataset, or their connections might not be present in the random sample of 500 nodes.

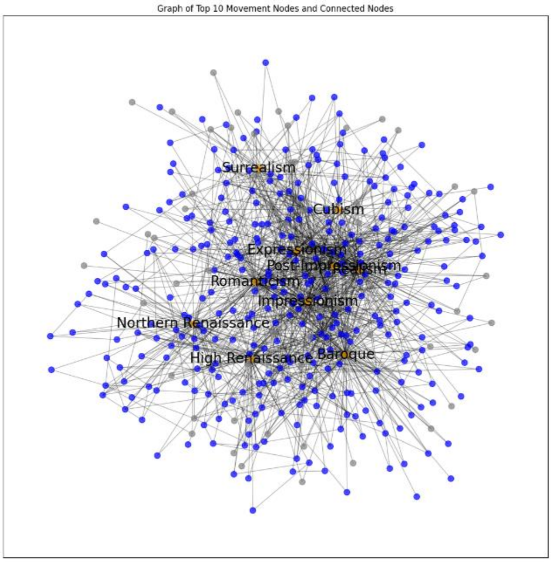


Fig.7. Movement and influenced artists

These indicate the top 10 movements, and the edges shows the artists that influenced them.

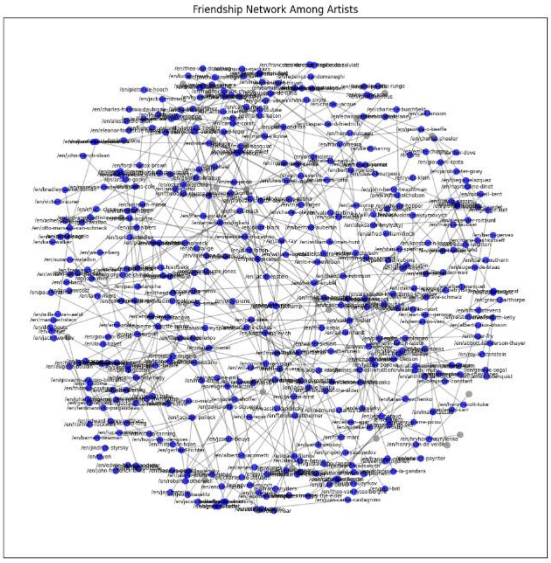


Fig.8. Friendship network

The above figure plots the nodes to indicate the friendship between the artists.

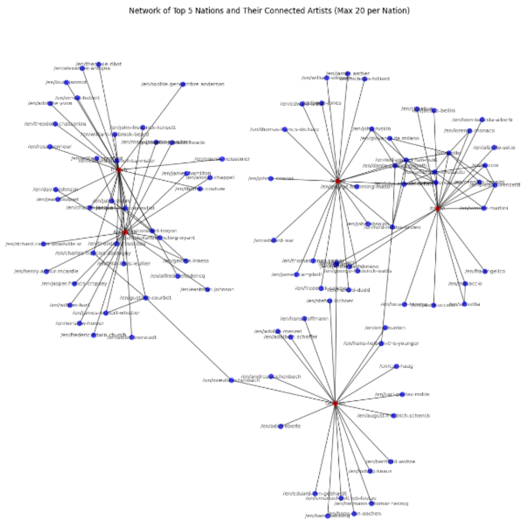

Network of Top 5 Nations and Their Connected Artists (Max 20 per Nation)

Fig.9. Top 5 nations and their influential artists

The above figure indicates the top 5 countries, and only shows 20 artists that is linked to them. This is done to make the graph visually clutter free and interpretable.

Interactive Visualization

Fig. 10.Interactive visualization

A dash app is used to create a interactive graph, where nodes are highlighted as per the selection. The user can choose the type of node to see it highlighted on the plot.

The approach towards finding the goals is by finding the degree centrality of each node for specific type. It is a floating-point number between 0 and 1. The higher the number more is the influence of the node. Also, eigenvector centrality is also implemented to compare the different output if any.

Degree Centrality: This metric simply counts the number of direct connections (edges) that an artist (node) has. An artist with many connections to other artists, movements, or institutions will have a high degree centrality, indicating major **direct influence.**

Eigenvector Centrality: This metric considers both the quantity and the quality of connections. An artist with ties to other influential or well-connected nodes will have a higher eigenvector score, indicating deeper **indirect influence.**

In brief,

Degree Centrality answers, "How many connections?" whereas,

Eigenvector Centrality answers, "How important are those connections?" Therefore, in my results, the output varies for degree and eigenvector centrality.

Results

Goal-1: To find most influential artists

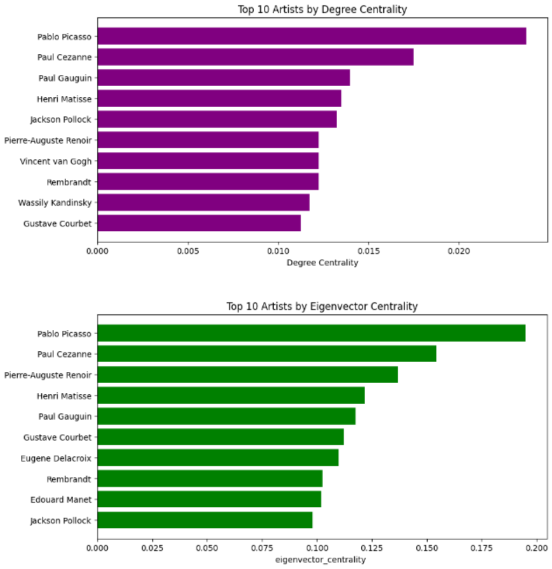


Fig. 11. Most influential artists by degree and eigenvector centrality

Pablo Picasso and Paul Cezanne were found to be the most influential artists, as per my analysis. There might be some discrepancies with the degree centrality as the preprocessing included some extra entries due to the splitting comma separated values the influenced_on and influenced_by columns.

Goal 2: Most influential movements

This was calculated on the basis on influenced by and influenced on columns. The nodes having the topmost degree centrality were chosen.

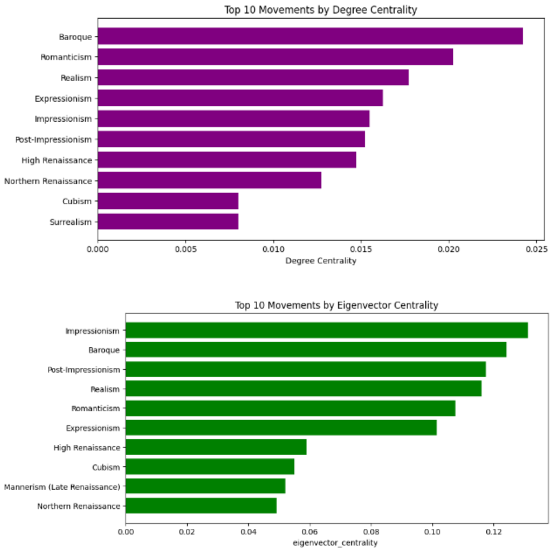


Fig. 12. Most movements artists by degree and eigenvector centrality

As per degree centrality, Baroque, Romanticism, Realism, Expressionism Impressionism were most influential movements. Whereas as per Eigenvector centrality Impressionism, Baroque, Post-Impressionism Realism, Romanticism are the most influential movements.

From this, it can be inferred that degree centrality is more accurate as per the preliminary analysis in the EDA. Also, with high eigenvector centrality Impressionism can be considered to have more deeper connections to other influential artists or institutions as well.

Goal 3: Most influential institutions

This was calculated on the basis on influenced by and influenced on columns. The nodes having the topmost degree centrality were chosen.

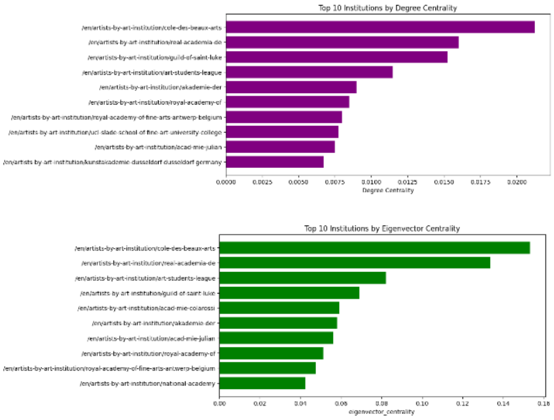


Fig. 13. Most influential institutions by degree and eigenvector centrality

Cole des Beaux Arts was the most influential institution, followed by Real Academia de Bellas Artes de San Fernando. The 3rd place as determined by degree centrality differs from eigenvector centrality.

For eigenvector centrality analysis it's almost the same, except for few entries.

Goal 4: Nationality having majority of artists

This was found out having artists and nations and finding out the degree centrality of each nation node. The nodes having most degree would be the one associated with maximum number of artists.

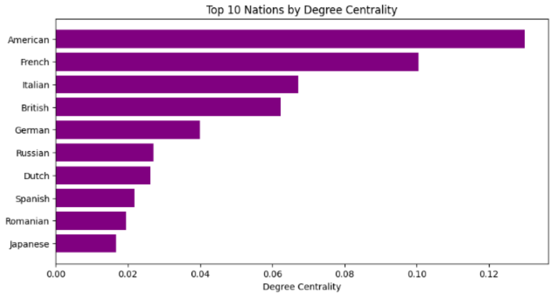


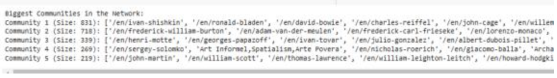
Fig 14. Nationalities concentrating majority of artists

It is found out that America, France, Italy, Britain has the greatest number of artists associated with them.

Goal 5: Community Detection

A community is a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network.

The **greedy modularity community** method is utilized to find the communities. The below visual is based on artist nodes.



```
Biggest Communities in the Network:
Community 1 (Size: 831): ['/en/ivan-shishkin', '/en/ronald-bladen', '/en/david-bowie', '/en/charles-reiffel', '/en/john-cage', '/en/willem
Community 2 (Size: 718): ['/en/frederick-william-burton', '/en/adam-van-der-meulen', '/en/frederick-carl-frieseke', '/en/lorenzo-monaco',
Community 3 (Size: 339): ['/en/henri-motte', '/en/georges-papazoff', '/en/ivan-tovar', '/en/julio-gonzalez', '/en/albert-dubois-pillet', '
Community 4 (Size: 269): ['/en/sergey-solomko', 'Art Informel,Spatialism,Arte Povera', '/en/nicholas-roerich', '/en/giacomo-balla', 'Archa
Community 5 (Size: 219): ['/en/john-martin', '/en/william-scott', '/en/thomas-lawrence', '/en/william-leighton-leitch', '/en/howard-hodgki
```

Fig. 15. Output of community detection

The above picture indicates the biggest communities in the network. Top five communities had a size of 831, 718, 339, 269 and 219 respectively.

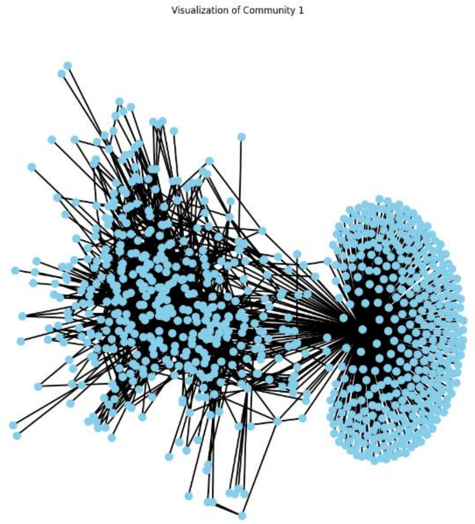
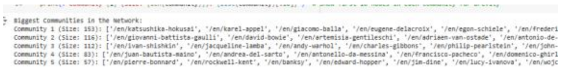


Fig. 16. Largest community

The above figure 16., depicts the largest community which has 831 nodes, connected to each other. This contains all the nodes in the graph, irrespective of type.

Then, a subgraph is created with only the artist nodes, and then, a community is created based on greedy modularity approach. The size of top 5 communities are 153, 116, 112, 83 and 57.



```
Biggest Communities in the Network:
Community 1 (Size: 153): ['/en/katsushika-hokusai', '/en/karel-appel', '/en/giacomo-balla', '/en/eugene-delacroix', '/en/egon-schiele', '/en/frederi
Community 2 (Size: 116): ['/en/giovanni-battista-gaulli', '/en/david-bowie', '/en/artemisia-gentileschi', '/en/adriaen-van-ostade', '/en/antonio-de-
Community 3 (Size: 112): ['/en/ivan-shishkin', '/en/jacqueline-lamba', '/en/andy-warhol', '/en/charles-gibbons', '/en/philip-pearlstein', '/en/john-
Community 4 (Size: 83): ['/en/juan-bautista-maino', '/en/andrea-del-sarto', '/en/antonello-da-messina', '/en/francisco-pacheco', '/en/domenico-ghirl
Community 5 (Size: 57): ['/en/pierre-bonnard', '/en/rockwell-kent', '/en/banksy', '/en/edward-hopper', '/en/jim-dine', '/en/lucy-ivanova', '/en/wojc
```

Fig. 17. Output of Community in Subgraph containing Artist Nodes only

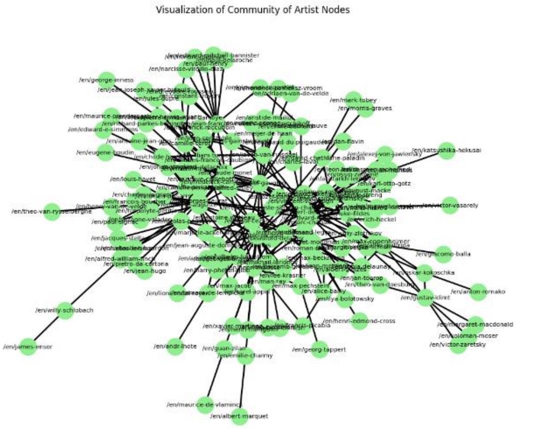


Fig. 18. Visualization of community of size consisting of artist nodes.

Conclusion

The main goal of this project was the successful construction of a graph using NetworkX, which effectively represented the relationships between artists, movements, institutions, and nations. By analysing centrality measures such as degree centrality and eigenvector centrality, the most influential nodes within the network were identified. The primary goals of finding influential artists, movements, institutions were fulfilled.

The results are almost like expected outcomes, which was noted from the preliminary EDA analysis. But EDA is not a definitive measure to find the influence as it only takes the counts of artists, not the actual connectedness to different nodes. Also, plotting of large graphs was a challenge due to memory errors, and session timeout. Therefore, the visualization is only done for random section of nodes. There were a lot of relationships between nodes, and establishing and showing it accurately on a graph was very difficult.

The future work involves using other relationships in the graph to generate more insightful analysis.