Social Computing, Evaluation 3

CS60081, Autumn 2021

8:00 am to 9:00 pm, 19th November 2021
Full marks: 36
Answer ALL questions

IMPORTANT INSTRUCTIONS

Taking the exam: You need to log into zoom, keep your video on during taking the test (so that we can monitor you during the exam). You will use pen and paper to write the exam,
Decorum: Throughout the examination, you are strictly expected to have their cameras on, directing towards their workspace including themselves. Arrange your laptops/desktops/mobiles beforehand to save time during the examination. Disconnecting video for a long duration will be grounds for suspecting malpractice.

You need to keep your workplace, your hands and your mobiles visible to us. We are trying to avoid the visibility of your answers in the papers to the rest of them. Once you open your question paper, refrain from using your PC/laptop from searching for anything or typing during the exam.

Tip: Install Adobe scan and, MS Teams on your phone to make the whole process easier. In that case, your laptop acts as a camera, while you are using your mobile for checking the questions, scanning and uploading the answers.

Submission: You can do either of two things (i) take pictures of your answer script pages, name the pictures page1.jpg, page2.jpg, page3.jpg etc., zip the pictures and upload the zipped file via CSE Moodle. (ii) Put all the pages sequentially in a pdf file and upload the pdf to KHARAGPUR Moodle. YOU HAVE TO USE PEN AND PAPER TO GIVE THE EXAM.

- Name your zip file as <your roll no>.zip

Policies: Note that, if we face problems with your answer script e.g., cannot open your submitted zipped file, cannot read the text in pictures (due to bad resolution), cannot determine the page order from the file names (or the pages in the pdf is jumbled up), or we find you copying, it will affect your marks.

Malpractice: If any group of students is found to have similar work in their answer sheets, all of them will receive the maximum penalty with no grace. We expect you to not take help from the internet, your copies, textbooks, slides or video recordings during the exam. Note that this is not an open-book exam. If found otherwise, you will be penalized.

PLEASE WRITE YOUR NAME AND ROLL NO. ON THE TOP OF THE FIRST PAGE OF YOUR ANSWER SCRIPT. WE WILL NOT EVALUATE YOUR ANSWER SCRIPT WITHOUT IT.

Question 1. [7 marks]

You founded a new search engine **Saraswati.** In Saraswati you decide to implement a new way to group together similar pages for showing the users--- by identifying Link Communities, which generate overlapping communities of nodes. This is done by calculating the dual of the page-hyperlink-graph (each edge becomes a node in the dual; and if two edges were incident on a common node then the corresponding nodes in the dual are joined by an edge). The similarity between two nodes in the dual is calculated by the similarity between the two edges in the original graph, whose formula is given below:

$$Similarity(e_{ik}, e_{jk}) = \frac{|\, n_+(i) \cap n_+(j) \,|}{|\, n_+(i) \cup n_+(j) \,|}$$

$n_+(x)$represents the set of nodes x and its neighbors in the original graph. Consider the similarity between nodes in the dual that are not connected as **zero.** Also note that similarities are computed only for those edge-pairs that have one node in common. An agglomerative clustering is then used on the dual graph, with single linkage clustering (i.e. similarity between two communities is the maximum similarity between all pairs of nodes between the two communities).

Draw the dual graph of the network given in Figure 1 and then find out 3 overlapping communities of nodes. Draw the graph after merging at each step, along with the new similarities. [If there are more than one pair of communities with equal similarity, you may join them in a single step.]

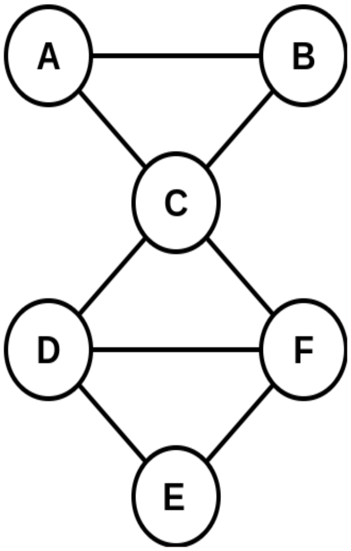

Figure 1

Question 2. [15 marks]

Please answer the following questions below (for each no marks without justification):

2.1. [1 + 3 = 4 marks]

The "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups" paper by Saha et al. mentioned that "We first form the co-occurrence network of emojis where the nodes are individual emojis and edges represent that they co-occur within a window of 5 characters at least once."

Consider emojis *i* and *j* which co-occur F_{ij} times in a window and occur individually F_i and F_j times

- what is the maximum and minimum value of an edge in this co-occurrence graph (with both F_i and $F_{j>0}$)? Why?
- Propose a new weight function expression whose minimum and maximum value will vary between [-1, 1] for each edge and capture the non-random co-occurrence of emojis. Justify your proposal.

2.2. [2 marks]

The "What happened? The Spread of Fake News Publisher Content During the 2016 U.S. Presidential Election" paper by Budak performed a cross-correlation analysis.

- Describe briefly the steps of how a cross-correlation analysis is done in two time series datasets S_t^1 and S_t^2 as used in this paper.

2.3. [1 + 4 = 5 marks]

The "Uncovering Social Network Sybils in the Wild" paper by Yang et al. ultimately proposed four features to detect Sybil or fake nodes in their deployed detector for Renren

- Name the four key features that you can use to detect Sybils or fake accounts
- Argue why or why not each of these features are robust to manipulation by Sybils in a practical deployment.

2.4. [2 + 2 = 4 marks]

The "First Women, Second Sex: Gender Bias in Wikipedia" paper by Graells-Garridoet al. used the metric pointwise mutual information (PMI) to know which words are most associated with which gender

- What is the *maximum* value of this PMI? What does it mean for gender bias when this PMI is reached? Why?
- What is the *minimum* value of this PMI? What does it mean for gender bias when this PMI is reached? Why?

Question 3. [5 + 3 = 8 marks]

Assume that you have a database which contains data about at least 10000 people and their number of views on a video advertisement on Facebook (included in the database) varied between 10 to 55 views. You want to query the database and report the average view of the people in the database using differential privacy. Consider a mechanism where the reported answer is obtained by adding Laplacian noise to the correct average;

(i) Using the definition of differential privacy, we need to find a distribution where adding noise based on that distribution in the given database would make the answers 3-differentially private. You came up with a sketch for the distribution

$$Ae^{-B|x|}$$

Now calculate A and B (which are constants). Show your calculations (no marks without showing the calculation)

(ii) For the minimum $\boldsymbol{\epsilon}$ define the error introduced in the results due to the noise addition mechanism (as mentioned in the class). Calculate the error for this mechanism.

Question 4. [2 + 2 + 2 = 6 marks]

Please answer the questions below using the privacy definitions covered in the class. (Don't use more than 3 sentences for each of the questions, otherwise you won't be given any credit)

(i) Which of the Fair Information Practice Principle (FIPP) is violated if an online shopping site put their user data (shopping history, reviews), accessible to any internet user without any password? Why?

(ii) Assume, you are marked as an Engineer in a phone directory service, but you are a doctor. You come to know about this issue and contact the phone directory service for changing your expertise to Doctor. However, they asked 1,000 inr for changing this entry. Which Fair Information Practice Principle (FIPP) is this phone directory service violating? Why?

(iii) If you realize that social bots deployed by "evilaggregator.inc" is scraping Facebook data you shared with "public" setting (e.g., your profile picture), how can theory of exposure control explain the privacy violation?