BIA 658 Social Network Analysis DEVILA BAKRANIA
Midterm Exam Fall 2020

There are two parts to this exam. Please note that you should refer to the Golbeck (book/videos) up to week 7 materials. If you use a reference online instead, be sure to include the reference. Wikipedia is not a valid reference.

Part I – Open Ended Questions (50 points total; 10 points per question)
Note: Remember your answers should aim to be four sentences in length at minimum. Use more space for your answers as needed. You can hand draw examples if needed. I am a big supporter of sketches.

1. What is the topic you have selected for your semester project and why? List three other topics that would be related to your semester project topic. You do NOT need to have three actual datasets. I am looking for just the topics so you can use them for your examples in questions 2-5 below.

Answer1 : In the analysis of networks we often look for the impact of nodes on the whole network. Since we are living in gloomy times, the idea of a pandemic fascinated us. Every person in this world is connected and that person to another and so on, which leads to a huge network of people meeting in social places and on platforms, our goal for the project is to simulate that experience and showcase how a virus has the potential to change how we interact. Be it a social gatherings, social distancing, wearing a masks, test results before and after cover infection and in between days where one cannot be sure, transmission range, effect as per the age and various analysis.

Second topic:
Prediction when people change jobs based on social media data for example; LinkedIn- In this project, I am going to predict when people will change their jobs based on LinkedIn profiles and jobs dataset. Also, planning to connect the same university or work people together and their common sharable learning source so that people can learn from each other and stay connected in this virtual world of the pandemic.

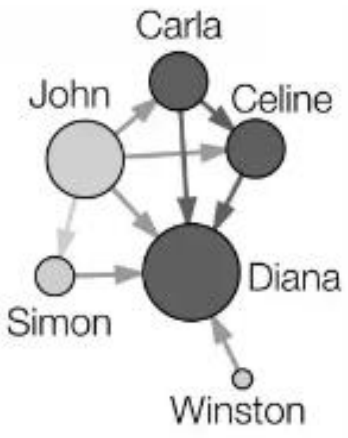
Lastly, I wanted to scrape my friend list and create a directed network, this can give me insights like which skincare, cosmetics, clothes, shoes my friends use or buy, from where they buy and comparison of price, discounts including coupon code discount data and notification when it is right time to buy on basis of offers that repeats yearly and compare the same product bought in different prices and which is cheaper.

2. Explain what makes a social network dataset different than a traditional dataset such as Excel? Provide an example (from the 3 you listed in #1) that compares a social network dataset and a traditional dataset. Be sure to clearly mark the key points.

Answer2 : A traditional dataset is about an entity with its several attributes. Traditional data sets such as in excel consists of rows and columns where rows are keys and columns are data associated within them.

While the social network data set is different as instead of using rows and columns it is composed of nodes and edges where nodes are the key points and edges are the relationship between different nodes. Unlike traditional datasets, Social network data sets can be directed or undirected.

	A	B	C	D	
1	**Last Name**	**Sales**	**Product Type**	**Company**	**Contact**
2	Smith	$1,675.00	EEE-312	Wok N Roll	Adams
3	Johnson	$1,480.00	DC-1	Wok N Roll	Rogers
4	Williams	$1,064.00	EE-2	Peace A Pizza	Evans
5	Jones	$1,390.00	DF-3	Kung Food	Webb
6	Brown	$4,865.00	EEE-45	Peace A Pizza	Fields
7	Williams	$1,243.00	FD-2	Kung Food	Mccoy
8	Johnson	$9,339.00	DC-1	Kung Food	Hansen
9	Smith	$1,891.00	EEE-312	Wok N Roll	Hamilto
10	Jones	$9,213.00	FG-5	Wok N Roll	Woods
11	Jones	$7,433.00	DF-7	Kung Food	Cunning
12	Brown	$3,255.00	FD-2	Pancakes on the Rocks	Myers
13	Williams	$1,486.00	A-34	Wok N Roll	Ford
14	Williams	$1,930.00	A-34	Pancakes on the Rocks	Edwards
15	Smith	$9,698.00	F-3334	Peace A Pizza	Murphy
16					



Nodes

Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2

Edges

Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6

Data set in the #1 proposed project consists of
"**r0**": 2.28, = basic reproduction number represents the transmission of disease
"**incubation**": 5, = Incubation period of the disease
"**percent mild**": 0.8, = percentage of mild symptoms for some infected to get
"**mild recovery**": (7, 14), = recovery from mild symptoms to recovering
"**percent severe**": 0.2, = percentage of getting severe symptoms
"**severe recovery**": (21, 42), = recovery from severe symptoms to recovering
"**severe death**": (14, 56), = chances of death from severe symptoms
"**fatality rate**": 0.034, = fatality rate of the disease
"**serial interval**": 7 = serial interval of 7 days

With a total population of 4500, analyzing the spread of disease for 365 days, we can observe how a disease transfers from an infected person to a susceptible person. Our aim is to stimulate the transmission of diseases. This analysis is impossible in traditional data as the data here governs the transmission.

Reference: https://art-bd.shinyapps.io/nCov_control/
https://www.washingtonpost.com/graphics/2020/health/coronavirus-how-epidemics-spread-and-end/
https://www.acpjournals.org/doi/10.7326/M20-0358

3. What is a matrix and how does a typical matrix differ from an adjacency matrix? Provide an example of both a typical matric and an adjacency matrix (from the 3 you listed in #1).

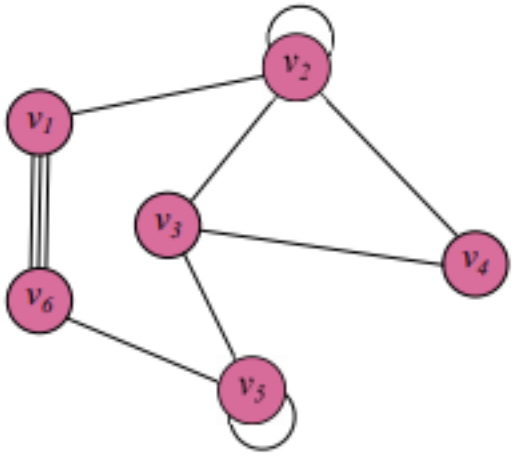
Answer3 : Matrix synonymously a table, which contains horizontal rows and vertical columns.

Example :

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

An $m \times n$ matrix: the m rows are horizontal and the n columns are vertical. Each element of a matrix is often denoted by a variable with two subscripts. For example, $a_{2,1}$ represents the element at the second row and first column of the matrix.

The adjacency matrix of a graph G, denoted by AG, is an n × n matrix with each entry aij indicating the number of edges between two vertices vi and vj. If there is no edge between the two vertices, the entry takes the value 0. Figure 1 shows an example of an undirected graph G.



$$A_G = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 3 \\ 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 1 \\ 3 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Figure 1: A graph G with its associated adjacency matrix AG

Accordingly, the associated adjacency matrices of graphs representing real-world networks are also sparse.

The **difference** between adjacency matrix and the typical matrix is that in a typical matrix the rows and columns can represent an arrangement of numbers while in an adjacency matrix the number represents edges and network between rows and columns.

For example if we create an edge list for Ag = ((v1, v2), (v2, v2), (v2, v3), (v2, v4), (v3, v4), (v3, v5), (v5, v5), (v5, v6), (v6, v1), (v6, v1), (v6, v1))

In an adjacency matrix, the rows are the same as the columns, while in a typical matrix, the rows are different from the columns.
Reference: https://static1.squarespace.com/static/559921a3e4b02c1d7480f8f4/t/59899cd33e00be69ff356841/150219081 0894/Kernschmidt+Letitia_713.pdf

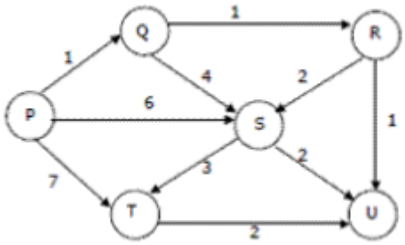
4. Define what is meant by the shortest path. Explain why finding shortest path between nodes is important for social network analysis. Provide an example (from the 3 you listed in #1) of the shortest path.

Answer4 : A path is a series of nodes that can be traversed followed by edges between them. The shortest path is the shortest distance from one node(vertices) to another.

The Shortest Path can help us to analyze the information spreading performance and research the latent relationship in the weighted social network, and so on. The length of the path represents the speed of information spread. The shortest path spreads the information fastest. In the case of a weighted graph the path with the least weight between the origin and destination represents the shortest path in the network.

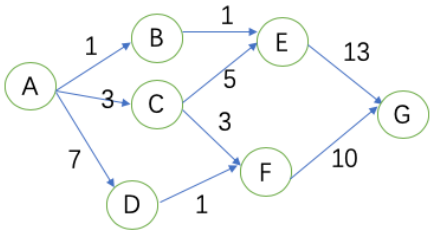
The Shortest Path is also important to find the cluster center in a social network.

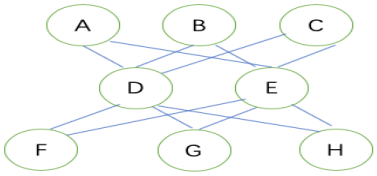
Example from listed 3,



From the above figure, we start from P and see the minimum weight of the edge then move to Q and so on till we reach U, to find the shortest path between P to U is P—>Q—>R—>U

If the path is the shortest, it means less time and cost. Below is another example of finding shortest path from A to G.





In the project from 3, Coronavirus is transmitted as per nearness of the space or we can say distance. Suppose at center of our network, the patient zero is considered. Using a Polar plot, when a node is in contact with the patient, there is a fair chance of infection in case they are susceptible, symptoms are driven by probability if they show mild or severe symptoms. In this case, people infected are the patient, and nodes near them get infected. Infected nodes have a chance of recovering or dying. This is determined by the characteristics in Question 2.

5. Find one news article from October 2020 related to the elections, COVID-19, or Hurricane Zeta. Include the link to the article you selected. Make a list of 10 data points (attributes) you think are important. Create a social network graph (nodes and edges) from 2-3 of your data points. You do not need to use any software, refer to the early chapters of the nodes/edges.

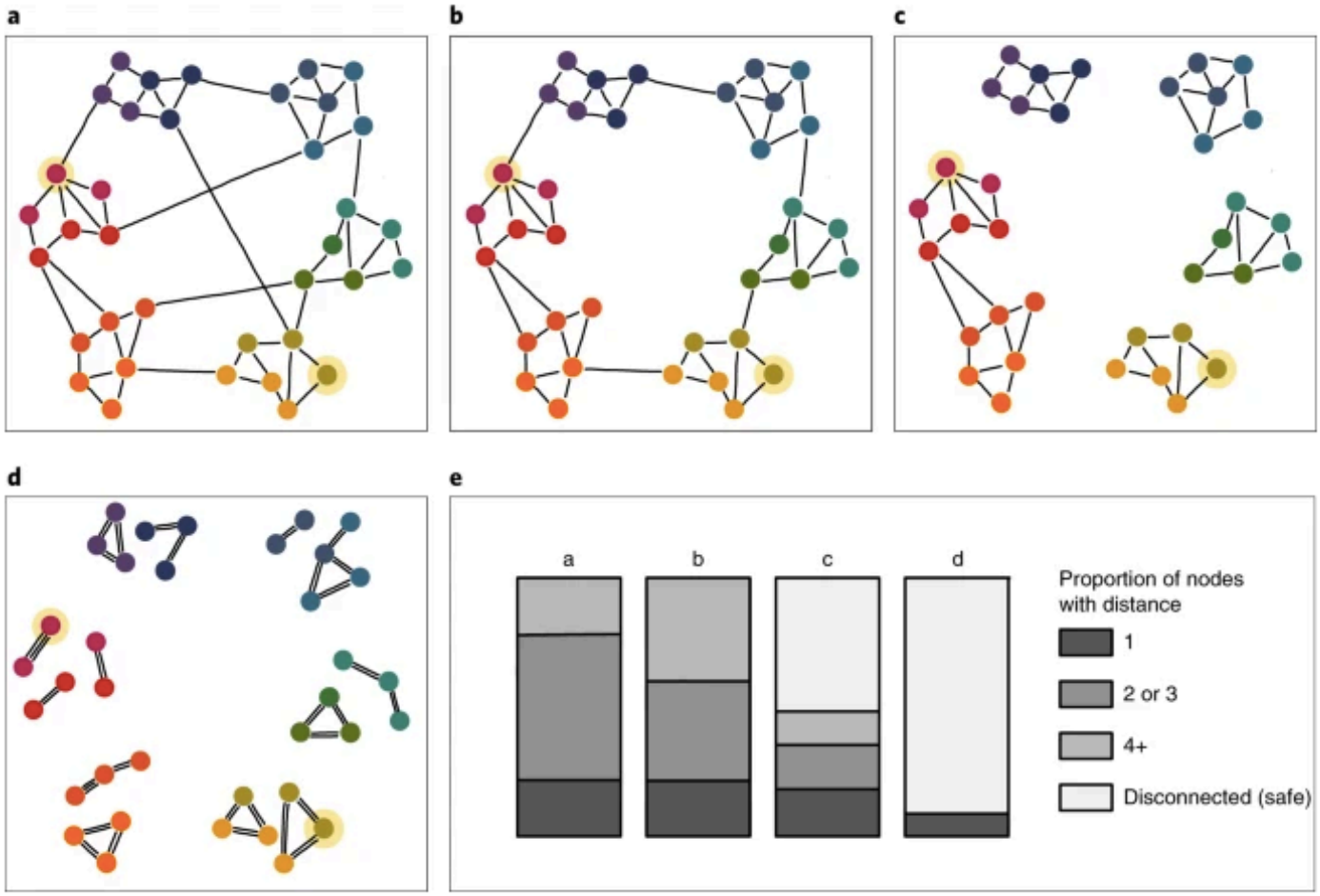
Answer5 : Reference:
https://www.scientificamerican.com/article/coronavirus-news-roundup-october-3-october-9/

Smart, useful, scientific stuff about COVID-19 and various factors responsible for it discussed with ambiguity in this article. However, we can analyze the data as per the information. The attributes from the article are as follows:

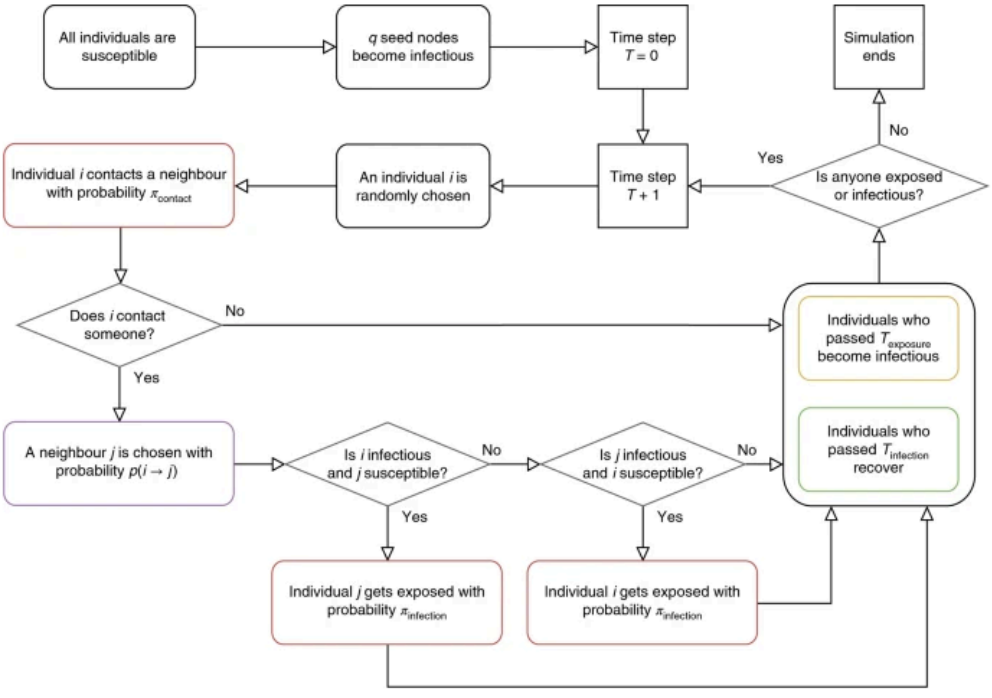
1) Saliva Tests
2) Infection rates
2) Experimental Drugs
3) Vaccine
4) Immune response as per age
5) Test Types (doesn't guarantee one is infected)
6) Transmitting Range/ Tie Reduction
7) Susceptible
8) Severity of infections
9) Dependability on Face-Mask
10) Evidence (ambiguity)

Below graph describes attribute 6: Tie Reduction Strategies

a–**d**, Based on an initial small-world network (**a**), example networks are mapped based on removing ties to dissimilar others who live far away (**b**), removing non-embedded ties that are not part of triads or four-cycles (**c**) or repeating rather than extending contact (**d**). Node colour represents an individual characteristic, where similarity in node colour represents similarity in this characteristic. Node placement represents geographic location of residence. Ties to dissimilar others who live far away are indicated by ties substantially longer than the average (that is, to nodes that are placed distantly and have

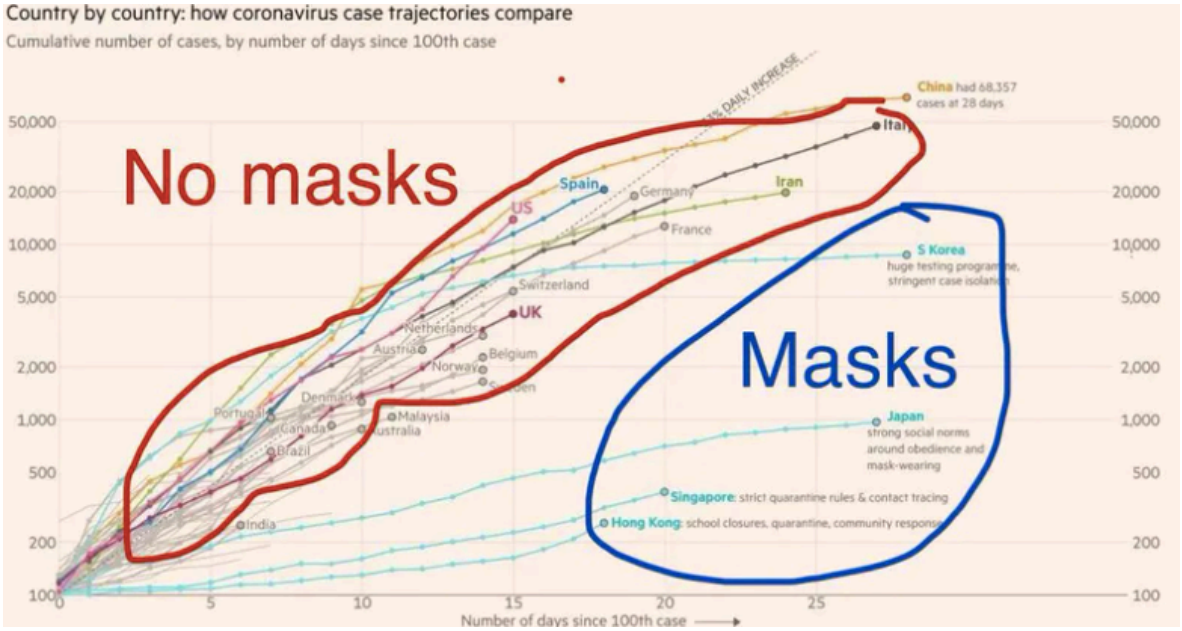


very different colours). **e**, Bar graph showing network distances from the infection sources (highlighted in yellow in **a**–**d**) for the different scenarios.



Link: https://www.nature.com/articles/s41562-020-0898-6

Below graph describes point 9 : Dependability on Face-Mask:



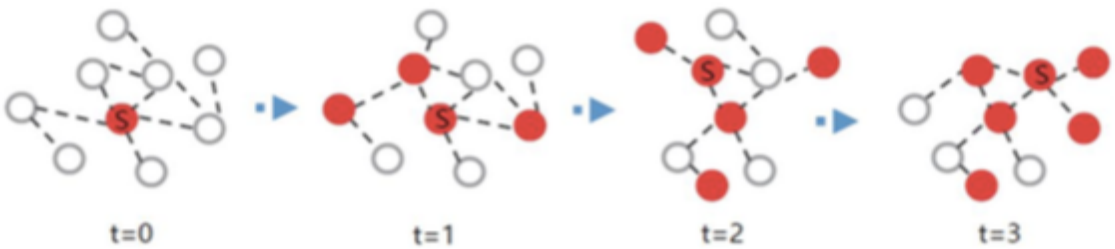
Link : https://www.dietdoctor.com/should-you-wear-a-homemade-mask-in-public

Here, Creating the Network by understanding various ways the disease is transmitted.

Adjacency matrix:

$$\text{Virus} = \begin{matrix} & P0 & P1 & P2 \\ P0 & 1 & 1 & 1 \\ P1 & 1 & 0 & 0 \\ P2 & 1 & 0 & 0 \end{matrix}$$

Here, P0 is patient 0 and is infected by virus so (P0, P0) = 1 , P1 and P2 are (attribute 7) susceptible to the disease so P0 can infect them i.e (P0, P1) = 1, (P0, P2) = 0, since P1 and P2 aren't infected yet so (P1,P1)=0 and (P1, P2) = 0.



Here, t = 0 is day 0, t = 1 is day 1 and so on S is patient 0

Above figure shows the most reduced infected graph.

Part II – Practical Application (50 points total)

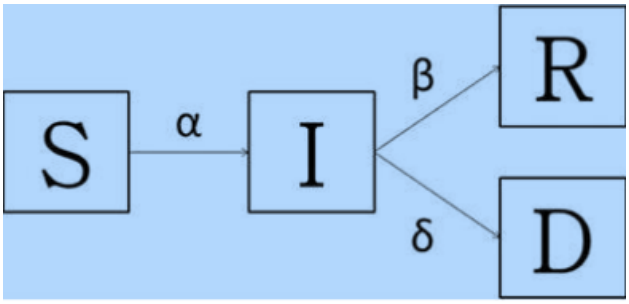
The part requires you to think critically about your dataset. If you do not have a dataset yet, please find a temporary dataset to answer the questions below. You can also draw a sketch if it helps explain.

1. Why did you select your dataset for your semester project topic (i.e. what is the problem you are hoping to solve from the dataset you selected)?

Answer1: **Problem:** Pandemics are rare in this day and age, with the advancements in health tech and vaccines readily available, Impact of epidemics can be deeply studied and countered but this is a double-edged sword, while our advancements provide us with safety this also leads to some organizations using harder and harder antibiotics to compensate for bad livestock practices lead to the evolution of superbugs, the bacteria and virus that are not affected by most powerful antibiotics. Also, with the emergence of Coronavirus, the unpreparedness of our health care system has come under attention. The problem of understanding this is very important to counter the effects. Understanding the transmission of disease allows us to set up counter measures to flatten the curve, it is also important to use these insights to have counter measures in place for further epidemics in the future.
Project Goal:
The goal of this project is to simulate how the virus spreads from a patient zero to a community, to better understand the impact, recovery and death that can occur because of the epidemic.

2. Explain your dataset in terms of basic demographics (descriptive statistics with up to 5 attributes). What type of statistical analysis do you plan to perform and what software will you use?

Answer2 :



The project closely follows the SIRD model – **Susceptible, Infectious, Recovered, Death**. The equation the model is governed by:

$$S(t) = S(0)e^{-R_0\frac{(R(t)-R(0))}{N}}$$

Where:
S(0) = Initial number of susceptible subjects
R(0) = initial numbers of removed subjects
N = Population
R0 = basic reproduction number
S(t) = number of susceptible individuals as a function of time R(t) = number of removed individuals as a function of time

Transition rate: $\frac{d(\frac{S}{N})}{dt} = -\beta SI/N^2$

Where:
β is the average number of contacts per person per time,
SI/N^2 is the probability of disease transmission in contact between a susceptible and infectious subject.

We cannot use data on cases and tests performed in different cities, states or countries as the stimulation provides the impact and recovery time of the coronavirus. In the statistical analysis, we use a predictive mathematical model to understand the transmission of pandemic and prepare for control strategies.

Software used in the project are Visual Studio Code 2019, Gephi to tell the cluster's center of the infected, Excel, RStudio to plot various graphs, Python 3.7, Libraries - NumPy, Matplotlib.

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7108749/

3. Using your dataset as a starting point, provide an example of unsupervised learning and a second example of supervised learning. You can either begin with unsupervised or supervised learning and then add more data to your example if needed.

Answer3 : This project is an example of Unsupervised learning, as simulating a coronavirus - it's spread, there is no definite answer, however, it is governed by various factors which make it closer to predictions. Another example of unsupervised learning is the categorisation of the people when travel is enabled form one place to another and that's how the infection spreads, or when all the nodes have to travel to a center point location such as a common place like buying day-to-day groceries from a store, going to gym, or medical store, or at traffic place, how this can impact the people and the solution, can be undertaken, to create solutions to counteract or limit this activity.

Supervised learning requires providing real world data to train the simulation to closely resemble real world, using data from traveling within cities, states and countries can allow us to create cases of potential spreads and prepare for risk associated with such activities, however, this comes at a cost where the data needs to be reliable and collected close to accurately, mishandling of data can lead to ineffective analysis.

We can also predict the future data, so one can know the safety measures and precautions to be taken before spreading the infection.