Social Network Analysis

Project Report – COS70008

Student Name: Rajkishor Singh Naruka

Student ID: 104216694

Group Number: G13

Date of Submission: 22 May 2024

Word Count: 2410

# Table of Contents

[Table of Contents 1](#_Toc1090996460)

[Introduction 2](#_Toc820479941)

[Project Background and Motivations 3](#_Toc1229021466)

[Design concept 3](#_Toc1863277885)

[Problem statement 3](#_Toc1072375766)

[Design Specifications 3](#_Toc2038341266)

[Design Implementation, Processes (Hardware and Software) 4](#_Toc155812000)

[Design benefits: 5](#_Toc123793412)

[Project reflection 6](#_Toc737461948)

[Group work reflection: 7](#_Toc813902897)

[Effective Strategies: 7](#_Toc1646335434)

[Approaches that did not work: 7](#_Toc1069176558)

[Future Improvements: 7](#_Toc2129424522)

[Outstanding Events/Actions 7](#_Toc637381054)

[Individual work reflection 8](#_Toc83022618)

[Project Tasks 8](#_Toc73136802)

[Contributions To the Group 9](#_Toc1708748263)

[Conclusion and Recommendations 9](#_Toc547240201)

[References 10](#_Toc1050026311)

[Appendix 10](#_Toc1797682946)

# Introduction

## Project Background and Motivations

The project focuses on understanding social connections through Social Network Analysis (SNA), a method to study networks like groups of friends or business partners by analyzing their connections. The aim is to create an interactive web application for social network analysis, allowing users to collect and upload network data, perform descriptive network statistics, visualize networks, and use statistical models like ERGMs or ALAAMs. This tool will benefit researchers, data analysts, and social scientists by providing an efficient and interactive solution for social network dynamics.

## Design concept

### Problem statement

The challenge is to make SNA more accessible and user-friendly for researchers, data analysts, and social scientists. The main objective is to create an interactive and easy-to-understand tool that integrates advanced statistical models like ERGMs and ALAAMs, as well as network visualization capabilities. The tool should provide a user-friendly interface that allows users to perform in-depth statistical analysis and generate reports, empowering researchers and analysts with a more efficient and effective tool for social network dynamics.

### Design Specifications

The tool has a user-friendly interface with clear navigation and intuitive controls and instructions.

* **Uploading the data:** Users canupload data in CSV or Excel format. The app guides users on file structure. CSV files can include network data (connections), while Excel files can hold multiple networks and node attributes (additional information).
* **Network Visualization:** The app creates a dynamic interactive network visualization with node statistics displayed as color density and annotations. Users can select specific networks and node statistics to view.
* **Descriptive Statistics:** The app automatically calculates and displays various network statistics. Users can choose which statistics to see.
* **Statistical Modelling:** Users can analyze networks using three ERGM models: Bernoulli, NodeMatch, and Node Covariate. Model selection depends on uploaded data (network-only files can only use Bernoulli). Goodness-of-fit results help users assess model accuracy (optional display).
* **Community Visualization:** User can select to visualize different communities present in the network. The app calculates all the communities and community statistics in the background and allows user to visualize it. Each community is represented by different colour and community statistics can be hidden/shown by the descriptive statistics pane.
* **Report Download:** The App generates downloadable reports summarizing all visualized elements and more. Reports include an Excel file with selected network attributes, model summaries, and goodness-of-fit results presented in separate sheets. Additionally, users can download self-contained HTML files with network and graph visualizations, including all necessary CSS and JavaScript code.

### Design Implementation, Processes (Hardware and Software)

The app is deployed on Streamlit cloud and is accessible through any device with internet access. It uses server-side computing and storage, so users do not need to install anything on their devices. Users can either download the app or use it from the website.

The app's design implementation includes the following processes and tools:

* **Landing Page:** Displays file upload instructions, including CSV (network-only) or Excel (multiple networks/node attributes).
* **Data Upload:** Users upload CSV/Excel files, which the app recognizes and loads using pandas. The app then creates a networkX graph from the selected network, performing data cleaning/preprocessing.
* **Network Visualization:** Using a selected network graph and visual matrix, the app generates a network visualization with Pyvis, saved on the server and displayed on the dashboard. Users can download the visualization as a standalone HTML file.
* **Descriptive Analysis:** The app calculates descriptive statistics in the background using NetworkX and allows users to select which stats to display. If the selected network is not fully connected, a warning is shown, but all other metrics are calculated and displayed.
* **Community Visualization:** Similar to Network Visualization, the app calculates communities using NetworkX, colors them with Matplotlib, and displays the visualization on the dashboard. Users can download the HTML file.
* **Statistical Modeling:** The user can choose one of the three available models. The app then uses an integrated R environment and rpy2 package to run an R script and save results on the server. Then, it loads those results and shows them on the dashboard. If it is a network-only file, the user can only use the Bernoulli ERGM model. If node-level attributes are available, the user can choose NodeMatch or Mode Covariate. If there are any issues with the attribute selection, it will be shown as a warning (e.g., ERGM Node Covariate Analysis is only supported for numeric attributes). It also allows users to perform the Goodness of Fit Test, which uses the same process and allows users to show those on the screen.
* **Report Download:** The app generates an Excel file with different sheets for report sections, automatically extracting usable results from the model summary and converting them into dataframes. The report is built dynamically based on user interactions.

**Note:** At boot time, the app runs a separate R script using the python subprocess module to install R dependencies and create the R environment, which is not accessible to users.

The app uses the following hardware and software:

* Streamlit cloud for deployment - Server-side computing and storage on Streamlit cloud.
* Python 3.11
* R 4.1.0
* RPy2 library for integrating R with Python.
* NetworkX library for network analysis.
* Pyvis library for network visualization.
* Matplotlib library for community coloring.
* Pandas library for data manipulation and analysis.
* NumPy library for numerical computing.
* Openpyxl library for reading and writing Excel files.
* Subprocess module for running a separate R script at boot time.
* xlsxwriter and pandas for report generation
* HTML, CSS, and JS for visualization downloads

### Design benefits:

The design of this social network analysis tool offers several benefits to researchers, data analysts, and social scientists. Here are some of the key benefits:

1. **User-Friendly Interface:** Minimizes training time for researchers with clear navigation and intuitive controls.
2. **Flexible Data Import:** Supports CSV and Excel formats, with plans to incorporate more in the future.
3. **Interactive Network Visualization:**
   1. Zoom and drag nodes for better understanding of the connections between
   2. Select node statistics for annotation on the graph.
   3. Gain deeper understanding of network structure and connections.
4. **Advanced Statistical Models:** ERGM family (Exponential Random Graph Models) for complex hypothesis testing (Bernoulli ERGM, NodeMatch ERGM, and Node Covariate ERGM).
5. **Descriptive Network Statistics:**
   1. Automatically calculates a range of statistics.
   2. Users can choose to display or hide specific statistics based on research focus.
6. **Community Visualization:** Identifies clusters and groups within the network (Using the Louvain algorithm for community Detection).
7. **Comprehensive Report Generation:**
   1. Download a report of everything on screen and more.
   2. Excel file with selected attributes, model summary, and goodness-of-fit results.
   3. Share findings with ease.
8. **Visualization Download:**
   1. Download network and graph visualizations as self-contained HTML files.
   2. Share visualizations conveniently.
9. **Cloud-Based Access and Server-side computing and storage**:
   1. No need for user installation, accessible from any device with internet access. Accessible at: https://yetanothertechproject.streamlit.app/
   2. Uses server-side computing and storage and reduces user machine dependency
   3. Option to download dedicated app on Phone/Tablet and PCs
10. **R Integration:**
    1. Integrated R environment and rpy2 package for running R scripts for statistical modeling.
    2. Flexibility for users who prefer R analysis.

Overall, the design of this social network analysis tool provides users with a comprehensive, user-friendly, and flexible solution for analyzing and visualizing network data. By offering advanced statistical models, descriptive network statistics, community visualization, and report generation, the tool empowers researchers and analysts to gain a richer understanding of network data and uncover hidden patterns and connections.