IS4241 SOCIAL MEDIA NETWORK ANALYSIS

Academic Year 2019/2020 Semester 2
Project "This Network Surprises You!"

Objective:

This project aims to showcase students' understanding and application of what has been taught in IS4241 Social Media Network Analysis. The emphasis is on network data analysis as well as data presentation, as opposed to "crawling" of data. Hence, three databases will be provided for students to choose from.

The project is titled, "**This Network Surprises You!**" For your presentation, you may change the title to better capture the essence of the network data analysis and hence, the story you wish to tell.

Details:

01. Team size – Between 5 – 6 students

02. Presentation of project – Date : 8 April 2020 (Wednesday)
Time : 8 am – 11 am
Venue : COM1 VCRM

03. Deliverables – Executive summary (6 pages,
A4 paper, normal margin,
Times New Roman Font 12)
Powerpoint (10 – 40 slides,
Note: not all slides need to be presented
e.g., extra slides for reference)

04. Total marks – 45 %

Presentation	: 5%
Executive summary	: 5%
Social network analysis	: 15%
Interesting insights	: 15%
What If	: 5%

Q: What is meant by "social network analysis"?

A: In the lectures/tutorial/lab, you will be taught numerous network concepts and metrics (e.g., network size, network diameter, network density, degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, clique, clan, club, cluster, cohesion, etc.). You may choose to perform some of these analyses (based on the story you intend to tell). You do NOT have to perform all the analyses of the various network concepts and metrics.

(Note: Feel free to ask me any question during consultation, after class, or via email)

Q: What is meant by “interesting insights”?

A: This is the chance for you to pique your curiosity and demonstrate your creativity! In the lectures/tutorial/lab, you have been urged to often think critically “Why …”, “What else …”, etc. You have also been provided with many real life examples and applications. The following are some things you can challenge yourself:

(Note: you are NOT limited to these, the following are just my humble suggestions; feel free to conceive your own wild suggestions! You will gather more during consultation too.):

e.g., What happen if some essential nodes have been deleted?! Reason?! Outcomes?!
e.g., What happen if some essential edges have been deleted?! Reason?! Outcomes?!
e.g., What happen if there are new nodes and/or edges being added?!

e.g., Are there any anomalies?! Outliers?! etc.
e.g., Are there any serendipitous findings?!
e.g., Are there seemingly conflicting findings (e.g., paradoxes)?!

e.g., Do you want to split the data along some dimensions and compare/contrast the groups?!
e.g., Do you want to compare two or more partial networks?!
e.g., Do you want to test some predictions or challenge established theories/findings?!

At first glance, this may look challenging; but do NOT feel stress.
This is really the section that allows your entire team TO HAVE FUN and really EXPLORE the data!

(Note: Likewise, feel free to ask me any question during consultation, after class, or via email)

Q: What is meant by “what if”?

A: In real life, researchers and data analysts are seldom able to secure ALL the data (or variables, in terms of “columns”) that they want. Hence, feel free to brainstorm one, two or at most three variables that you think if made available, can help elucidate more surprising/interesting/useful findings.

Do enlighten us (either through explanation or simulation/mock data) why these one, two, or three variables (if available) will make your story more compelling!

(Note: Likewise, feel free to ask me any question during consultation, after class, or via email)

Q: May we have some initial details about the presentation?

A: I will tell you more nearer the date, but here are some things you may be interested:

– Duration	: about 20 min per team
– Does everyone have to speak?	: NO
– Will there be a Q&A?	: NO
– Is there a dress code?	: please look professional

Q: May we know what to do if a team mate loafs and does NOT contribute?

A: Your team may email me.
Be rest assured of confidentiality.
Also, there will definitely be a fair investigation and fair grade.

Databases:

Students are free to choose from 1 out of 3 provided databases, namely:
(i) Database 1 – Facebook Network (see p.4)
(ii) Database 2 – Wikipedia Network (see p.5)
(iii) Database 3 – Amazon Network (see p.6)

Noteworthy:
(i) The three databases comprise only a partial network.
(ii) They have all been "anonymized" (in view of privacy and confidentiality issues).
(iii) Missing data has also been "cleaned" (i.e., in an academic way).

DATABASE 1 – FACEBOOK NETWORK

Facebook connects individuals from different locations, ages and gender to build and maintain a social network. The first database 1A comprises information about individuals (i.e., nodes) identified by their unique User ID. The second database 1B comprises information linking (i.e., edges) individuals (i.e., nodes) to individuals (i.e., nodes). Specifically, the edges denote two individuals being friends on Facebook, with weights being the number of messages exchanged.

Table 1A. Facebook Database - Node

Column	Description
User ID	It denotes the unique key for each individual in Facebook network.
Gender	It denotes the gender of the individual: 0 = Female 1 = Male
Age	It denotes the age range of the individual: 1 = 13-20 4 = 50-65 2 = 21-35 5 = >65 3 = 36-50
Horoscope	It denotes the Horoscope of the individual: 1 = Capricorn 7 = Cancer 2 = Aquarius 8 = Leo 3 = Pisces 9 = Virgo 4 = Aries 10 = Libra 5 = Taurus 11 = Scorpio 6 = Gemini 12 = Sagittarius
Language (Native)	It denotes the native language that the individual uses on Facebook: 1 = English 3 = Hindi 2 = Spanish 4 = Others
Locale (Current)	It denotes the current location the individual has is staying at: 1 = USA 3 = Asia 2 = Europe 4 = Others
Education (Attained)	It denotes the highest level of education the individual has specified: 1 = High School 3 = Postgraduate 2 = Undergraduate 4 = Others

Table 1B. Facebook Database - Edge

Column	Description
Source*	It denotes the User ID.
Target*	It denotes the User ID.
Conversation Messages	It denotes the number of messages exchanged between the two users.

Note: Even though there is a column "source" and column "target" (i.e., to facilitate the input into Gephi), you may assume that the edges in this database is "undirected". Reason: When two individuals on Facebook are friends with each other, the number of messages exchanged can be considered shared (hence same number) between the two.

DATABASE 2 – WIKIPEDIA NETWORK

Wikipedia has nowadays become an important source of information and learning on the internet. The first database 2A comprises information on Wikipedia pages (i.e., nodes) identified by their unique Page ID. The second database 2B comprises links (i.e., edges) from one Wikipedia page (i.e., nodes) to another Wikipedia page (i.e., nodes). Specifically, the edges denote hyperlinks from one page to another page, with weights being the number of clicks or access.

Table 2A. Wikipedia Database - Node	
Column	Description
Page ID	It denotes the unique Page ID for each Wikipedia page in the Wikipedia network.
Categories	The Wikipedia pages are classified into six categories: 1 = Geography 4 = Nature 2 = History 5 = People 3 = Health 6 = Others
Year of First Edit	It denotes the year that the first edit was recorded on the page, which indicates the year which the page was created.
Edit Counts	It denotes the number of times a page has been edited throughout its lifetime.
Page Views	It denotes the number of views that a Wikipedia page has had over its lifetime.
Number of Words	It denotes the number of words on the Wikipedia page.
Number of Images	It denotes the number of images in the Wikipedia page.

Table 2B. Wikipedia Database - Edge	
Column	Description
Source*	It denotes the Page ID.
Target*	It denotes the Page ID.
Clicks	It denotes the number of clicks activated on the hyperlink from Source page to Target page.

* Note: You may assume that the edges in this database is “Directed”. Reason: Page A that hyperlinks to page B may not be guaranteed a reciprocal hyperlink back to page A. Even if there is, the number of clicks may differ.

DATABASE 3 – AMAZON NETWORK

Amazon.com has been very successful in selling various products (including Books) on their website. The first database 3A comprises information about books (i.e., nodes) identified by their unique Product ID. The second database 3B comprises information linking (i.e., edges) a book (i.e., nodes) to another book (i.e., nodes). Specifically, the edges denote the "co-purchase" behaviour of consumers, and the weights reflecting co-purchases. In other words, it is often reflected in Amazon website as "Customers Who Bought This Product Also Bought ...".

Table 3A. Amazon Database - Node

Column	Description
Product ID	It denotes the unique key for each book.
Genre	Books are classified into the following genres: 1 = Non-Fiction 4 = Romance 2 = Sci-Fi & Fantasy 5 = Others 3 = Mystery and Suspense
Number of Pages	It denotes the number of pages in the book.
Price	It denotes the listed retail price of the book in US dollars.
Sales Rank	It denotes the Amazon Sales Rank of the book, which is an inverse measure of sales performance.
Average Rating	It denotes the average rating of the book out of 5. An average rating of 0 indicates no prior ratings.
Number of Reviews	It denotes the number of reviews the book has.

Table 3B. Amazon Database - Edge

Column	Description
Source*	It denotes the Product ID.
Target*	It denotes the Product ID.
Frequency	It refers to the number of co-purchases that has been made between two books.

* Note: Even though there is a column "source" and column "target" (i.e., to facilitate the input into Gephi), you may assume that the edges in this database is "undirected". Reason: For co-purchases, the products are typically bought together without indication of which product leads to the purchase of the other product.

ALL THE BEST TO YOUR PROJECT!