Mini Project 3 Report

Maša Ćirković

I. Introduction

Processing large amounts of data has become a norm nowadays, given that a great number of devices generate data. Collecting the data, cleaning and standardizing it, and then visualizing it are all part of the standard process when working with data.

Visualization is arguably the most important part, because humans respond to and process visual data better than any other type of data. In fact, the human brain processes images 60,000 times faster than text, and 90 percent of the information transmitted to the brain is visual, as said in this article. Different visualizations types work better for different target groups.

This project works with data from WikiArt, an extensive online art encyclopedia. This dataset provides an opportunity to explore and analyze various aspects of the art world, including artists, movements, institutions, nationalities, as well as examine their interrelationships and associations.

Graphs provide an intuitive way to visualize relationships between different entities, and to follow paths between entities in order to make conclusions. In this project, library *networkx* was used to create a directed graph with entities from the dataset and to present different observations, such as the most influential artists, movements, and institutions.

II. Methodology

A. Data Processing and EDA

Data processing involves understanding what the data is about, exploring different data properties and transforming it into a suitable representation for further analysis. Exploratory Data Analysis (EDA) gives an overview of data statistics and plots which help visualize mentioned statistics.

1) Data Explanation: The dataset consists of 4 csv files: artists, schools, institutions, and relationships. Their characteristics are given below:

- **artists**: URL of the artist at WikiArt, id, image URL, nation, name, total of art work, interval of active years. In total 2996 rows and 7 columns.
- **schools**: Name and school URL at WikiArt. In total 220 rows and 2 columns.
- **institutions**: City, country, name, URL of the institution at WikiArt. In total 73 rows and 4 columns.
- **relationships**: URL of the artist at WikiArt, list of friends, list of artist that they were in influenced by, list of artist that they influenced, list of art institutions that the artist studied, list of schools that was part of, type (artist or collection). In total 2996 rows and 8 columns.

2) Data Cleaning: First presence of NaN values is given in the table I, as a sanity check for future results.

Dataset	Feature	Missing Values
Artists (2996, 7)	url	0
	id	0
	image	0
	nation	32
	title	0
	totalWorksTitle	0
	year	1
Institutions (73, 4)	city	2
	country	2
	title	0
	url	0
Schools (220, 2)	title	0
	url	0
Relationships (2996, 8)	artistUrl	0
	friends	2580
	influenced_by	2512
	influenced_on	2637
	institutions	2362
	movements	40
	schools	1966
	type	1

TABLE I: *Summary of missing values in datasets*

Each of the loaded dataframes were manually looked through and cleaned. Changes are given below.

- **artists**: The column totalWorksTitle was changed to a numerical column after words 'artworks' and 'artwork' were taken out. Nationalities which are the same but differ in the sense of singular or plural form were merged into one, using the singular form.
- **schools**: No cleaning was done.
- **institutions**: Country names were standardized and all American states were changed to USA.
- **relationships**: No cleaning was done.

3) EDA:

- **artists**: Nationalities with more than 50 artists plot is given in 1. And top 10 nationalities versus the total works is given in 2.
- **schools**: Top 10 schools by the number of artists are given in 3.
- **institutions**: Number of institutions per country is given in 4.
- **relationships**: Movements with more than 50 artists are given in 5. Institutions that more than 10 artists attended are given in 6.

B. Network creation and analysis

Directed graph was chosen for the network in order to represent influenced_on relationship, since other relationships

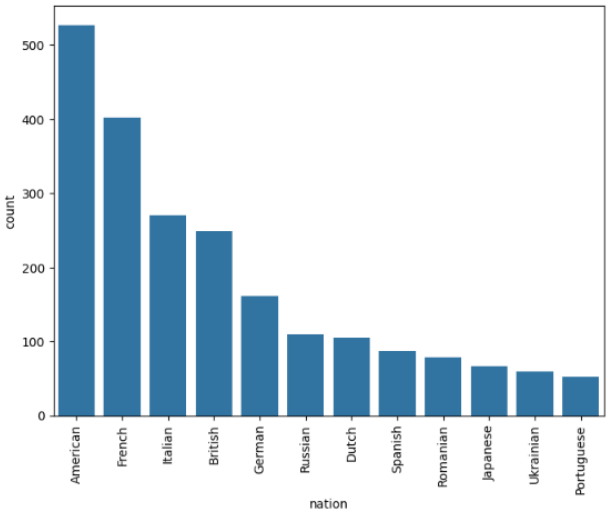

Fig. 1: *Nationalities with more than 50 artists*

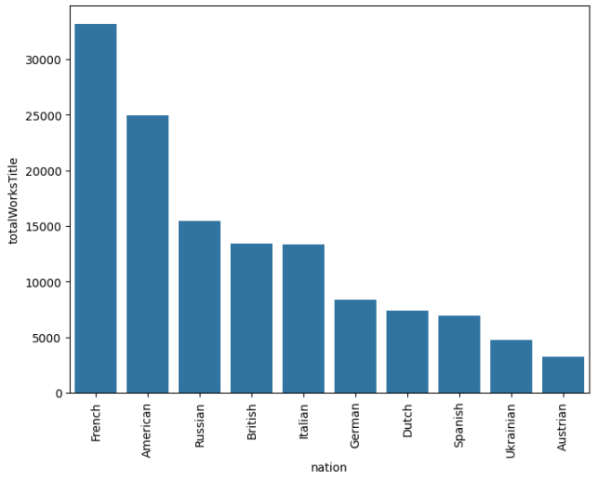

Fig. 2: *Nationalities VS total works*

School	
	0
École de Paris	75
New York School	51
Degenerate art (exhibition, held by the Nazis in Munich in 1937, named to inflame public opinion against modernism)	35
Peredvizhniki (Society for Traveling Art Exhibitions)	27
Mir Iskusstva (World of Art)	26
Dutch School	25
Abstraction-Création	24
Florentine School	22
Zero	21
Flemish School	21

Fig. 3: *Top 10 schools by the number of artists*

can be represented as bidirectional relationships as well.

There are 4 different types of nodes: artists, institutions, schools, and movements.

The only bidirectional relationship is the friend relationship between artists. Influenced_by was swapped so that it can become influenced_on, in case there are some missing values and not every influenced_by corresponds to influenced_on. Artists

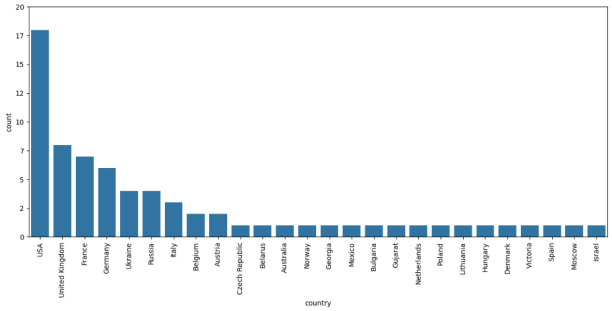

Fig. 4: *Institutions per country*

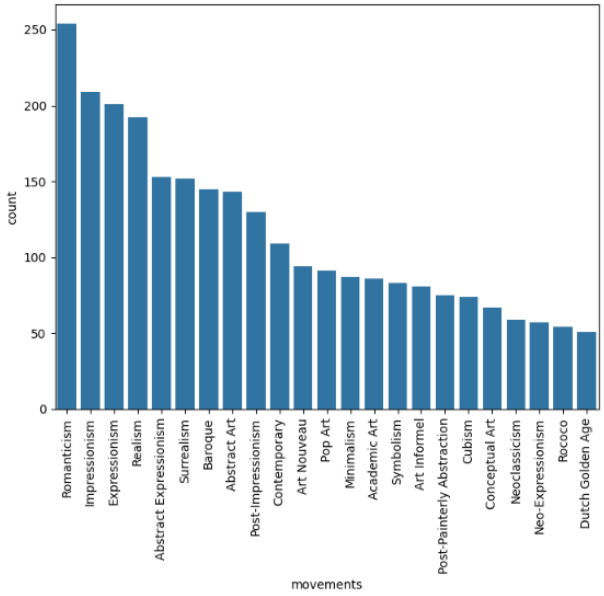

Fig. 5: *Movements with more than 50 artists*

are connected to institutions with a studied_at relationship, to schools with is_part_of, and to movements with movement relationship.

In the end the graph had **3467** nodes and **8154** edges.

Interactive graph was created which allows users to choose the node types they want visualized, and the choices are: all, artist, institution, school, movement. All of the subgraphs created include artist nodes because all of the relationships are based on them and the directed relationships wouldn't show up otherwise. The other criteria is the minimal degree a node has to have in order to be included in the subgraph, and this controls the readability of the graph. Degree of the node is represented by its size and color, with nodes with a higher degree being bigger and darker. Colorway is Oranges and a color bar is given on the side of the plot.

Artists only with a minimum degree of 20 (all possible relationships combined) are given in 7, and with a minimum degree of 30 in 8. Institutions with minimum degree 15 are shown in 9, and with a minimum degree of 3 (where we can also see the artists who attended at least 3 institutions) in 10. Schools with the minimum degree of 20 in 11. Movements

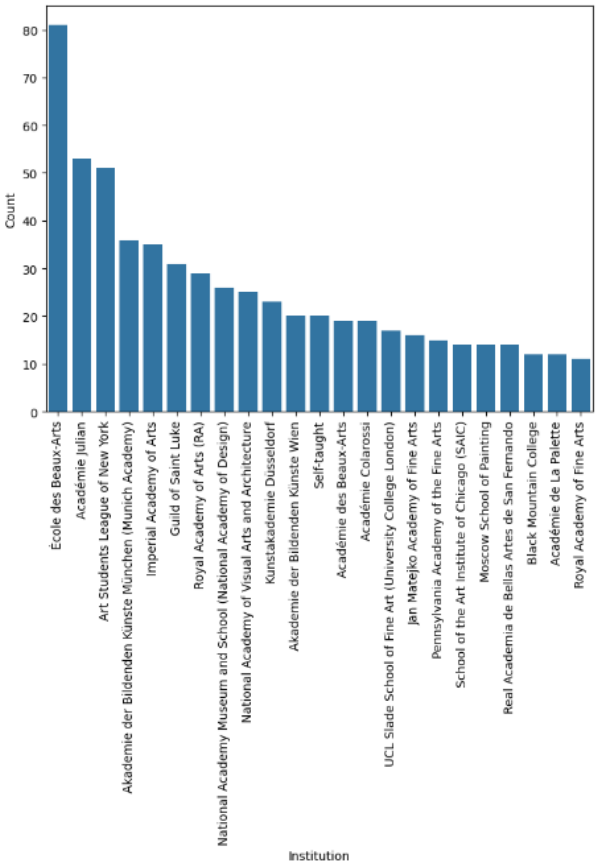

Fig. 6: *Institutions attended by more than 10 artists*

with minimum degree of 3 along with some artists who were part of at least 3 movements in 12, and movements with minimum degree of 30 in 13.

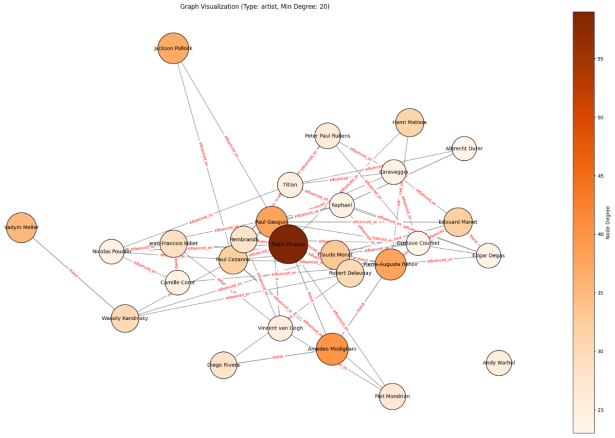

Fig. 7: *Artists only with a minimum degree of 20 (all possible relationships combined)*

III. Results

1. Which were the most influential artists? The top 5 most influential artists and their influence count are given in

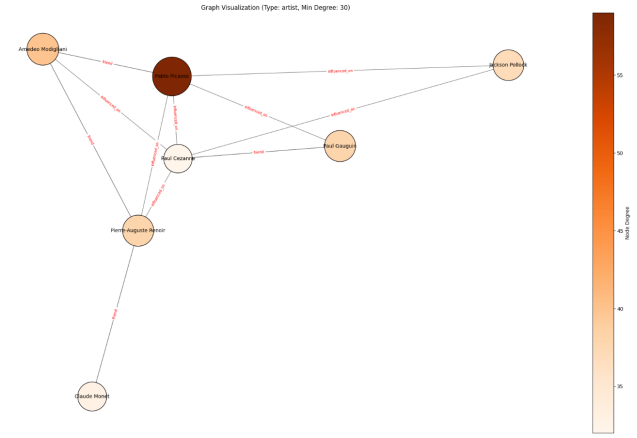

Fig. 8: *Artists only with a minimum degree of 30 (all possible relationships combined)*

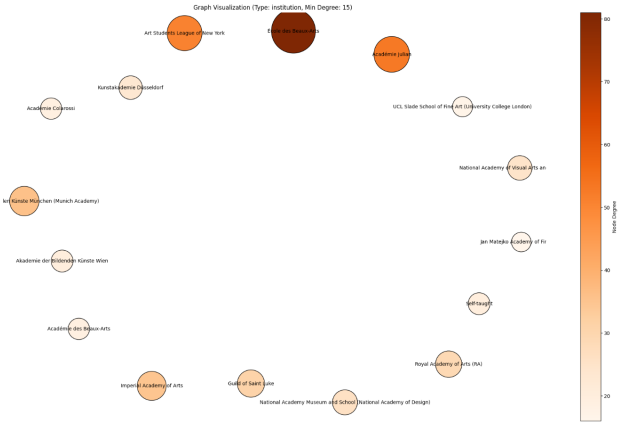

Fig. 9: *Institutions with minimal degree 15*

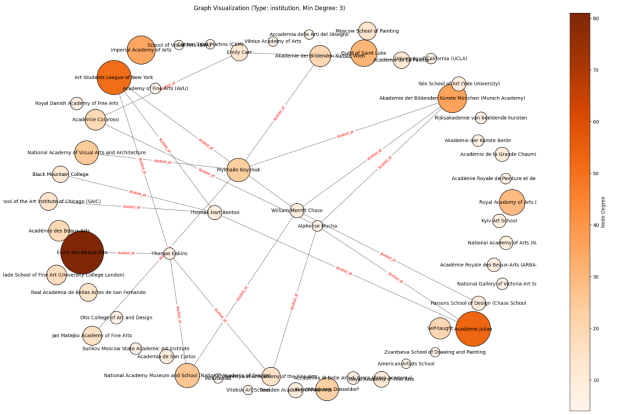

Fig. 10: *Institutions with minimal degree 3*

the table II. This output is to be expected. Influence count was calculated by summation of influences_on and friend relationships, since one is a direct influence and the other is a potential indirect influence. Some influencers were schools, but those were not included in the final table. Visualization of all the relationships and nodes which these relationships

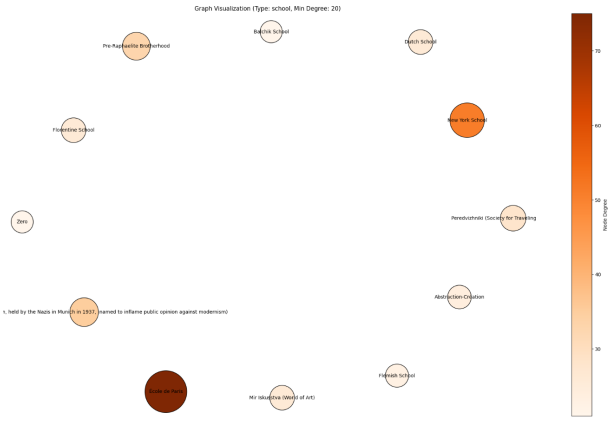

Fig. 11: *Schools with the minimum degree of 20*

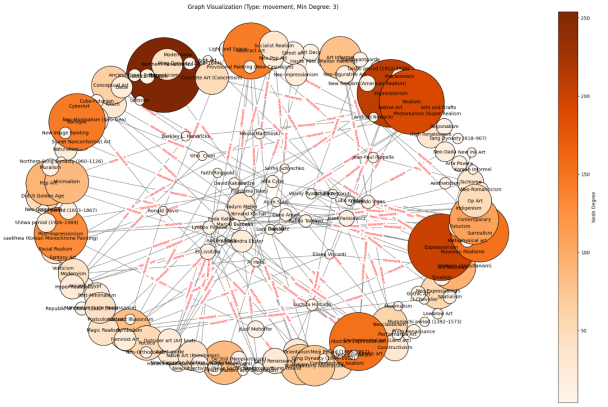

Fig. 12: *Movements with min degree of 3 along with some artists who were part of at least 3 movements*

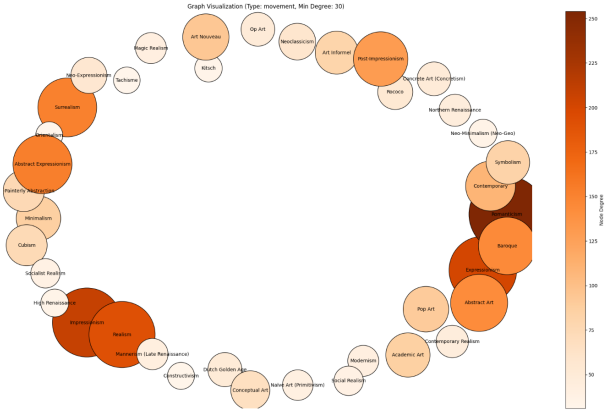

Fig. 13: *Movements with min degree of 30*

connect for Pablo Picasso, the most influential artist, is given in 14.

Artist	Influence Count
Pablo Picasso	30
Paul Cezanne	23
Caravaggio	21
Rembrandt	20
Titian	18

TABLE II: *Top Influential Artists and Their Influence Count*

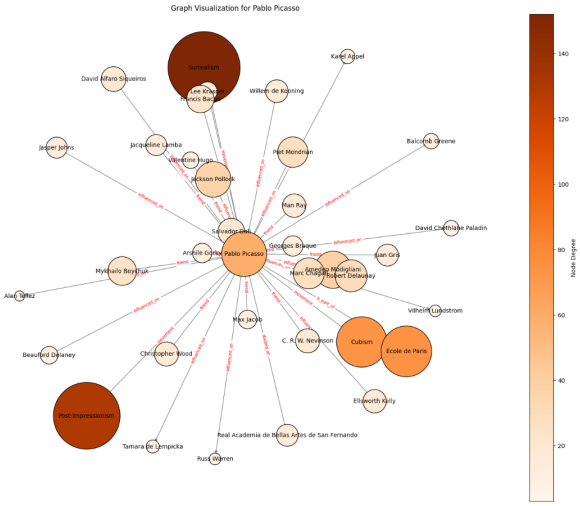

Fig. 14: *Pablo Picasso's network - The most influential artist*

2. Which were the most influential movements? The top 5 most influential movements and their influence count are given in the table III. Influence count was calculated based on how many artists had a relationships to this particular movement (were part of the movement), summed with the influence of those artists, which was discussed previously. Visualization of all artists which were part of Romanticism, the most influential movement, is given in 15.

Movement	Influence Count
Impressionism	430
Romanticism	378
Realism	360
Expressionism	359
Baroque	348

TABLE III: *Top Influential Movements and Their Influence Count*

3. Which were the most influential institutions? The top 5 most influential institutions and their influence count are given in the table IV. Influence count was calculated based on how many artists had a relationships to this particular institution (studied at this institution), summed up with the influence count of those artists, which was previously discussed. Visualization of all artists who studied at École des Beaux-Arts, the most influential institution, is given in 16.

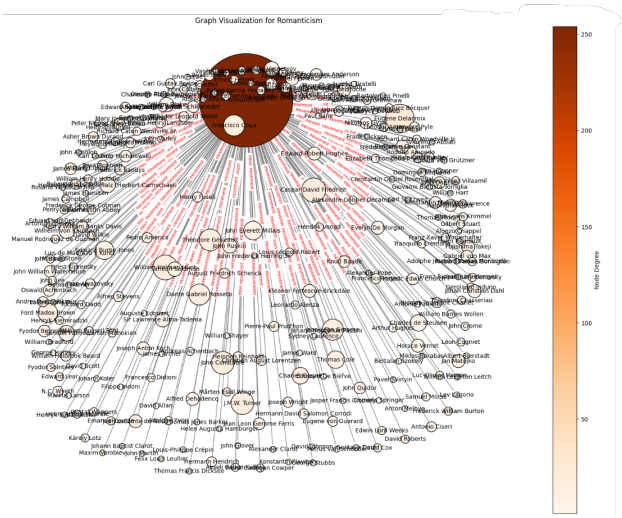

Fig. 15: *Artists belonging to Romanticism - The most influential movement*

Institution	Influence Cou
École des Beaux-Arts	244
Académie Julian	136
Akademie der Bildenden Künste München (Munich Academy)	123
Art Students League of New York	123
Guild of Saint Luke	81

TABLE IV: *Top Art Institutions and Their Influence Counts*

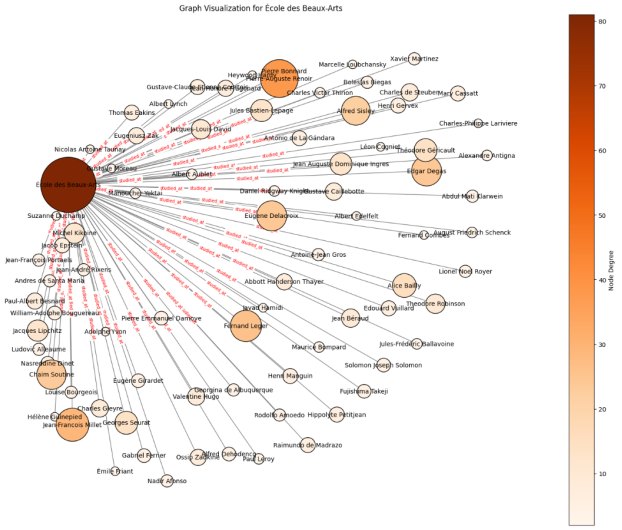

Fig. 16: *Artists belonging to École des Beaux-Arts - The most influential institution*

4. Which nationalities concentrate the majority of artists? The top 5 nationalities which concentrate the majority of the artists are given in the table V. Visualization of Italian artists and their inter relationships with a minimum degree of 5 is given in .

Nationality	Count
American	527
French	402
Italian	270
British	249
German	161

TABLE V: *Top Nationalities of Artists and Their Counts*

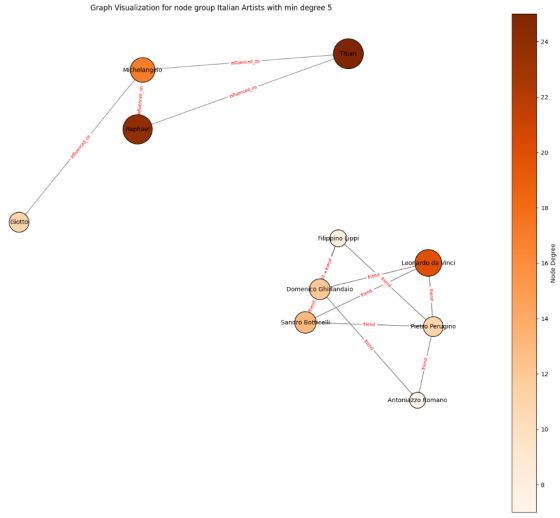

Fig. 17: *Italian artists with a minimum degree of 5 and their inter relationships*

5. Which are the biggest communities in the network? Communities were performed on the subset containing only artists nodes. A few different algorithms were compared and those are: louvain_communities, greedy_modularity_communities, and fast_label_propagation_communities. The number of communities was inspected and 5 largest communities for each algorithm, along with the most influential person from a community, are given. Number of communities for each algorithm is given in VI. Top 5 biggest communities with their most influential person for each algorithm is given in VII. Influence of artists is calculated the same way as discussed previously in the most influential artists. **Note that the outputs are not the same each time**, even though the data provided is the same, which suggest some randomness in these algorithms. As it can be seen, the largest communities tend to be similar, in terms of the influential people they include. When run multiple times, most of the largest communities include: Pablo Picasso, Caravaggio, Raphael, Jackson Pollock, Titan, and Wassily Kandinsky, Paul Cezanne.

The largest community for each method is visualized. Louvain largest community 18, Greedy Modularity largest community 19, Fast Label Propagration largest community 20.

Algorithm	Number of Communities
Louvain	2256
Greedy Modularity	2260
Fast Label Propagation	2331

TABLE VI: *Number of Communities Found by Each Clustering Algorithm*

Algorithm	Community Size	Most Influential Artist	Influence
Louvain	112	Caravaggio	21
	105	Pablo Picasso	30
	60	Raphael	16
	55	Vincent van Gogh	13
	54	Jean-Francois Millet	17
Greedy Modularity	153	Pablo Picasso	30
	116	Caravaggio	21
	112	Jackson Pollock	14
	83	Raphael	16
	57	Edgar Degas	12
Fast Label Propagation	64	Pablo Picasso	30
	59	Pierre-Auguste Renoir	16
	44	Jackson Pollock	14
	31	Raphael	16
	20	Caravaggio	21

TABLE VII: *Most Influential Artists in the Top 5 Largest Communities for Different Clustering Algorithms*

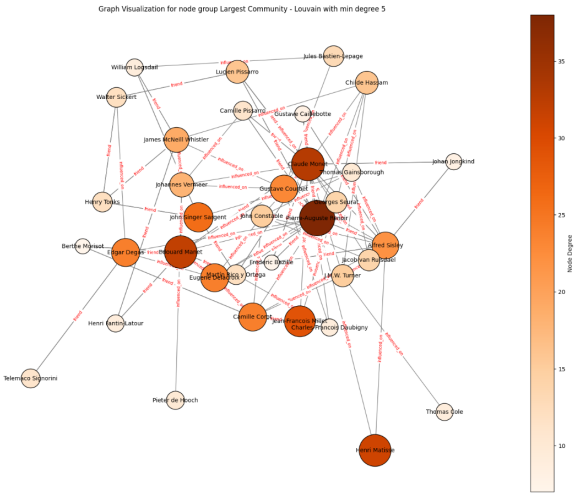

Fig. 18: *Largest community from Louvain method with minimum degree 5*

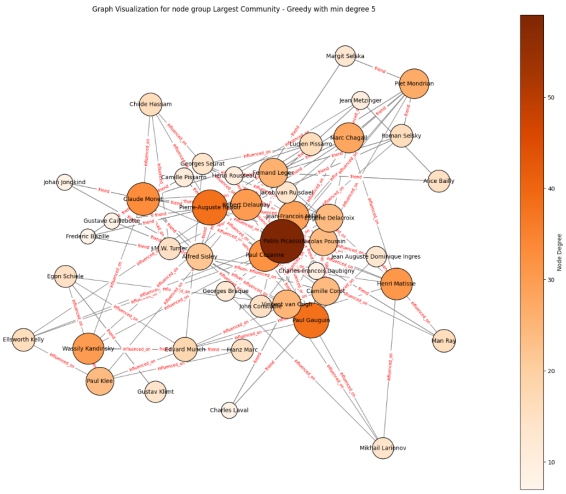

Fig. 19: *Largest community from Greedy Modularity method with minimum degree 5*

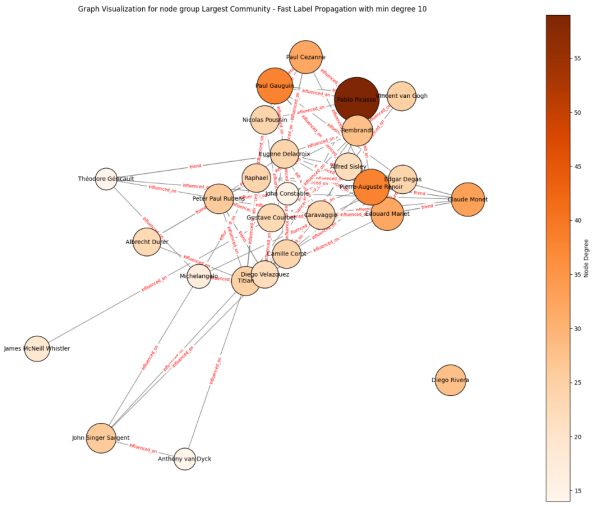

Fig. 20: *Largest community from Fast Label Propagation method with minimum degree 10*

IV. Conclusion

In general graphs provide a very convenient way to visualize data, because of the flexibility they offer in terms of which entities and relationships users want visualized. The easier part is the code behind this project, but the harder part is the logic for modeling entities and relationships. Specifically, determining what should be an entity and what can just be an attribute to a relationship or a node attribute. Those are some design choices that are up to engineers writing the code.

The results obtained support the initial expected outcomes for anyone who is familiar with the history of art. Future work could relate to exploring transient relationships. If person A influenced person B and person B influenced person C, did person A also influence person C and how should that be accounted for? Graphs again are a good way to do this because all possible paths to node X can be explored in order to find transient connections and sum them up.

Finally, graph visualization, although informative and captivating, should be used with caution since it can have problems with large networks. Using subgraphs or limiting the minimum degree of nodes can be different ways to account for this problem.