CS 579: Online Social Network Analysis

Project 1 – Social Media Data Analysis

Deliverable 1 due on January 25, 2024 at 11:59pm (google form)
Short progress report due on February 6, 2024 at 11:59pm (google form)
Final report due on February 20, 2024 at 11:59pm (Blackboard)

This project is to be completed in 2-person teams. The team members should work together on each aspect of the project, including the writeup of the report. Each deliverable should be submitted once per team.

Project Objectives

You will learn how to crawl social media data, consider privacy and data usage implications, process, model and analyze the data. You will write a detailed written report and give a short oral presentation summarizing your results.

Project Outline

1. Data Collection
2. Data Visualization
3. Network Measures Calculation

Guidelines

Data Collection – Your initial task is to choose a social media platorm to collect data from. Some example platforms include instagram, dblp, Reddit, arXiv, ResearchGate, Stackoverflow, Stackexchange, Wikipedia, etc. Figure out how you can crawl data from these websites. Some of these platforms provide an API for collecting data. Make sure you have the needed credentials for scraping the data (i.e. API key).

You should collect enough data to create a social network with 100-500 nodes. Some representative network types are described as follows.

- Friendship Network. A user's friendship network can be represented as a graph that the nodes are the users and the edges show whether there is a friendship relationship between them. Example: Users and connections in LinkedIn.
- Co-authorship Network. The nodes are scientists and two scientists are connected if they have co-authored a paper. Example: An authorship network in the Computer Science category of papers in arXiv.

- Diffusion Network. A node represents an entity which can publish, receive and propagate information. A directed edge between nodes represents the direction of information propagation. Example: Fake news propagation when the nodes are users and the edges are re-tweets/replies/likes.

Your report will include a description of how you crawled your chosen platform to collect the data. Please also describe any challenges you faced, how you overcame the challenges and how the challenges impacted the data that you were ultimately able to collect. Your report should also include the user privacy policy for your chosen social media platform and data usage policy. If you cannot find these policies, please describe where you looked for them.

Data Visualization – Once the data is collected, the next step is to utilize a graph analysis software to visualize your network as a graph. There are many software packages available including networkx [link], snap [link], Gephi [link], NodeXL [link] and graph-tool [link]. Choose one and read the instructions to determine how to input and visualize your graph. Each package may require a particular format (i.e., adjacency matrix, adjacency list, edge list) for input of the graph data.

Your report will include a short description of the graph analysis software that you used, your reasoning for choosing the software and the format of the data input file. You will incude a screenshot of your visualized graph along with any information needed for the reader to understand the visualization.

Network Measures – You will learn different network measures in class (Degree Distribution, Clustering Coefficient, PageRank, Diameter, Closeness, Betweenness, etc.). Use your chosen graph anaysis software to obtain degree distribution and plot it as a *histogram*. In addition to this, choose two other network measures to report on. Choose any two from those that we've learned about. Report on these measures in an appropriate format.

Your report will include a description of how you used the graph analysis software to get each of the three measures along with the measures and corresponding visualizations as appropriate.

Discussion of Results – Your report will include a discussion of the results of the data visualization and network measures. What insights do these results provide? What further questions do these results raise? What would your next step to investigate further be?

Reference – Your report will cite all tutorials, packages, software and libraries you used in your data collection and analysis.

Video – Each team will submit a video (no longer than 4 minutes) where each team member talks about the most significant challenge they faced working on the project.

Submission

We will run your code to see if it works for all of the steps. You should put all of your files including your raw data, your cleaned data, source code files, a report in pdf format and your short video into a .zip folder named LASTNAME1_LASTNAME2_PJ1 (Instead of LASTNAME1 and LASTNAME2 type the lastname of each team member). Submit your zip folder to Blackboard. One submission per team.

Academic Integrity

You must develop your own code for data scraping. It is **NOT** okay to use a publicly available dataset.

You can refer to others' code and use libraries, software and packages but it is **not** okay to copy existing code from others. Be sure to cite any sources you use. Failure to cite sources will be considered plagiarism.