PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

Department of Computer Science & Engineering Session : Jan-May, 2023

UE20CS344 – NETWORK ANALYSIS AND MINING

Lab Evaluation 01

Budget : *for a 4 member team, 3x1 hour slots are set aside for this assignment. 12 man hours should be enough for this exercise*

English novels are available in the Project Gutenberg site. Choose one for your assignment. Do not choose a very big novel (Like War and Peace by Tolstoy or Mahabharata as there are too many characters) in it. Do not choose a novel that has only a very limited number of characters. Then your analysis will not be interesting. Similarly, do not submit anything of which the analysis is available in github (and we all know). Choose something that you have probably read and quickly learn the key points in the story from the internet , if not known.

Download the text version of it (not HTML) as you don't want to spend time in unnecessary pre-processing. **Start after knowing the story at a top level.** Now do the following :

1. **Implementation (6 marks) :** ***you are expected to reuse and modify sample code provided. If you do that, it won't take much time***
    a. Make a list of characters in the novel. You need to decide whom to include. For example, for Mahabharata, there is no point in including a character representing a random soldier (;-
    b. Extract a social graph of the manually identified characters in the text (as shown in the hands-on session). To do this, you need to use a co-occurrence algorithm as discussed and shown in the demo in class. Also, plot the graph using networkx (it may be a very dense graph and that is okay).
    c. Calculate the four types of centrality of main protagonists i.e. degree, betweenness, closeness, PageRank . (Ref : Unit 1 – centrality analysis)
    d. Calculate the global clustering coefficient of your graph and local clustering coefficient of the main protagonist nodes. (Ref : Unit 2 – Measures of cohesion)
    e. Detect communities using the following methods: (Ref : Unit 2 – Measures of cohesion)
        i. K - clique (percolation method)
        ii. Louvain community detection
        iii. Girvann Newman
    f. Find the degree distribution, average shortest path, and size of the largest component. Also create equivalent generative models to compare against the social graph that you extracted (Ref: unit 3 - Generative models)
        i. G(n,p) and G(n,m) generated graph
        ii. Preferential attachment
        iii. Small-world model

Department of Computer Science & Engineering Session : Jan-May, 2023

UE20CS344 – NETWORK ANALYSIS AND MINING

Lab Evaluation 01

2. *Analysis* (4 marks) - ***While the implementation above should be fast if you reuse the sample code provided, spend quality time in this section as a team***.

   Theme of the analysis: What you know of the story and is it matching with what you got from your network analysis? Have you got any insight to offer ?

   a. Who are **the protagonists as per your analysis**? If the **4 centralities are not having high correlation, how do you interpret them**?
   b. What do the **clustering coefficients, discovered communities, extracted ego network of protagonists and average shortest path** tell you about the dynamics in the story? How is clustering coefficient related to transitivity of nodes?
   c. Compare all the generated graphs (from (f)) to the actual graph. Is there a difference, and if yes, what can it be attributed to? Also, analyze the differences between the 3 generated graph's attributes.
   d. Feel free to do any appropriate visualization using Gephi **only to substantiate your analysis**

Questions 1a, 1b, 1c and 2a can be attempted after unit 1 has been completed.
Questions 1d, 1e, and 2b can be attempted after unit 2 has been completed.
Questions 1f and 2c can be attempted after unit 3 has been completed.

Shared code can be found here

Submission Instruction:

1. **Submit a Python Jupyter notebook**.
   a. The first cell should have SRN of your team members .
   b. Use Python 3.x and Networkx 2.5 or above
   c. Refer to the networkx API reference while coding as this API keeps changing
2. **In the submitted Jupyter Notebook**
   a. We expect **2 sections in the notebook with appropriate subheadings** (a,b,c, and so on)
   b. Do not add unnecessary details in the **analysis, keep it brief and to the point**.
3. **Late submission will invite penalty**