Project Report: Building a Network of Painters

Summary

The aim of our project was to create a dynamic painter network for analysis, stored in a graph-based database. By integrating two sources of data: the PainterPalette [1] dataset, which provides information about painters, including quantitative data on styles and movements, and coexhibitions of artists based on the e-flux website [2], we can create a network of artist coexhibitions. Using KNIME [3], the datasets were standardized, enriched through the Wikidata API [4], stored in a Neo4j [5] database for analysis, and visualized in Gephi [6]. Key research questions focused on identifying artist communities based on communities, understanding which regions are most influential in the network, and exploring common attributes over the graph. This project provides a framework for analyzing painter networks.

Technical choices:

The pipeline to extract, transform and load (ETL) and analyze data was created in KNIME, involving four major stages: data reading, data cleaning, API enrichment, and Neo4j-based analytics. Neo4j is the most prominent graph database, ideal for storing our network using their AuraDB Server, for free. Additionally, Gephi was used for visualization to get the insights into artist communities, regional influence, and collaboration trends.

The first stage in the pipeline involves data import and preprocessing, where quantitative data was imported from the PainterPalette dataset and artist-exhibition pairs from e-flux announcements using File Reader and JSON Reader nodes. We used a JSON file from GitHub created recently that collected the exhibition and artist pairs. The process generated two raw datasets: one with stylistic and quantitative painter attributes and another with artist-exhibition links. These represent the nodes and edges in our network. Filters and transformations using nodes such as Column Filter and GroupBy, were applied to ensure the data was structured well. The data cleaning and integration stage focused on standardizing names for joins, increasing join matches by 140. The two artist tables were then joined into a single one, including all attributes. Processed outputs are saved and loaded in, to increase recurrent runs.

The API enrichment stage used the Wikidata API to add artist details such as nationality, birth date and place. A Python script handled SparQL queries and processed data in batches to address API performance limitations. This method provided detailed metadata essential for network analysis and visualization. The cleaned and enriched dataset was integrated into a Neo4j graph database, where artists were represented as nodes and co-exhibitions as edges. Neo4j's querying capabilities enabled network analysis to identify patterns of artistic collaboration, supporting interactive exploration and visualization of the painter network and achieving the project's goal of creating a dynamic and insightful system.

We have also designed the pipeline to make use of already processed data, and only query the API and load into the database new information, i.e. do not run the same artist queries again. Outside the pipeline, Gephi was used for network visualization to apply modularity clustering and centrality analysis. It was chosen for its ability to process large datasets and integrate with Neo4j, allowing clear visualization of artist connections and regional influences.

Data model:

The data model is designed as a graph structure with nodes representing artists (label: Artist) and edges representing coexhibitions (label: "COEXHIBITED_WITH"), to show connections between artists who participated in the same exhibition. Each artist is represented by a single node.

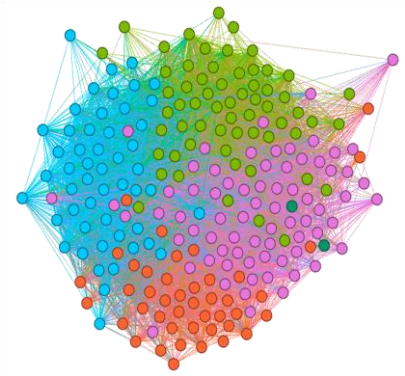
Analytics

The analytics conducted through Gephi and Neo4j provided a comprehensive understanding of the artist network. Gephi's modularity clustering highlighted distinct artist communities based on co-exhibition patterns

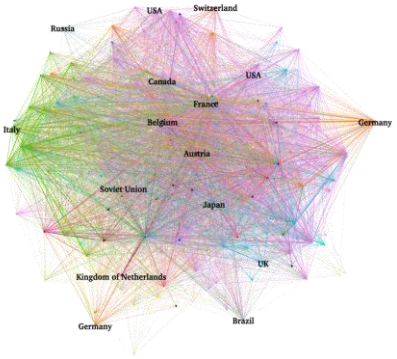
(Graph 1), while centrality analysis identified geographic regions, such as Germany, France, and the USA, as key hubs for artistic collaboration (Graph 2). The visualizations also showed the connections between clusters and how some artists link different groups.

Neo4j analytics offered deeper insights into the dataset. Queries revealed that Andy Warhol, Liam Gillick, and Hito Steyerl were the most frequently exhibited artists, collectively participating in over 750 exhibitions. It was also found that most artists in the network died in New York, while the majority were born in Europe, highlighting the significance of Europe as a birthplace for artists and New York as a major cultural hub. Additionally, movements such as Conceptual Art, Expressionism, and Abstract Art were identified as creating the strongest connections between artists, while 20% of the artists showed minimal network engagement. These results highlighted important patterns in artistic collaboration and influence.

Network visualization of artists based on modularity clustering and citizenship centrality



Graph 1



Graph 2

Conclusion:

This project created a painter network by integrating datasets and analyzing co-exhibition patterns. Using KNIME, the data was cleaned, standardized, and enriched with additional details through APIs. The enriched data was stored in a Neo4j graph database, where artists were represented as nodes and co-exhibitions as edges. Gephi was used for visualization to identify artist communities and influential regions. The results showed key hubs like Europe and New York, frequently exhibited artists such as Andy Warhol, and strong connections in movements like Conceptual Art and Expressionism. This network provides a practical framework for analyzing artistic collaboration and studying trends in the art world.

References:

[1] 'PainterPalette - extensive dataset of painters'. Accessed: Dec. 04, 2024. [Online]. Available: https://github.com/me9hanics/PainterPalette/

[2] 'e-flux online journal'. [Online]. Available: https://www.e-flux.com/

[3] M. R. Berthold *et al.*, 'KNIME - the Konstanz Information Miner: Version 2.0 and Beyond', *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, 2009, doi: 10.1145/1656274.1656280.

[4] 'Wikidata: The Free Knowledge Base'. 2024. [Online]. Available: https://www.wikidata.org/

[5] Neo4j, *Neo4j - The World's Leading Graph Database*. 2012. [Online]. Available: http://neo4j.org/

[6] M. Bastian, S. Heymann, and M. Jacomy, 'Gephi: An Open Source Software for Exploring and Manipulating Networks', in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2009. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154