Social Media Analytics Exam, Spring 2022

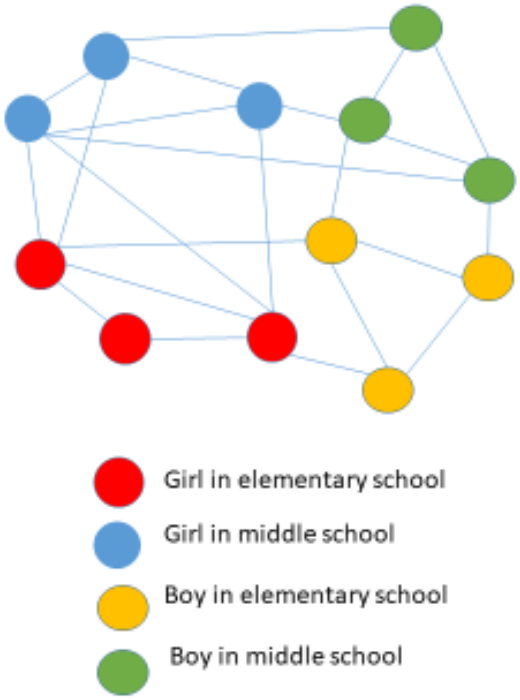
Name Rohitashwa Chakraborty

Time: **2 hours 30 mins**
Maximum points: 100

Submit your answers on Canvas by 7:30 p.m.tonight. Only two files (Word or pdf with answers, and an Excel file, if deemed necessary) should be submitted. You can write answers by hand (but they must be legible) and take pictures. Please write your name inside each document you submit. Questions run from 1 to 5.

Unlike other group tasks in this course, this exam is a strictly individual task. Do not discuss the questions and/or answers with a class- or group-mate (or anyone for that matter), for that would constitute a clear violation of the University honor code. Such cases are required by University rules to be reported to the Office of the Dean of Students.

1. Consider the friendship network below (a separate ppt file with the network diagram is attached for your convenience) between boys and girls in an elementary and a middle school.
Red: Girl in an elementary school
Blue: Girl in a middle school
Orange: Boy in an elementary school
Green: Boy in a middle school



From this data, what test(s) of homophily can be performed? Show detailed calculations. (20 points).

Answer:

Homophily refers to the tendency of similar people to stick together. "Birds of same feather, flock together".

In the above scenario, there are two attributes associated with each node/person -their gender (Boy/Girl) and their Schooling (Elementary/Middle)

From the network, we observe:

- Probability of boy = probability of girl. Thus, p_gender = q_gender = 0.5
- Probability of elementary = probability of middle. Thus, p_school = q_school = 0.5

First let us check for presence of homophily across gender

H0: Girls/Boys in elementary school are NOT closer to girls/boys in middle school

Ha: Girls/Boys in elementary school are closer to girls/boys in middle school

Let 'r0' be the fraction of cross-gender edges in a random network, generated from the probabilities given above.

Probability of a boy-girl connection is (r0) = 2*p_gender*q_gender = 0.5

Let r be the fraction of actual boy-girl connection in the network: $5/23 \approx 0.217$

Since r << r0, p-value of our hypothesis test will be very small we can reject the Null Hypothesis. Therefore the network shows Homophily when it comes to gender

Let Hypothesis: Girls/boys of elementary/middle school more closer to each other than other group. We can statistically test for homophily according to gender as follows: Let μ be the fraction of cross-gender edges in the random network R. Homophily Test: H0 : $\mu \geq \mu 0$ against H1 : $\mu < \mu 0$. If the fraction of cross-gender edges is significantly less than $2p(1 - p)$, then there is evidence for homophily.

Similarly, Testing for homophily by schooling.

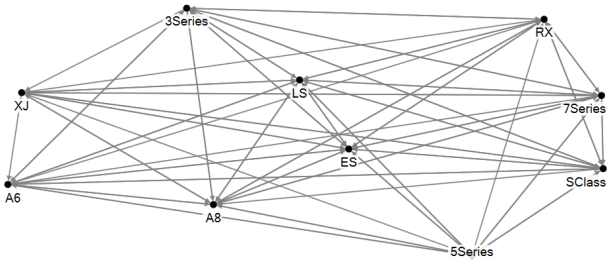
Expected probability of a middle-school – elementary school connection:

- 2p_school*q_school = 0.5

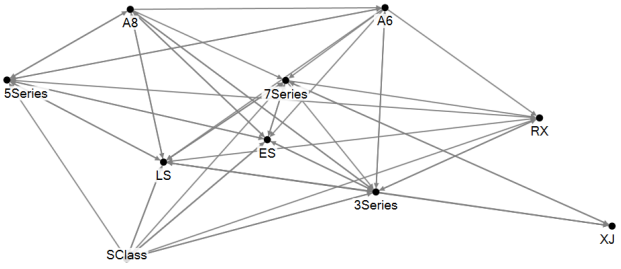
Actual observed probability of a cross school connection: $6/23 \approx 0.26$

Since Actual probability is far lesser than expected probability, homophily exists in Schooling also.

2a. Consider two product preference networks (A and B) shown below involving 10 products.



Preference network A



Preference network B

Choose the correct answer below:

[ ] PageRank scores from network A will have higher correlation with sales than those from network B.

[ ] PageRank scores from network B will have higher correlation with sales than those from network A.

[ ] It is not possible to tell from the diagrams whether PageRank scores from network A or B will have higher correlation with sales.

Provide justification for your choice. (10 points)

Answer:

PageRank scores from network B will have higher correlation with sales than those from network A.

In order to gain information from a network, we need to assume the criteria for a link between two nodes.

Let us assume, in both networks, a link is made if two products are compared in a consumer forum. Therefore, a node with a high degree means a lot of products compare themselves against this product.

Network A, is densely connected. Therefore, the transition matrix will have very similar values. When we find the limit of the transition matrix transformations, i.e: calculate the page rank, we will find very similar values for every product. (Extrapolation: every node in a fully connected network will be equally important)

However, Network B is more sparse, thus the importance values will be skewed/concentrated in a select few products. The interpretation of this importance score is that consumers are talking about that product. Thus, a high page rank on Network B implies that people are comparing other products with that product, i.e: the latter is the gold standard/desirable product of sorts.

If people talk about it, they are likely to buy it, therefore having higher correlation with sales.

Therefore, we can conclude, PageRank scores from network B will have higher correlation with sales than those from network A.

2b "It is possible to guess reasonably accurately whether a post on social media such as Twitter is coming from a spammer from the **individual's** (and not network or group level) social network centrality metrics." Do you agree with this statement? Justify your response. (10 points)

Yes/No: YES

Metric(s) used: Degree Centrality

Justification:

Spammers usually operate in closely connected groups, creating an echo chamber to reach as wide an audience as possible. Thus, one would most likely expect to see a very high out-degree, i.e: connecting to a lot of people (to maximise probability of clickbait).

Similarly, you need a very high in-degree because spammers are usually connected to each other, retweeting and endorsing each other and creating an echo-chamber

2c. Over time you kept on adding people to follow on Twitter without a careful assessment. One day you realized you were following over 1000 people, and that it's time to trim the list to be able to receive what really matters to you. You are looking for interesting and new information through Twitter that you may not easily get elsewhere. State two centrality metrics you would use to shorten your list. Explain. (10 points)

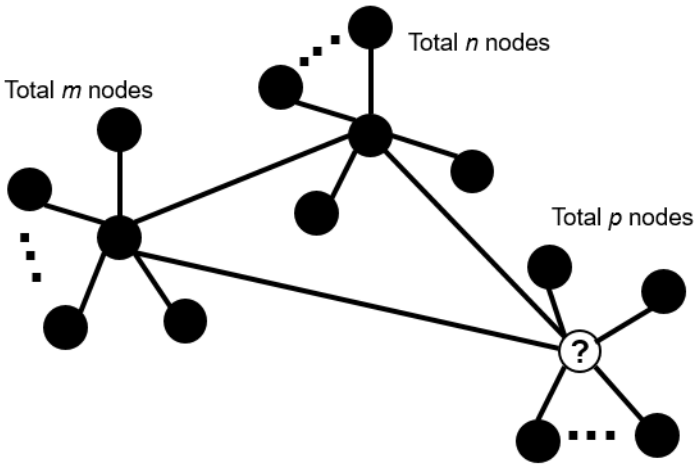
Metrics:

- High Betweenness and Low Degree

Justification:

- Betweenness implies a node is important to the flow of information across members in different communities (betweenness of a node increases if it lies on the shortest path between two other nodes)
- Therefore, We should retain connections with a high betweenness as it will add new and varied content to our feed
- A low degree value of closeness implies a node is sparsely connected. In other words, the data this node generates, will not be easily available.
- Thus, taking a combination of both, will give us an exposure to new, niche and grassroot level information that will not be easily found elsewhere.

3. In the diagram below, three connected sub-networks have a total of *m*, *n* and *p* nodes (including the central nodes). Calculate the normalized betweenness centrality of the central node of the subnetwork with *p* total nodes (the white node marked "?" in the diagram). Normalize with respect to the largest possible betweenness score of any network configuration. Show all calculations. (20 points)



Answer:

Betweenness of a node:

$$b_i = \sum_{s,t\ (s \neq i \neq t)} \frac{g_{st}(i)}{g_{st}}$$

Where g_{st} is the total number of shortest path from node s to node t
And $g_{st}(i)$ is the total number of shortest path from node s to node t, passing through i

Therefore, the betweenness of central node: $({}^{m+n}C_1 * {}^{p-1}C_1)+({}^{p-1}C_2)$
Normalised betweenness of central node: $({}^{m+n}C_1 * {}^{p-1}C_1)+({}^{p-1}C_2) / ({}^{m+n+p-1}C_2)$

(dividing betweenness of the node with max betweenness of a star node of size m+n+p nodes)

4. Two individuals in my personal Facebook network are Professor Deepa Mani of the Indian School of Business (who was my Ph.D. student at UT Austin, and who graduated about 7 years ago) and Professor Prabhudev Konana (Prof at UT Austin for 25 years, now Dean at U. Maryland). Their network centrality measures (raw scores) **in my network** are shown below:

	Prof. Mani	Prof. Konana
Degree	461	213
Betweenness	572	5478

How do you explain the differences in the above centrality metrics between Professors Mani and Konana? Specifically, how do you explain a lower degree but much higher betweenness for Prof. Konana (and the opposite for Prof. Mani)? You may find it useful that occasionally I visit the Indian School of Business as a guest professor, and teach a short course or two over there. Note that you don't need to know anything about these professors beyond the facts I have provided above. (10 points)

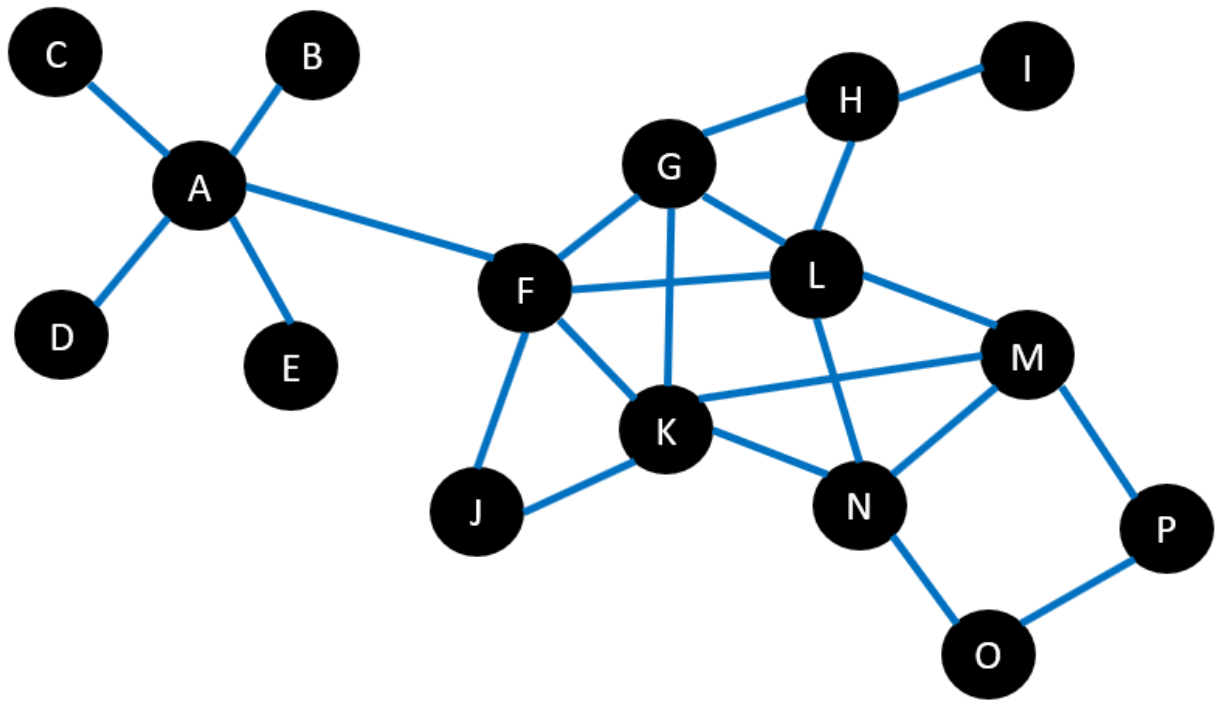
Prof. Mani

- Prof Barua and Prof Mani share a common pool of connections from their time at ISB. Thus, even though Prof Mani is extremely well connected with her peers, chances are, these connections are directly connected to Prof Barua too, thereby pushing down Prof Mani's betweenness score

Prof Konana

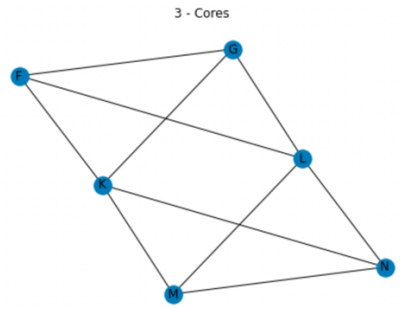
- The only common connection between Prof Konana and Prof Barua is that they spent time together in UT and then he shifted to Maryland.
- A high degree of Prof Konana implies he too is well connected with his peers, who are obviously not connected with Prof Barua, therefore driving Prof Konana's betweenness score high.

5a. From the network below, find the largest 3-core subnetwork. For your convenience I have included the PowerPoint file with the network diagram which you can edit instead of having to create the subnetwork from scratch.

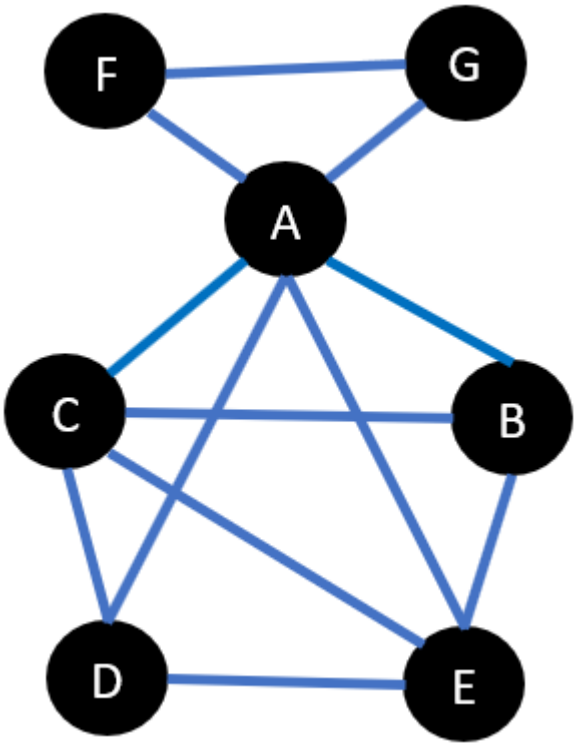


Answer:

Largest 3-core sub-network: ['F', 'G', 'L', 'K', 'M', 'N']



5b. In the network below, find the largest clique. What is the value of the clique? As in 5a, I have included the PowerPoint file with the network diagram which you can edit instead of having to create the subnetwork from scratch. (10 points)



Answer :

Largest – Clique = 4 ([‘A’, ‘C’, ‘D’, ‘E’] or [‘A’, ‘C’, ‘B’, ‘E’])