Artist Relationships in the WikiArt Collection

As museums and other archives are digitalizing their collections, data analysis can be used to retrieve insights from the collection. WikiArt is an extensive online encyclopedia for art, updated by anyone interested in the subject. This paper will examine its collection using four CSV files representing artists and collections, their relationships, institutions, and schools. These files were generated by scraping the WikiArt web page.

This data is used to derive connections between artists, institutions, movements, and nations.

1 Data Collecting

The data consists of four CSV files:

artist.csv: url of artist, id, imageURL, nation, name, total of art work, interval of active years

relationships.csv: url of artist, friend list, artists that have influenced them, artists they have influenced, institutions studied, schools studied at, type (artist/collection)

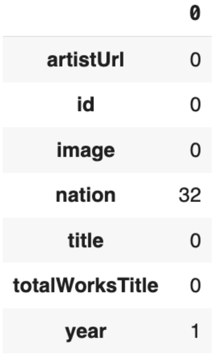
institution.csv: city, country, name, url of institution

schools.csv: name, school url

For this analysis, the data in the CSV files is read into four corresponding data frames: artists_df, relationships_df, institutions_df, schools_df.

1.1 Checking NaN-values

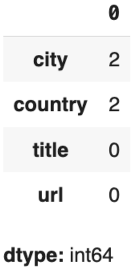
The length of artists_df is 2996, and some NaN values were observed: 32 missing nations and 1 year.



	0
artistUrl	0
id	0
image	0
nation	32
title	0
totalWorksTitle	0
year	1

Figure 1: NaN values in artists_df

Institutions_df has some missing values.

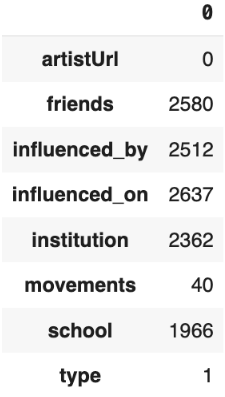


	0
city	2
country	2
title	0
url	0

dtype: int64

Figure 2: NaN values in institutions_df

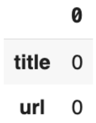
The most significant amount of NaN values is observed in the relationship_df data frame.



	0
artistUrl	0
friends	2580
influenced_by	2512
influenced_on	2637
institution	2362
movements	40
school	1966
type	1

Figure 3: NaN values in relationships

The schools_df data frame has no missing values.



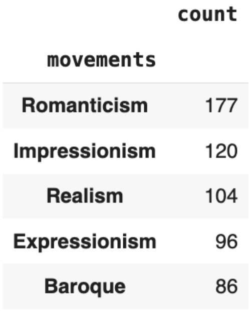
	0
title	0
url	0

Figure 4: NaN values in schools_df.

NaN values are handled by dropping them when creating nodes and edges, and adding “Unknown” when creating dictionaries for nodes.

1.2 EDA visual data analysis

The total amount of movements extracted from the relationships_df data frame is 717. The five most commonly noted movements are Romanticism, Impressionism, Realism, Expressionism, and Baroque.



	count
movements	
Romanticism	177
Impressionism	120
Realism	104
Expressionism	96
Baroque	86

Figure 5: NaN values in movements_df

The distribution of countries is visualized, with American, French, Italian, British, and German artists at the top. America seems to refer to the United States of America.

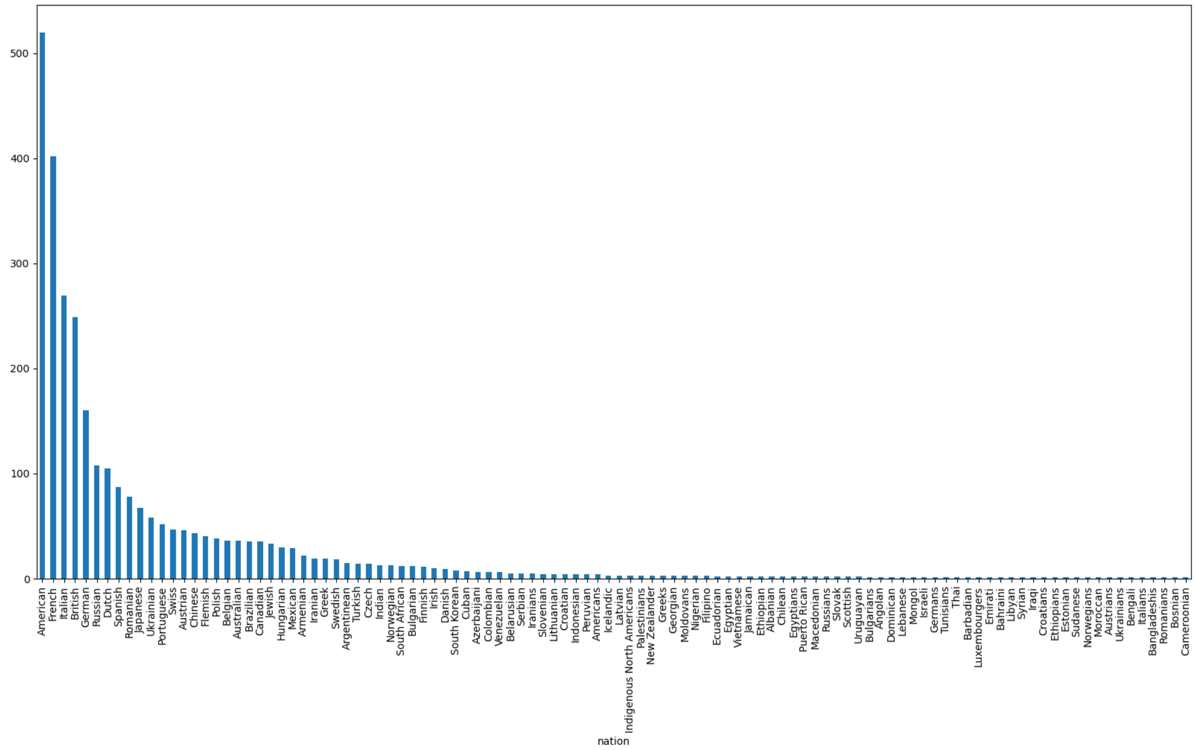


Figure 6: Distribution of artists per country

2 NetworkX Analysis

In this part of the analysis nodes and edges are derived from the data frames created in the data analysis part of this paper.

The influence will be measured by using the influence columns for each artist. Institution and movement influence is measured indirectly based on the artists they are associated with.

The influence will be measured in different degrees such as degree and out-degree, based on the graph type.

2.1 Influential artists

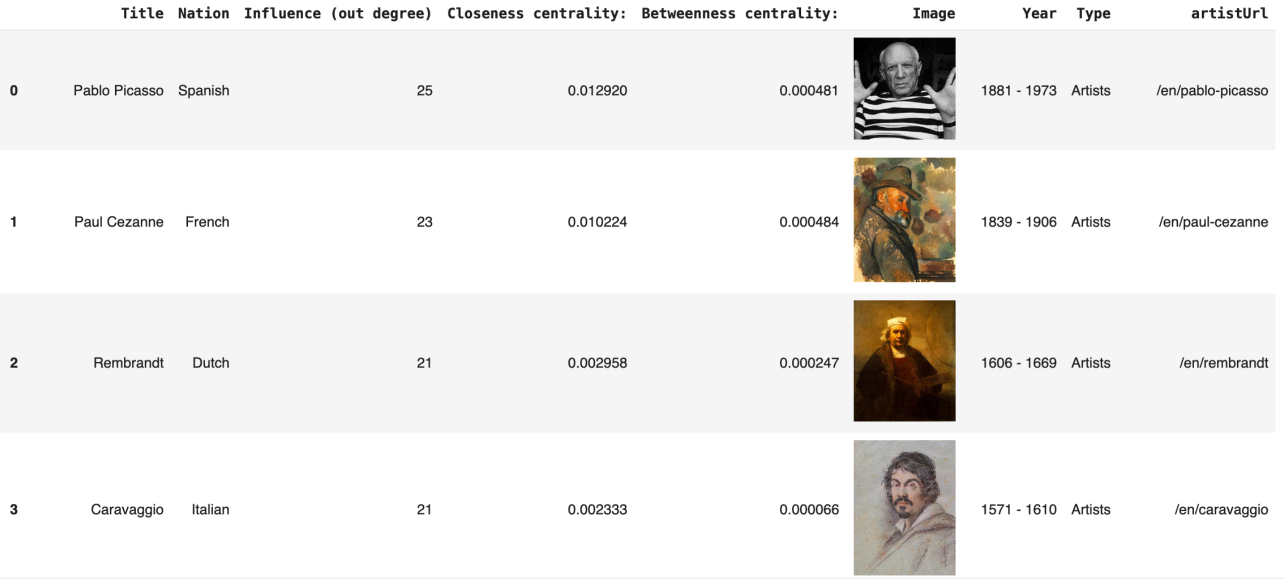
To analyze the most influential artists, nodes are created from the data frame artists_df, and the Type column is added to each artist from relationships_df, as artists_df contains both artists and collections.

Artists are distinguished by the artist URL attribute used in artists_df and relationships_df. Dictionaries are created to keep track of other artist attributes for the nodes, such as nationality and images.

Some cataloging errors were noted at this stage, as the total number of nodes before graph creation was 2996 nodes, increasing to 3138 after the graph was created. The difference is rooted in discrepancies between the artist_df and relationship_df, which both have a slightly different set of artists represented by artist URLs. Edges are created from the relationships_df. To include as many edges as possible, both influenced_on and influenced_by columns are used. This reduces the risk of losing edges caused by human cataloging mistakes. The relationship direction is reversed for influenced_by, to map the right direction of influence. Duplicates are dropped.

In this analysis, collections are not separated at this stage. Thus, the analysis and graph representation include relationships between artists and collections, as well as other artists.

As the direction of the relationship is essential, the most influential artists are determined by their outdegree. The relationships were examined with a directional graph.



	Title	Nation	Influence (out degree)	Closeness centrality:	Betweenness centrality:	Image	Year	Type	artistUrl
0	Pablo Picasso	Spanish	25	0.012920	0.000481		1881 - 1973	Artists	/en/pablo-picasso
1	Paul Cezanne	French	23	0.010224	0.000484		1839 - 1906	Artists	/en/paul-cezanne
2	Rembrandt	Dutch	21	0.002958	0.000247		1606 - 1669	Artists	/en/rembrandt
3	Caravaggio	Italian	21	0.002333	0.000066		1571 - 1610	Artists	/en/caravaggio

Figure 7: Most influential artists based on their out degree

The distinction between artist and collection was made when creating the data frame with the final results of the analysis. Some of the most influential artists were Pablo Picasso, with an outdegree of 25, Paul Cezanne with an outdegree of 23, and Rembrandt and Caravaggio with an outdegree of 21 each.

The highest closeness centrality was observed for Damien Hirst, 0.018660.

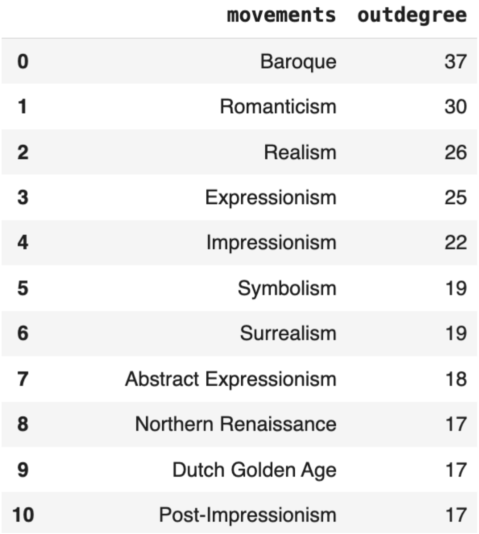
Paul Cezanne has the highest betweenness centrality with 0.000484.

The low centrality values imply that the network has a low density.

2.2 Influential movements

The most influential movements are examined by creating a new movements_df comprising the movement and influence columns from relationships_df. Unique movements are extracted as nodes.

The relationships are represented with a directed graph. The graph also features artists and collections that the movements have influenced.



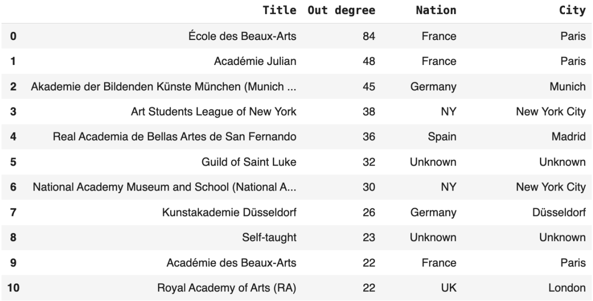
	movements	outdegree
0	Baroque	37
1	Romanticism	30
2	Realism	26
3	Expressionism	25
4	Impressionism	22
5	Symbolism	19
6	Surrealism	19
7	Abstract Expressionism	18
8	Northern Renaissance	17
9	Dutch Golden Age	17
10	Post-Impressionism	17

Figure 8: The most influential movements based on their outdegree

The movements and their outdegrees are represented in Figure 8 above. Baroque, romanticism, realism, and expressionism have the highest outdegrees.

2.3 Influential institutions

The influential institutions are measured like the movements, based on their indirect influence via the artist they are associated with.



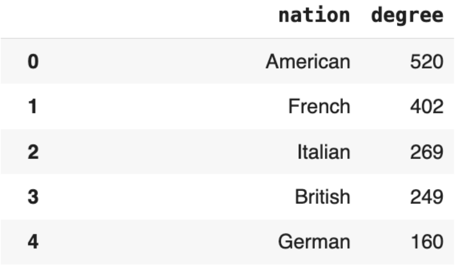
	Title	Out degree	Nation	City
0	École des Beaux-Arts	84	France	Paris
1	Académie Julian	48	France	Paris
2	Akademie der Bildenden Künste München (Munich ...	45	Germany	Munich
3	Art Students League of New York	38	NY	New York City
4	Real Academia de Bellas Artes de San Fernando	36	Spain	Madrid
5	Guild of Saint Luke	32	Unknown	Unknown
6	National Academy Museum and School (National A...	30	NY	New York City
7	Kunstakademie Düsseldorf	26	Germany	Düsseldorf
8	Self-taught	23	Unknown	Unknown
9	Académie des Beaux-Arts	22	France	Paris
10	Royal Academy of Arts (RA)	22	UK	London

Figure

École des Beaux-Arts has the highest outdegree of 84.

2.4 Nations

For the analysis of the nations with the most artists, nations are used as nodes, nations and artists as edges. The nations are sorted by their degree in Figure 9, which displays the top five nations by amount of artists.



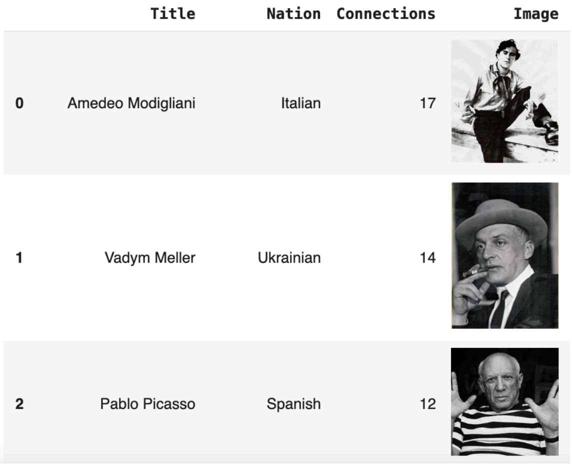
	nation	degree
0	American	520
1	French	402
2	Italian	269
3	British	249
4	German	160

Figure 9: Degree for the top five nations

Figure 9 displays the same distribution as Figure 6 in the EDA portion of this paper.

2.5 Largest communities

The friendship column is examined to find the largest communities. Artists are used as nodes, and the friendships are derived from the relationships_df.

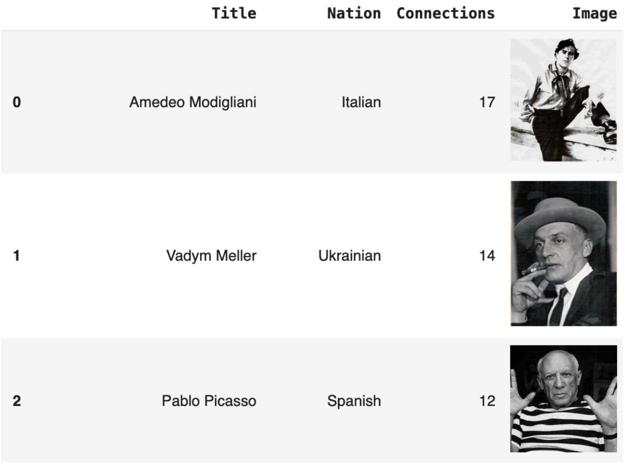


	Title	Nation	Connections	Image
0	Amedeo Modigliani	Italian	17	
1	Vadym Meller	Ukrainian	14	
2	Pablo Picasso	Spanish	12	

Figure 10: Artists with most friends in the entire network

The graph has 3025 nodes and 482 edges.

The largest component of the network is selected to examine the largest cluster.

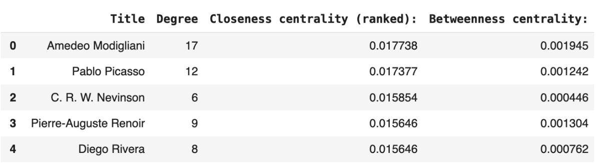


	Title	Nation	Connections	Image
0	Amedeo Modigliani	Italian	17	
1	Vadym Meller	Ukrainian	14	
2	Pablo Picasso	Spanish	12	

Figure 11: Most friends in the subgraph

The artists with the most friends in the complete network also have the most friends in the largest subgraph.

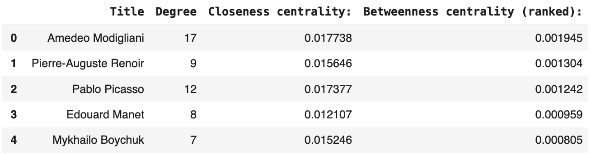
The subgraph has 191 nodes and 282 edges, making it easier to visualize. The density of the entire graph is 0.00010538, while the largest subgraph's density is 0.0155414, making it significantly denser. However, both graphs are sparse.



	Title	Degree	Closeness centrality (ranked):	Betweenness centrality:
0	Amedeo Modigliani	17	0.017738	0.001945
1	Pablo Picasso	12	0.017377	0.001242
2	C. R. W. Nevinson	6	0.015854	0.000446
3	Pierre-Auguste Renoir	9	0.015646	0.001304
4	Diego Rivera	8	0.015646	0.000762

Figure 12: Artists with the highest closeness centralities in the entire network

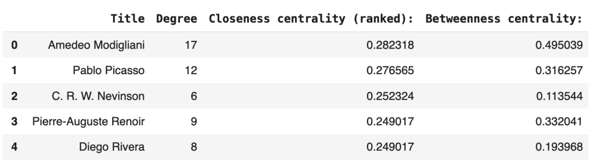
Amadeo Modigliani has the highest closeness centrality in the entire network: 0.017738.



	Title	Degree	Closeness centrality:	Betweenness centrality (ranked):
0	Amedeo Modigliani	17	0.017738	0.001945
1	Pierre-Auguste Renoir	9	0.015646	0.001304
2	Pablo Picasso	12	0.017377	0.001242
3	Edouard Manet	8	0.012107	0.000959
4	Mykhailo Boychuk	7	0.015246	0.000805

Figure 13: Artists with the highest betweenness centralities in the entire network

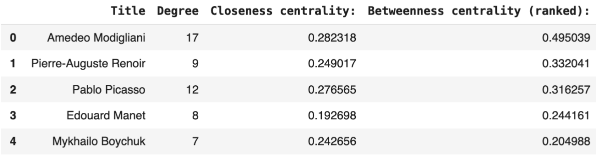
Modigliani also has the highest betweenness centrality in the entire network, 0.001945. Both closeness and betweenness centralities are low, implying that the entire graph is loosely connected.



	Title	Degree	Closeness centrality (ranked):	Betweenness centrality:
0	Amedeo Modigliani	17	0.282318	0.495039
1	Pablo Picasso	12	0.276565	0.316257
2	C. R. W. Nevinson	6	0.252324	0.113544
3	Pierre-Auguste Renoir	9	0.249017	0.332041
4	Diego Rivera	8	0.249017	0.193968

Figure 14: Artists with the highest closeness centralities in the subgraph

The same analysis is conducted on the subgraph. Amadeo Modigliani's closeness centrality is 0.282318 in the subgraph, which is significantly higher than in the entire network.



	Title	Degree	Closeness centrality:	Betweenness centrality (ranked):
0	Amedeo Modigliani	17	0.282318	0.495039
1	Pierre-Auguste Renoir	9	0.249017	0.332041
2	Pablo Picasso	12	0.276565	0.316257
3	Edouard Manet	8	0.192698	0.244161
4	Mykhailo Boychuk	7	0.242656	0.204988

Figure 15: Artists with the highest betweenness centralities in the subgraph

Such is the case for Amadeo Modigliani's betweenness centrality as well, which is 0.495039 in the subgraph. Modigliani, along with Pierre-Auguste Renoir and Pablo Picasso are important connections in the subgraph.

3 Conclusions

Data from the WikiArt site was used to explore the art collections in this analysis. Discrepancies in the artist data were noted, as the artists in the artists_df and relationships_df were not identical. For further analysis, it is beneficial to ensure that the data is well prepared.

NaN values were handled by discarding the values when nodes and edges were created and replaced with an "Unknown" label when dictionaries for the nodes were created. Artist nodes that did appear in the relationships_df were not properly labeled since no metadata was provided for them in the artists_df.

To get as many significant edges in 2.1, the influenced_by and influenced_on directions were considered when edges were created. A further way to prevent data loss could be to examine whether the analysis can be better conducted in a way that minimizes the risk of losing data due to cataloging mishaps.

For the analysis part, a beginner-friendly approach was chosen. The networks were analyzed with the degrees of the nodes and basic centrality and density measures. The networks of this analysis were generally loosely connected, except for the subgraph when communities were analyzed.
The size of the graphs resulted in performance issues in the Colab notebook, as the graph size had to be increased to inspect nodes properly. For further analysis, methods to draw interactive graphs can be explored. Currently, the graphs need to be downloaded and zoomed in for a better viewing experience.

Some difficulties were noted while graphs were created, such as overlap of the labels, significantly decreasing the readability of the graphs. Some features were added to improve the readability, such as labels, node sizing, and a node color that ensures a higher contrast between text and background. The edges are visualized with a light grey color to improve the contrast between them and the text. For the nation's graph, labels were dropped for artists.

The relationships were explored, showing that the network of artists, institutions, and movements they are associated with are loosely connected. This was expected, given the high counts of NaN values for influenced_on and influenced_by columns that were core values of this analysis.

As WikiArt is a platform where any art enthusiast can contribute to the categorization of artworks, there is always a risk of missing or inaccurate cataloging. However, their goal to cover art as a global phenomenon is ambitious, and it has the potential to provide even more insight with larger data sets as more artists are added.