Assignment 1

Amirali Abari
Faculty of Business and IT
Ontario Tech University

January 15, 2025

1 Introduction

This assignment focuses on analyzing social network and information network datasets. For social networks, we focus on small datasets collected/surveyed in the class. For information networks, you will collect your own data from the Web. You may use Gephi[1] for visualization and analyses of small datasets. For large datasets, you can use any API or libraries. The instructor recommends NetworkX[2] or SNAP[3] for large network analyses. Remember that these are only suggestions, if you use any other tools that get the job done is still acceptable.

2 Social Networks Datasets

We have collected 10 social network datasets during the second lecture. The data can be found at this url.[4] The goal is to analyze these datasets by Gephi (or any other software) to extend our understanding of these social networks. You will use the concepts that you have learned so far (e.g., visualization, diameter, degree distribution, etc.) in your analyses.

Pre-processing data. Your data in the current format might need some pre-processing to be fed in Gephi. You need to create CSV file by Excel in a specific format to upload them in Gephi. See here for more details.[5] You can create a shared *node table*, but for each social network you need its own *edge table/list*. This pre-processing phase of the assignment can be done in collaboration of all of your classmates through Discord/Slack. This is the only part of this assignment that you can team up :)

Analyses of 6 networks. You will choose 6 of those collected social networks. For each selected social network, you need to make some representation decisions such as if the network is directed/undirected, weighted/unweighted, etc. You need to justify this decision in your writings. After making these assumptions, you will visualize each social network. Choose the coloring themes, positioning layout, and sizes for nodes so as to able observe some interesting patterns. Make sure to label your nodes too; this makes your analyses more interesting. After visualization, you need to report the following statistics on the networks: number of nodes, number of edges, edge density, degree distribution, average clustering coefficient, number of

[1] https://gephi.org/

[2] https://networkx.github.io/

[3] https://snap.stanford.edu/snap/index.html

[4] Unprocessed: https://docs.google.com/spreadsheets/d/1hcWSzib7foEnlPC7K6OYPipY1a1L_AUNhBoWe4q8ZqE/edit?usp=sharing

[5] https://gephi.org/users/supported-graph-formats/spreadsheet/

nodes in strongly connected component (SCC), number of nodes in weakly connected component (WCC), average path length in SCC, diameter of SCC. Also, run community detection algorithm to detect communities. Then, visualize different communities with different color. Try to understand if the detected communities make sense!

For each network, you will choose two notions of centrality (e.g. Degree, Closeness, or Betweenness) and then report two most central/influential persons under each selected notion of centrality.

The most important part of your analyses is *the interpretation of your statistics and visualizations*. You should explain which new insights you have learned from each statistics, and how your analyses shed lights on more important questions that you can ask about the friendships/relationship of your classmate.

Insights from the collection of networks. You are also expected to have a section for explaining some new insights that you could gain by looking at your six networks together. Specifically, what you can learn from your collection of networks, that you could not learn if you look at each network individually.

3 Information Networks

You are going to collect your own data from Ontario Tech webpages. You are asked to write your own crawling scripts in your favorite programming language. I suggest using Python. You need to be familiar in implementing BFS and working with some Python libraries such as BeautifulSoup and urllib.[6] Using your script, you are going to crawl all web pages that have the domain of "ontariotechu.ca" or "uoit.ca" You can ignore crawling urls which are for images, videos, or pdf files. Just focus on actual web pages.

Task 1. Put your crawling code in your report. Then, you need to write a very thorough explanation of each lines of code. Explain what each line does. Also explain every libraries used at high level. Your explanation should convince us that you have fully understand this code and you can write similar code in future.

Task 2: Analyses. You do the similar analyses that you did in the previous section. You first need to make some representation decisions such as if the network is directed/undirected, weighted/unweighted, etc. You need to justify this decision in your writings. You need to report the following statistics on this network: number of nodes, number of edges, edge density, degree distribution (with and without log-log scale), average clustering coefficient, number of nodes in strongly connected component (SCC), number of nodes in weakly connected component (WCC), average path length in SCC and WCC, diameter (in SCC and WCC). Also, run community detection algorithm to detect communities. Then, try to understand what each community is representing. Try to understand if the detected communities make sense!

Consider all notions of centrality (e.g. Degree, Closeness, Betweenness, and PageRank) that you have learned. Then compute the centrality measures for all nodes. Report top 10 most (and least) central webpages under each notion of centrality.

The most important part of your analyses is the interpretation of your statistics. You should explain which new insights you have learned from each statistics, and how your analyses shed lights on more important questions about UOIT webpages.

Task 3: Visualization (Optional, Bonus Mark). Come up with a creative way to visualize this information network. You might need to first find a way to summarize the network to a smaller network and then visualized the summery network. Of course, for such summarization,

[6] I suggest that you use pip `https://pip.pypa.io/en/stable/` for installing third party libraries.

you need to write some code :-)

4 Grading Scheme and Deliverable

This assignment has 15% of your final grades. You can get up to 2% extra mark. The grades are distributed as follows:

- Social network analyses (detailed in Section 2): 9%
  - Report on statistics: 1.5%
  - Visualizations: 1.5%
  - Interpretation of statistics: 5%
  - Insights from the collection of networks: 1%
- Information network analyses (detailed in Section 2): 6%
  - Crawler: 3%
    - Correct code: 2%
    - Good explanation of the code: 1%
  - Report on statistics: 1.5%
  - Interpretation of statistics: 1.5%
- Bonous Question: 2%.

Deliverable. You should submit the report of your assignment electronically. There is no page limit (or minimum page requirement) for your reports. It will be assessed based on its quality :-)

Policy on Using GPT and Other LLMs. The use of GPT or other language models is *not forbidden* in this assignment. However, I strongly encourage you to *avoid* it or use it *responsibly*. This means you should ensure that any AI-generated content is well-understood and correct. Be aware that the instructor will run the assignment through various LLMs (including GPT-4 and others) to identify any unusually incorrect answers. If your incorrect answers *match* the surprising incorrect answers generated by these models, you will *lose all 15%* allocated to this assignment. Additionally, any use of GPT or similar tools *must be cited* appropriately.

Good luck! and have fun!