CS9.435 Computational Social Science
Homework #2

Deadline: 23:55 IST on March 10, 2022
Max marks: 30

Instructions

- Please create a single **typed** PDF answering the following questions and submit it to **Homework #2** assignment on MS Teams. Name the file **<roll-number>_hw2.pdf**.
- This assignment is to be done individually. Please cite any sources you use. All submissions will be tested for plagiarism, and if found, the institute's policies will be followed.
- State any assumptions you make. Please reach out to the TAs on MS Teams in case you have any queries.

Questions

Question 1 [1+3+5+2]

A) What is a representative dataset?

B) You are tasked with building a dataset of Twitter users from your city and analyzing what are the main topics they tweet about. You can refer to this notebook for the basic code to scrape data and create your dataset. Refer to the definition of representative dataset you wrote above and explain why this dataset does or does not fit into that category. If not, list the challenges in building a representative dataset for this task.

C) Outline various kinds of biases you observe in this dataset. Substantiate your observations with tables/graphs to effectively visualize the biases.

D) Propose some methods (at least two) by which these kinds of biases can be avoided while building real-world datasets (for any task).

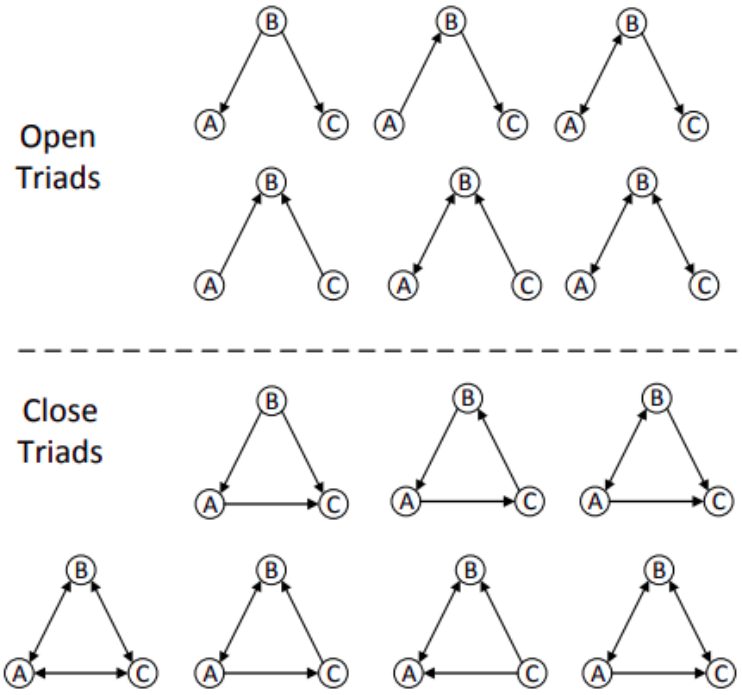
Question 2 [1+9 (3 marks for the graph, 3 marks for each of the two sub-questions)]

A) How do platforms like Twitter, Facebook and Instagram suggest friends to users? List down a few techniques / methods they may be using.

B) Social media platforms greatly connect our physical daily life with the online world. They produce a huge volume of data, including not only the information about user activity and communications, but also user behavioral logs. For instance, group formation – the process by

which people come together, seek new friends, and develop communities – is an important research topic in CSS. The simplest group structure is called a triad, consisting of three people. In a closed triad, for any two people in the triad, there is a relationship between them. In an open triad, there are only two relationships, which means that two of the three people are not connected with each other. Perform the following activity to understand triads in a celebrity network.

Activity: Pick any Indian celebrity and get a list of 100 people they follow on Twitter. Represent each person as a node and connect the nodes with edges. For instance, an edge from node A to node B represents that A is followed by B (you are welcome to use any network analysis tool for this).



Answer the following questions based on your graph. (Please attach your graph in the submission document)

1. Identify the number of open and closed triads in the graph. Explain any emerging patterns and differences between these two kinds of triads. You can read this paper to gain some useful insights about triads and triadic closures.
2. Write your observations (at least 3) on the nature of the graph (any strongly connected subgroups, commonalities among the nodes, possible biases because of the network structure, echo chambers, etc.)

Question 3 [2+2]

Keeping Sahana's Extreme / Hate Speech lecture in mind, please review this paper and list down the following:

A) At least 2 limitations / challenges (if you find more, please feel free to list) in the conclusions that the paper makes, i.e., paper is claiming more than what it really is.

B) At least 2 ways by which you can address the limitations you mentioned about; be precise and explain the data / methodology changes that you will make to fix the limitations.

Question 4 [3+2]

Telegram groups are a common messaging platform and are used as discussion forums for various topics. Multiple groups are created and content/links/images etc. are extensively shared/forwarded on a daily basis. Consider that we wish to collect data pertaining to the spread of misinformation during Covid-19 on telegram.

1. What methodology can be used to exhaustively collect all data from such telegram groups for this? Discuss the details of the same.
2. Mention 2 other use cases where this methodology can be employed.