1 Project Architecture

The aim of the project is to analyze a large social network to uncover its latent properties. You should implement a tool that integrates the full pipeline of social data from data collection to analytical visualization. The project should follow an architecture similar to the one depicted in Figure 1. Below, we describe in detail all the components of the project.

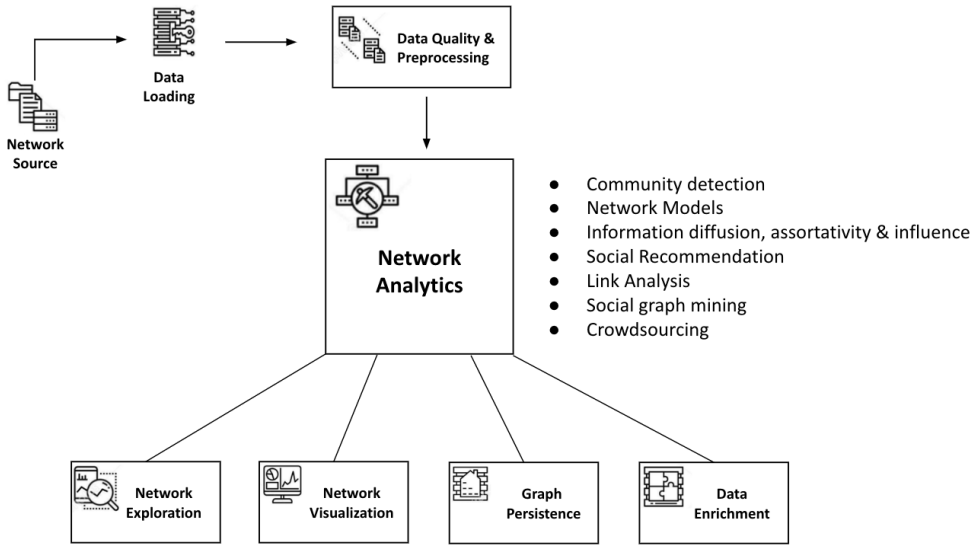

Figure 1: Architecture of the SMA Project. The pipeline consists of network analytics as a main component supported by different components

2 Project Components

2.1 Main Component: Network Analytics

Network analytics is important for extracting interesting findings and inferring useful knowledge from a network. Along the course, you will be exposed to a range of network analytics algorithms that are applicable to real-world scenarios. Examples of analytics include *community detection or information diffusion* which are used to design effective marketing tactics and influencer outreach campaigns. Another important class of network analytics algorithms is *graph mining*, which is applied to expose unusual behavior in social networks, from rumor and mis-information detection to malicious activities such as online trolls and bots. Furthermore, *recommendation algorithms* provide insights that are tailored to the interests of individuals.

The main task here is to implement a network analytics algorithm to uncover interesting patterns and reveal key properties of a network's behavior. You are encouraged to select from a wide range of algorithms you will study throughout the course, e.g., community detection, social mining, social recommendation, information diffusion, etc. The syllabus of the course can be useful in this regard. You can also implement a new interesting network analytics task, not covered in the course, that you deem relevant as a project idea (e.g., sentiment analysis, truth inference, etc.).

2.2 Data Loading

Social networks capture information about social entities, locations, dates, interactions, and much more. Social data can be extracted from a variety of sources. We provide you with a non-exhaustive list of such sources for real world social data: snap repository, network repository, aminer repository, recommendation repository, kaggle repository.

The main task here is to select a network source that fits your analysis and load the network into your tool. Make sure that the size of your network maps to the needs of your analysis. A too large network might carry lots of information, but would incur a time-consuming analysis. A too small network will allow fast analytics, but would limit the quality of the results.

2.3 Data Quality and Preprocessing

Oftentimes, the data provided for data mining is not immediately ready. When preparing data for use in data analytics algorithms, various data quality problems can occur. The existence of such problems might severely hinder the accuracy of the analytics task. Those data quality problems include: i) *Noise* which represents the distortion of the data, ii) *Outliers* which represent instances that are considerably different from other instances in the dataset, iii) *Missing Values* which represent feature values that are missing in instances, and iv) *Duplicates* which represent multiple instances with the exact same feature values, e.g., duplicate tweets or profiles on social media sites with duplicate information. Depending on the context, these data problems might be either fixed or ignored.

Once quality checks are performed, the next step is preprocessing to prepare the data for mining. This step includes Aggregation, Discretization, Feature Selection, Feature Extraction, and Sampling. Several sampling techniques can be applied in this context such as Random sampling, Sampling with or without replacement, or Stratified sampling.

The main task here is to implement data quality checks on your data and apply preprocessing steps to transform your data. Once the data inconsistencies have been fixed and the raw data is properly pre-processed, you can feed the resulting data to your analytics.

2.4 Network Exploration

Exploring data interactions between nodes reveals important hidden patterns. Network measures can be used to untangle the interactions between nodes to understand structural properties and similarities. Two main factors might impact the choice of the exploration method. First, the choice of exploration should not be agnostic to the task in hand. For your network analytics, it might be useful to understand the influence of nodes and how they are related, which can be explored using, for example, centrality or similarity measures.

Second, when applying network measures, you might find that your algorithm is too expensive to execute, as you need to explore all possible relations between nodes. A common way to facilitate your data exploration is to operate on a simulated smaller size network that preserves the properties of the original network. If needed, you can operate on a simulated smaller scale graph with similar properties to the original one.

The main task is to implement network measure algorithms to explore the network and provide revealing insights. It is important to choose appropriate network measures that fit your analysis. Depending on your network size and analytics, you might need to operate on a simulated smaller graph.

2.5 Network Visualization

Visualization tools are important to understand and reveal data patterns. For effective visualization, it is important to draw attention to specific observations that reveal key findings. This can be achieved in various ways. First, plotting data distributions helps you formulate hypotheses about the network's

behavior. Second, projecting different analytical dimensions together can reveal hidden data associations that were difficult to detect before. Finally, displaying the output of an algorithm might help assess the quality of its results. For example, visualizing discovered communities or visualizing the spread of information through a network, can facilitate understanding the results.

The main task here is to use visualization tools to explore your network, bring clarity to your analytics, and highlight the findings you would like to present.

2.6 Graph Persistence

Using a graph database throughout your project gives you two main advantages. First, You can execute a wide range of graph queries to facilitate your graph processing operations. Examples of these queries include traversing your network, retrieving information about nodes, or updating labels and properties of nodes and edges. Second, graph databases allow to efficiently execute large-scale analytics. They follow a graph-oriented data model, where data is stored as nodes and edges, which improves the execution time of graph processing tasks. Examples of popular graph databases are Neo4j, OrientDB, Sparksee, and Titan.

The main task here is to select a graph database to store and query your data. Neo4j provides tutorials[1] to get started with the database. Such tutorials will expose you to a query language, and will walk you through a set of examples to explore a social network (e.g., movie) with the interactions between nodes (e.g., movies and people).

2.7 Data Enrichment

Data enrichment or augmentation is the process of enhancing existing information by supplementing additional data. Typically, data enrichment is achieved by using external data sources or crowdsourcing platforms. While the types of data enrichment vary depending on the type of the social network, this step might have several benefits including i) improve the quality of existing data by inferring missing data or correcting outliers, ii) optimize data usefulness by adding new information, and iii) better understand the social entities and gain deeper insights into their connections.

The main task is to search for external data sources which can be merged with your current network. This task becomes relevant when the data quality and pre-processing did not achieve a high level of cleansing.

3 Requirements

Your project needs to comply with the pipeline described in Figure 1. It needs to implement the analytics algorithms from scratch and compare its performance against other existing algorithms (from a the course or other sources).

You can use helper libraries throughout the project to facilitate implementing your tool. For example, libraries can be used for data loading, graph manipulations, etc. Here are examples of graph manipulation libraries: NetworkX, SNAP and iGraph and visualization libraries: Matplotlib, vis.js, d3.js and Gephi.

4 Evaluation Criteria

- Originality of your project.
- Fulfillment of the basic requirements as detailed in this document.
- Quality of your code (structure, design, consistency, scalabilty, correctness, etc).

[1] https://www.cl.cam.ac.uk/teaching/1617/DatabasesA/graph-db-setup.html

- Working demo of your tool.
- Quality of the report. Analysis of the efficiency, effectiveness, and the correctness of the implementations.

All team members should be present during the discussion. Each team will be composed of **two-three** students and members can be graded individually.

5 Deliverables

- All scripts and Readme file.
- A report of minimum 4 pages that includes:
  - Introduction of the task and its applications.
  - Pipeline outlining the components of the project.
  - Description of the data.
  - Description of the project components and implementation choices.
  - Analysis of the results and findings.
  - Evaluation of the performance of implementation and its correctness.
  - Potential limitations of the method and new ideas/directions to improve it.
  - Github link to your code.

6 Project Idea Presentation

- The project idea presentation will take place on March 12, 2024.
- Each team should prepare a 10 minutes presentation describing the project idea.
  - Include the names of team members.
  - Introduce the main idea and components of your project.
  - Describe the data and type of analytics that will be used.

7 Project Checkpoints

- Discussions with each group about the current stage of the project and eventual problems.
- Take place after the lab on April 16, 2024 and May 7, 2024.

8 Timeline

Date	**Task**
March 11, 2024	Upload presentation of the project idea to ILIAS
March 12, 2024	Project idea presentation
April 16, 2024	Project checkpoint 1
May 7, 2024	Project checkpoint 2
May 24, 2024	Upload report to ILIAS
May 28, 2024	Final project discussion