**CSE3024- Web Mining**

***Digital Assignment - I***

**Application of Social Network Analysis to Investigate COVID-19 Transmission Among Contacts**

*By*

20BCE1182 Sheral Simon Waskar

20BCE1320 Anjali Jain

B.Tech CSE

*Submitted to*

**Dr.A.Bhuvaneswari,**

Assistant Professor Senior,

SCOPE, VIT, Chennai

**School of Computer Science and Engineering**



*February 2023*

*­­­*

**School of Computing Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

WINTER SEM 22-23

**Worklet details**

|  |  |  |
| --- | --- | --- |
| Programme | B.Tech CSE | |
| Course Name / Code | Web Mining - CSE3024 | |
| Slot | A2 | |
| Faculty Name | Dr.A.Bhuvaneswari | |
| J Component Title | Application of Social Network Analysis to Investigate COVID-19 Transmission Among Contacts | |
| Team Members Name | Reg. No | Sheral Simon Waskar | 20BCE1182 |
| Anjali Jain | 20BCE1320 |

**Team Members(s) Contributions – Tentatively planned for implementation:**

|  |  |
| --- | --- |
| *Worklet Tasks* | *Contributor’s Names* |
| Topic selection & Dataset Gathering | Sheral |
| Preprocessing | Sheral |
| Model building | Anjali |
| Visualization | Anjali |
| Technical Report writing | Sheral and Anjali |
| Presentation preparation | Sheral and Anjali |

**ABSTRACT**

Contact tracing data is used from the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic to estimate basic epidemiological parameters. Contact tracing data may also be used to assess heterogeneity of infection at the individual patient level. Characterizing individuals based on different levels of infectivity can better inform field-level contact tracing measures.

It could help identify, test, and confine close contacts to prevent further spread. The effectiveness of this approach has been amply established in the monitoring and management of COVID-19 (coronavirus disease 2019) epidemics. The goal of this study is to use contact tracing data to examine the level of dissemination and the emergence of transmission cascades made up of numerous clusters.

This study aims to leverage contact tracing data to investigate the degree of spread and the formation of transmission cascades composed of multiple clusters.This study examines why it is important to study all combined interventions against the infection to analyse the infection outbreak size. In addition to this, it also aims to minimize the spread of virus in the given network by analysing the communities and their impact on others.

We have constructed a graph in which individuals were considered as nodes and links as the infection relationships between them. It was a directed graph in which the links will be directed from the covid positive patient who came in contact with a covid negative patient which resulted in the latter being infected with covid.

The node with the highest degree centrality will be considered as the node responsible for the spread of the virus on a large scale. Betweeness centrality measure will be used to find which node was responsible for passing on the virus from one network to another.

We will be clustering the nodes based on their level in the infection tree and analysing the impact of each level on the spread of covid.

***Keywords -*** *Degree Centrality. Closeness Centrality, Betweenness Centrality, EigenVector Centrality,*

**INTRODUCTION**

Social Network Analysis (SNA) is a method used to visualize and understand the relationships and interactions among individuals in a network. With the outbreak of the COVID-19 pandemic, SNA has become an important tool to help track and control the spread of the virus. In the context of investigating COVID-19 transmission, SNA can be used to identify and analyze the connections between individuals who have been in close contact with one another and have tested positive for the virus.

This information can then be used to inform public health interventions, such as contact tracing and quarantine measures, to reduce the transmission of the virus. SNA can also help to identify potential super-spreader events, where a single individual is responsible for transmitting the virus to multiple others. This information can be used to target public health resources to the individuals and communities most in need, and to reduce the overall impact of the pandemic.

Overall, the application of SNA to investigate COVID-19 transmission is an important tool in the fight against the pandemic, and can provide valuable insights into the spread of the virus and inform effective public health strategies.

There are several challenges faced in the application of Social Network Analysis (SNA) to investigate COVID-19 transmission among contacts. Some of these challenges include:

Data Collection: One of the biggest challenges in SNA is obtaining accurate and comprehensive data on the relationships and interactions among individuals. This is particularly difficult in the context of COVID-19, as many individuals may not be willing to disclose information about their contacts or may not have a complete understanding of who they have been in close contact with.

Data Quality: The quality of the data used in SNA is critical to the accuracy of the results. In the case of COVID-19, data on close contacts may be incomplete or unreliable, as individuals may not remember all of their contacts or may not accurately report them.

Privacy Concerns: Privacy concerns are a major challenge in the use of SNA for investigating COVID-19 transmission. Individuals may be hesitant to disclose information about their close contacts due to concerns about privacy and confidentiality.

Bias in Data: SNA results can be influenced by biases in the data collected. For example, if individuals are more likely to report close contacts with certain individuals, such as family members or friends, this can skew the results and lead to a biased understanding of the spread of the virus.

Technical Challenges: SNA can be a complex and technical process, and requires specialized software and expertise to implement effectively. This can pose a challenge for public health officials and others who may not have the necessary technical skills or resources to implement SNA effectively.

Despite these challenges, SNA remains an important tool in the fight against COVID-19, and efforts are being made to overcome these challenges and ensure that SNA is used effectively to track and control the spread of the virus.

**LITERATURE SURVEY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Title** | **Author / Journal name / Year** | **Technique** | **Result** |
| 1 | Social network analysis methods for exploring SARS-CoV-2 contact tracing data | Nagarajan, K., Muniyandi, M., Palani, B. et al.  BMC Med Res Methodol 20, 233 (2020) | degree centrality and betweenness centrality | Network component analysis identified nineteen connected components comprising of influential patient’s which have overall accounted for 26.95% of total patients (1959) and 68.74% of epidemiological contacts in the network. |
| 2 | Using high-resolution contact networks to evaluate SARS-CoV-2 transmission and control in large-scale multi-day events | Rachael Pung et al.r    Nature Communications, 2022 | Used contact tracking device data to analyze the risk of SARS-COV-2 outbreaks, which can provide a clear-cut evaluation of the interactions such as contact distance and duration of time of contacts | The 95th percentile of the epidemic size is around three times greater than the projected outbreak size, resulting in the largest number of infections from all other permutations of interventions |
| 3 | A social network analysis of the spread of COVID-19 in South Korea and policy implications. | Jo, W., Chang, D., You, M. et al.  Sci Rep 11, 8581 (2021) | network analysis, hypothesis tests on the distributions of network indicators, and virtual structural changes in the network. | This study utilized actual data to provide limited but meaningful results. |
| 4 | The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data | [Chayanon Phucharoen](https://pubmed.ncbi.nlm.nih.gov/?term=Phucharoen+C&cauthor_id=32984782)  Clinical Medicine,2020 | The paper aims to explore the various factors that contribute to the transmission of COVID-19 or SARS-CoV-2 in Phuket. | Using the Probit model, the paper analyzed the high-risk contacts of the public health office in the region who were involved in the confirmed cases of COVID-19. It also looked into the impact of quarantine measures on the probability of infection among individuals. |
| 5 | Mining relationships between transmission clusters from contact tracing data: An application for investigating COVID-19 outbreak | Tsz Ho Kwan, Ngai Sze Wong, Eng-Kiong Yeoh, Shui Shan Lee  Journal of the American Medical Informatics Association, (2021) | An algorithm on mining relationships between clusters for network analysis is proposed with 3 steps: horizontal edge creation, vertical edge consolidation, and graph reduction. | The proposed algorithm could contribute to in-depth epidemiologic investigation of infectious disease transmission to support targeted non pharmaceutical intervention policies for COVID-19 epidemic control. |
| 6 | Social network analysis of COVID-19 transmission in Karnataka, India | [S. Saraswathi](https://pubmed.ncbi.nlm.nih.gov/?term=Saraswathi%20S%5BAuthor%5D),  National Library of Medicine,2020 | Analyzed which area was most affected by COVID-19 by using the visualization tools like Cytoscape and Gephi, through a network and found out all the centrality measures for the data collected in India. | Bangalore had the highest number of cases recorded where more healthcare facilities were provided and priority was given to this region |
| 7 | Social network analysis of tourism data: A case study of quarantine decisions in COVID-19 pandemic | Fatma Altuntas, Serkan Altuntas, Turkay Dereli,  International Journal of Information Management Data Insights  (2022) | [social network analysis](https://www.sciencedirect.com/topics/social-sciences/social-network-analysis) (SNA) based on tourism data to make the right quarantine decisions in the COVID-19 pandemic. | Quarantine authorities in each country can utilize SNA to make the right quarantine decisions to reduce the impact of the pandemic on tourism. |
| 8 | The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology | [Kyra H. Grantz](https://www.nature.com/articles/s41467-020-18190-5#auth-Kyra_H_-Grantz),  Nature Communications (2020) | Analyzed the data collected using mobile phones to determine the spread of the COVID-19 pandemic. Non-pharmaceutical intervention (NPI) like social distancing was used as the primary control strategy during COVID-19. Analyzing how effectively the NPI is working to control the transmission of COVID-19 between people is a challenging task. | Call Detail Records(CDR) collected by the Mobile Phone operators which contain the timestamp and Global Positioning System(GPS) location of all subscribers is used to generate Origin-Destination matrices. This matrix helps to detect regular hotspots of attraction. The COVID-19 spread patterns analyzed using this matrix are used to inform future projections of disease risk and help in implementing travel restrictions. These patterns became the key features to detect the disease hotspots. This data can be further analyzed and put into decision-making for the restrictions of the areas which are detected as disease hotspots. |
| 9 | Constructing personal networks in light of COVID-19 containment measures | [Emanuela Furfaro](https://genus.springeropen.com/articles/10.1186/s41118-021-00128-4#auth-Emanuela-Furfaro),Springer , 2021 | The Family and Social Subjects (FSS) survey is analyzed. Due to epidemiological significance, contact tracing data helps to control the spread of disease in intervention settings | Includes rebuilding the ego-centered networks based on family relations, face-to-face contacts, and residential proximity of siblings, children, grandchildren, etc., to study social relationships |
| 10 | Network-based prediction of COVID-19 epidemic spreading in Italy | [Clara Pizzuti](https://pubmed.ncbi.nlm.nih.gov/?term=Pizzuti%20C%5BAuthor%5D),  National Library of Medicine ,2020 | If the person is infected with COVID-19, uploads the broadcasted chips which will be reflected in the diagnosis server, and whoever got in contact with the infected person will be easily identified using digital exposure notification by checking log data., | The classic SIR model is extended to analyze the infection rates based on the changing protocols and measures followed by the government. |

**DATASET AND TOOL USED**

The dataset that we are analyzing is openly available in its csv format on Kaggle.

Our dataset consists of 4 columns named reporting user (ID), contact user (ID), contact start time and contact end time.

For our analysis, we have used the first 2 columns of our dataset for forming the network.

We used google collab to perform the metrics and for visualisation also we will try the same visualisations using Gephi tool.

**ALGORITHMS/TECHNIQUES DESCRIPTION**

**Construction of graph**

Using the data from contacts.csv we constructed the graph using networkx package in python

1. import networkx as nx

2. G = nx.DiGraph()

3. with open('contacts.csv', 'r') as f:

4. data = csv.reader(f)

5. i=1

6. headers = next(data)

7. for row in tqdm(data):

8. i=i+1

9. if i==201:

10. break

11. G.add\_node(row[0]) #superhero in first column

12. G.add\_node(row[1]) #superhero in second column

13. if G.has\_edge(row[0], row[1]):

14. # edge already exists, increase weight by one

15. G[row[0]][row[1]]['weight'] += 1

16. else:

17. # add new edge with weight 1

18. G.add\_edge(row[0], row[1], weight = 1)

19. nx.draw(G, with\_labels=True)

20. plt.show()

Contact tracing data of different people is being collected in a specific area along with other details such as their gender, age, whether they had come in contact with any COVID-19 positive patient before they got infected, whether they have any comorbidities or not, whom all did they meet while they were COVID-19 positive and whether they had come in contact with any COVID-19negative person after getting tested as COVID-19 positive. SARS-COV-2 outbreak size will be analyzed based on all the possible interventions and factors which impact this outbreak size the most, to reduce the number of cases. The contact tracing data set is converted into a network with nodes and edges representing persons and their interactions among them. Various SNA metrices are analyzed to identify the spread among them. The node with the highest degree centrality will be considered as the node responsible for the spread of the virus on a large scale. Betweenness centrality measure will be used to find which node was responsible for passing on the virus from one network to another. Clustering the nodes will be done based on their level in the infection tree and analyzing the impact of each level on the spread of COVID-19. Tracing the patterns will be done in which the virus was spread in the network, in order to find a solution for curbing further spread of the virus.

**GITHUB REPOSITORY LINK**

<https://github.com/Sheral18/Web-Mining-CSE-3024---Digital-Assignment-1>

**REFERENCES**

[1] Nagarajan, K., Muniyandi, M., Palani, B. et al. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. BMC Med Res Methodol 20, 233 (2020).<https://doi.org/10.1186/s12874-020-01119-3>

[2] Pung, R., Firth, J.A., Spurgin, L.G. et al. Using high-resolution contact networks to evaluate SARS-CoV-2 transmission and control in large-scale multi-day events. Nat Commun 13, 1956 (2022). https://doi.org/10.1038/s41467-022-29522-y

[3] Jo, W., Chang, D., You, M. et al. A social network analysis of the spread of COVID-19 in South Korea and policy implications. Sci Rep 11, 8581 (2021).<https://doi.org/10.1038/s41598-021-87837-0>

[4] Chayanon Phucharoen, Nichapat Sangkaew, Kristina Stosic, The characteristics of COVID-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data, EClinicalMedicine, Volume 27, 2020, 100543, ISSN 2589-5370,<https://doi.org/10.1016/j.eclinm.2020.100543>.

[5] Tsz Ho Kwan, Ngai Sze Wong, Eng-Kiong Yeoh, Shui Shan Lee, Mining relationships between transmission clusters from contact tracing data: An application for investigating COVID-19 outbreak, Journal of the American Medical Informatics Association, Volume 28, Issue 11, November 2021, Pages 2385–2392,<https://doi.org/10.1093/jamia/ocab175>

[6] Sakranaik, Saraswathi & Mukhopadhyay, Amita & Shah, Het & Ranganath, T. (2020). Social Network Analysis of COVID-19 Transmission in Karnataka, India. Epidemiology and infection. 148. 1-30.  [https://doi.org/](https://doi.org/10.1038/s41467-020-18190-5)10.1017/S095026882000223X.

[7] Fatma Altuntas, Serkan Altuntas, Turkay Dereli, Social network analysis of tourism data: A case study of quarantine decisions in COVID-19 pandemic, International Journal of Information Management Data Insights,Volume 2, Issue 2, 2022, 100108, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2022.100108.

[8] Grantz, K.H., Meredith, H.R., Cummings, D.A.T. et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun 11, 4961 (2020).<https://doi.org/10.1038/s41467-020-18190-5>

[9] Furfaro, E., Rivellini, G., Pelle, E. et al. Constructing personal networks in light of COVID-19 containment measures. Genus 77, 17 (2021).<https://doi.org/10.1186/s41118-021-00128-4>

[10] Pizzuti, C., Socievole, A., Prasse, B. et al. Network-based prediction of COVID-19 epidemic spreading in Italy. Appl Netw Sci 5, 91 (2020).<https://doi.org/10.1007/s41109-020-00333-8>