HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 4 (2021)

Simulated Study of SARS-CoV-2: Contact Tracing and Potential Transmission Networks

Anjali Iyer[1] and Anupama Shankar[#]

[1]Denmark High School, Alpharetta, GA, USA
[#]Advisor

ABSTRACT

During a disease outbreak, contact tracing and epidemiological analysis are of critical importance to analyze disease sources and transmission. To perform this epidemiological analysis, effective data visualization is necessary. In this study, the outbreak of COVID-19 was simulated within three metro Atlanta counties. Data was generated using contact tracing forms and used to create node and edge lists. Each node in the network contained a unique ID representing either a location or individual, as well as any contact tracing information related to each node. Data visualization was performed using MicrobeTrace, an online program developed by the CDC. Visualization of the contact tracing network enabled us to effectively analyze the transmission dynamics of SARS-CoV2. In the simulated network, a singular person node appeared to be linked to the most positive COVID-19 cases in the network. Similarly, a restaurant was identified as the place node with the greatest number of direct connections to positive persons, highlighting it as a potentially major source of exposure to SARS-CoV-2. This study illustrates the benefits of data visualization and demographic analysis using MicrobeTrace, which helps to target mitigation and prevention efforts, while also emphasizing the importance of contact tracing to reduce the transmission of disease.

Introduction

Contact tracing is a process employed during a disease outbreak to identify any potential sources of the disease and its transmission. In the case of an outbreak, contact tracing is utilized to construct a contact tracing network that displays potential sources of outbreaks and transmission routes of the disease over a geographical area. Within a contact tracing network, various elements are used together to create a larger epidemiologic picture. Two of the primary elements are nodes and edges. Objects with attributes in a contact tracing network are referred to as nodes. For example, a node in a contact tracing network could be an individual who tested positive for the disease. Nodes have certain characteristics, or attributes, which can include gender, age, etc. Edges link two or more nodes from the disease source to the target which is used to analyze the transmission of a particular disease. A key benefit of constructing this type of network is that, upon its development, appropriate action can be taken to interrupt the spread of disease.

COVID-19 is caused by SARS-CoV-2, a novel virus belonging to the Coronavirus family, first identified in Wuhan, a city in central China. The first COVID-19 death was reported on January 7, 2020 (Alzu'bi et al., 2020). Since then, the number of COVID-19 cases rapidly exploded in the rest of the world, causing a global pandemic. As of September 1, 2021, there have been a staggering 193 million reported cases and approximately 4 million deaths. The identification of this novel virus prompted the implementation of new and efficient contact tracing measures by health organizations around the world.

In any outbreak investigation, just gathering data is not sufficient, the data must be visualized and effectively analyzed. Computer programs can be utilized to generate contact tracing networks. For this study, a publicly available data visualization tool called MicrobeTrace, developed by the Centers for Disease Control and Prevention (CDC), was used to develop the network. The program allows users to securely load data and offers various views for a visual

representation of the loaded data. The intuitive functionality of MicrobeTrace allows researchers to more easily view and interpret disease transmission trends.

COVID-19 data is very sensitive in nature due to a variety of factors, particularly due to privacy concerns. Consequently, collecting real-world COVID-19 data is challenging because many people are wary of sharing this data publicly. These challenges prompt the creation of simulated datasets, as was done for this study, instead. The dataset was simulated using MicrobeTrace, a web application that allows for visual analysis of data collected and their epidemiologic networks at the time of an outbreak (Campbell et al., 2020).

The ultimate purpose of this study was to perform a simulated data analysis of COVID-19 transmission dynamics within a limited geographical area using MicrobeTrace, a data visualization tool. The data analysis and visualization were performed in a manner resembling true contact tracing procedures. This study highlights the importance of contact tracing and shows the benefit of using the MicrobeTrace program to effectively construct contact tracing networks and aid in drawing the appropriate conclusions from the data before taking further action.

Methods

Software Training

Before beginning data generation, appropriate training on the data visualization tool MicrobeTrace was completed. Technical training included study of the MicrobeTrace user manual and information videos as well as running example data sets provided by the CDC through the program for a greater understanding of its views and functionalities (*MicrobeTrace User Manual*, 2020). There was an additional training session with one of the members of the MicrobeTrace team to demonstrate the tool as well.

Data Generation

The data for this project was simulated in a fashion so the data resembled true contact tracing information. While contact tracing is performed over larger areas, this project's main goal was to use MicrobeTrace to replicate the process of contact tracing so the geographical area included was smaller and data was collected on a much smaller scale. The contact tracing network analyzed 47 nodes and 66 edges. Node attribute data was randomly generated. For each attribute, an option from a predetermined list of options was randomly chosen. Each node was allocated between 3 and 4 links to other nodes for the purposes of this project. Source to target links were once again randomly chosen from the collection of nodes that had already been simulated. Information that would be collected came directly from contact tracing information collected by the CDC as listed on their index (*Appendices,* 2021). Data from this index was used to aid in the development of questions for the mock contact tracing form. Mock surveys were created with general questions about age, zip code, and gender, as well as contact tracing questions about most recent contacts, most recent places visited, symptoms presented, and severity of the disease (hospitalization required or not). These surveys were then used to "collect" data from different people.

Each survey received a unique ID number in the node list. For the purposes of the study, all person nodes in the node list were designated as previously testing positive for COVID-19. Node attributes included: Gender, race, age range, and residence zip code, symptom presentation, type of symptoms presented, ICU admission, number and types of pre-existing health conditions, date of first positive COVID test, date of first negative COVID test, method of COVID exposure, any healthcare settings worked in, any congregate settings worked in, types of group settings frequented, and 3-4 recent contacts. All of this information became the metadata for the study. No names were collected or used in studies, nodes were only identified by node ID. The completion of one survey was simulated for each person and with each category answered randomly. Each survey received a unique ID on the node list and information

for each category was recorded for that ID in the same list. Contact listed by each ‘person’ were also given unique IDs in the network. There were 47 total nodes in the contact tracing network.

ID	Node Type	Gender	Race	Age Range	Zip Code	Test positive for COVID?	Symptoms?
1	Person	F	Asian	18-60	30270	Yes	Yes
2	Person	M	White	18-60	30004	Yes	No
N1	Place				30061		
3	Person	M	Black/AA	18-60	30061	Yes	Yes
R1	Place				30004		
4	Person	F	Asian	18-60	30004	Yes	Yes

Table 1. Example Node List

In the Node List (Table 1), node IDs were recorded in the first column, and Node Type, which was either a person node or place node, was recorded in the second column. All columns after recorded information about the nodes which made up the metadata for the network. Nodes with IDs that corresponded to a person were referred to as person nodes. Nodes with IDs that corresponded to a place frequented by a person were place nodes.

Edges, which were the recorded links between nodes, were also simulated. Such links were determined by the contacts listed in each completed survey. Only three edge attributes were recorded about edges in the edge list: source, target, and link type. Edges were characterized by two link types: personLink and locationLink. There were 66 total nodes in the contact tracing network.

Table 2. Example Edge List

Source	Target	Link Type
1	2	PersonLink
1	3	PersonLink
1	4	PersonLink
2	N1	LocationLink
3	N1	LocationLink
3	R1	LocationLink
4	LG1	LocationLink

The edge list above (Table 2) displays source nodes which are identified in the first column. The target nodes, which were identified from the contacts provided in the mock survey, are located in the second column. A personLink describes a connection between two person nodes and a locationLink describes a connection between a person node and a place node.

Data from the surveys were saved as the node and edge lists (.csv format) before being loaded into Microbe-Trace. Three different views were used to visualize the data: 2D Network View, Map, and Gantt View.

Results and Discussion

2D Network View

In the 2D network view (Figure 1), nodes were shaped by node type and colored by the display of symptoms. The person node with ID 7 was directly connected to multiple other nodes. One cluster of interest contained place node LG9 which was directly connected to person nodes with IDs 27, 28, and 30.

Another cluster of interest contained place node R2 which was directly linked to five person nodes with the following IDs: 17, 19, 18, 8, and 6. Three more nodes were indirectly linked to R2, meaning they were directly linked to one of the five nodes previously mentioned, these nodes were 20, 16, and 5.

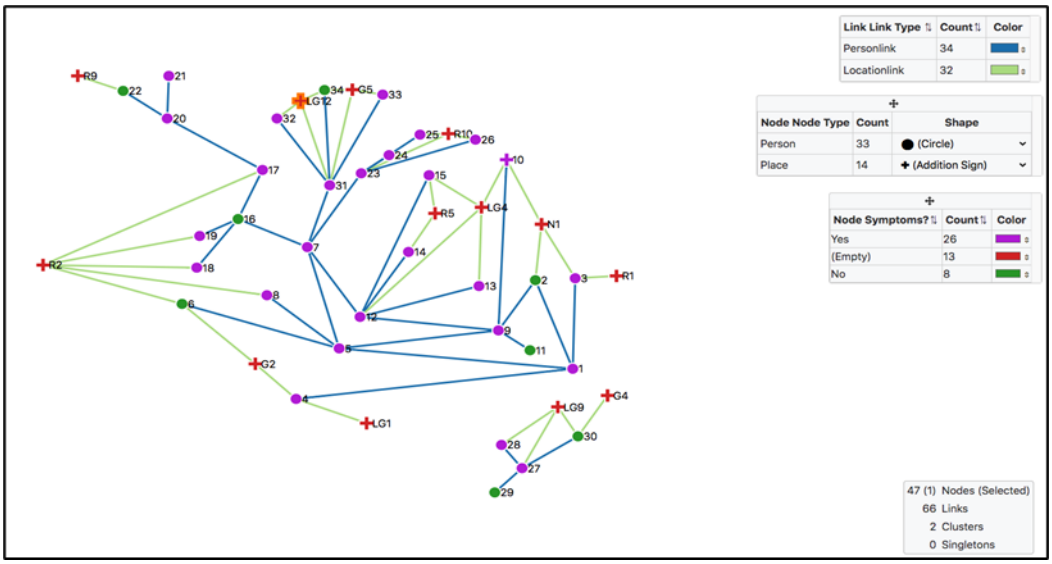

Figure 1. 2D network view: nodes shaped by node type, circle for person, and plus sign for place. Nodes were colored by whether symptoms were displayed, pink is yes, green is no, red is for place nodes that didn't have information about display of symptoms.

Gantt View

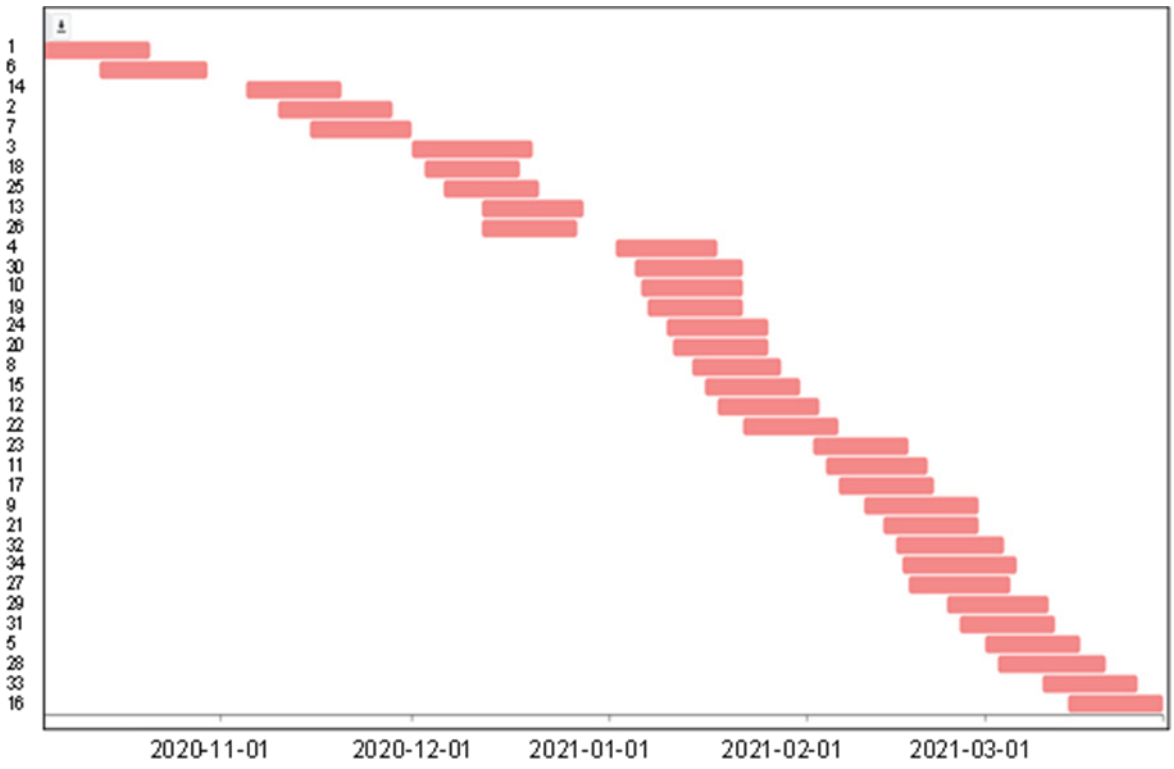

Figure 2. Gantt view: length of each bar represents length of infection for each node (node labels on left). From this graph, average length of infection was calculated.

Gantt View (Figure 2) was used to determine the average length of infection of all people within the network. The parameters entered to use Gantt View were the date of the first positive COVID test to the date of the first negative COVID test. The graph displayed that the length of time date of the first positive and negative test was at least two weeks for all people. The average length of time between the first positive and negative test was referred to as the average length of infection, and it was calculated to be 15.882 days.

Map View

Map view (Figure 3) was used to overlay the network onto a map using the geospatial coordinates in the metadata. The data utilized in the contact tracing network was meant to represent contact tracing data within a specified geographical area. The area used for this study of three Georgia counties and three different zip codes (30004, 30061, 30338). There was one connection between the 30061 and the 30338 zip codes, one connection between the 30061 and the 30004 zip codes, and one connection between the 30338 and 30004 zip codes.

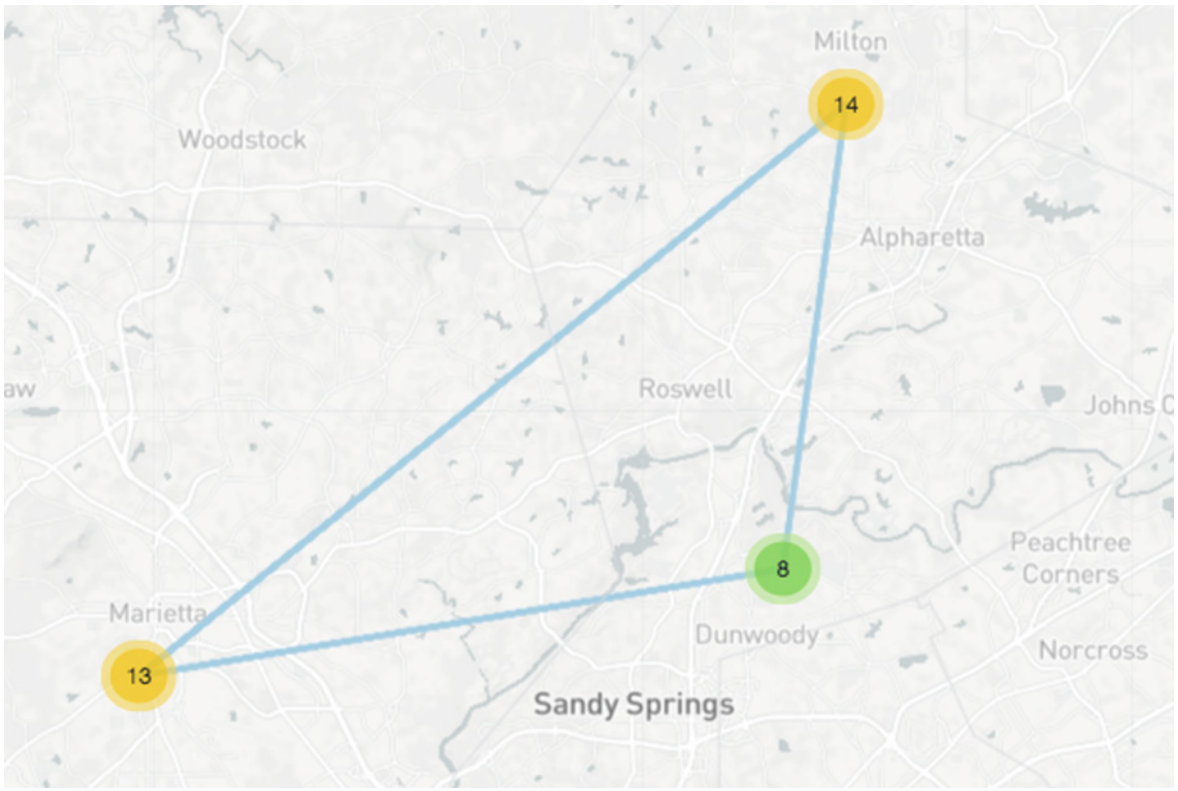

Figure 3. Map view: network across three different zip codes and three different Georgia counties was displayed. Showed that COVID had been transmitted across the different counties.

My observations highlight the importance of the development of a contact tracing network that will point to possible outbreaks are and who appears to be responsible for transmission so that the next necessary protocols, such as quarantining, can be effectively and accurately implemented within a community to reduce the transmission of the disease.

Conclusion

Of the thirty-four person nodes in the dataset, eighteen females had tested positive and sixteen males had tested positive. This data does not point to a significant influence of gender on becoming infected with COVID-19 in the simulated contact tracing network. Of the thirty-four person nodes in the dataset, twenty-six were in the age range from 18-60 and eight were in the age range over 60. This data could be explained possibly because the different age ranges were disproportionately represented in the sample used in this network, or it could point towards the particular age range from 18-60 as more susceptible to becoming COVID-19. Further analysis would have to be done, possibly comparing the results from these age ranges to results in other areas, or a repeat analysis of age could be done using MicrobeTrace with smaller age ranges.

This study primarily showed how utilizing data visualization software like MicrobeTrace provides the benefit of displaying large amounts of data in a manner that simply displays the spread of a particular disease. This software was very valuable in revealing the transmission dynamics of COVID-19 for this study which can be used as a template for further studies into the outbreak of infectious diseases.

From my analysis using the 2D Network View, I was able to see which individuals could have been responsible for the largest amounts of transmission and what places could have been the largest sources of these cases. The

central location of person node 7, as well as the multiple connections to other nodes in the network, flagged this node as a potential "superspreader" of COVID, or a person who may have been central to COVID transmission in this community. Similar conclusions could be drawn about a particular place that appeared to be a source of COVID exposure. In this simulated epidemiologic network, the place which appeared to possess the most connections to people, meaning it was linked to the most number of COVID cases, was R2. R2 was a place node that would represent a restaurant in a true contact tracing network. It was also visible that person nodes with IDs 27, 28, and 30 all visited the same large gathering (LG9). These nodes had also listed each other as contacts on the contact tracing survey, thus LG9 was flagged as a possible source of disease outbreak. Viewing these data with the 2D network view allowed these clusters of interest, an important feature to consider when identifying potential sources of COVID exposure or outbreak, to be identified quickly

The average length of infection which was calculated using the Gantt view and calculating the average length of time between the first positive COVID test and the first negative COVID test was 15.882 days. Length of infection data is particularly useful in defining the infectious period of a particular disease, especially if it is not known. In the case of COVID-19, which is a novel coronavirus, the infectious period could have only been defined by analyzing data such as those used in this simulated network. The infectious period determined by this study is consistent with current data regarding the infectious period of COVID-19, which is approximately two weeks.

In this study, individuals moved within zip codes that corresponded to various cities in the metro Atlanta area. Each zip code was connected by at least one link which demonstrated how individuals moved across these zip codes while simultaneously demonstrating the rapid transmission of COVID-19 across county lines. The network displayed that individuals that resided in different zip codes visited common areas such as restaurants, gyms, and general large gatherings in which people residing in different counties were exposed to one another. Consequently, it was the cross-exposure in these locations that caused them to be later marked as places of COVID-19 exposure. Data analysis revealed that the places that were sources of COVID exposure were large gatherings, restaurants, and gyms.

Limitations

Some limitations of the data generated are that the simulated sample size and the geographical area were very small. When compared to the contact tracing data that is collected about the United States as a whole nation, using only three counties does not provide data to the scale that would usually be seen. In addition, while possible patterns of infection were outlined, no concrete plan to minimize the spread of COVID-19 was developed. Replications of this study could be enhanced by utilizing a larger data sample as well as using the data to ultimately develop a plan of action. An additional improvement for this study would be to create a more robust randomization program to better simulate data. Despite these limitations, this simulated study effectively outlined the importance of contact tracing procedures as well as the benefits of using an effective data visualization tool to map transmission networks and inform of disease spread.

Acknowledgments

I would like to sincerely thank Dr. Anupama Shankar for her mentorship and guidance throughout my project and the drafting of my manuscript.

References

1. Alzu'bi AA, Alasal SIA, Watzlaf VJM. A Simulation Study of Coronavirus as an Epidemic Disease Using Agent-Based Modeling. Perspect Health Inf Manag. 2020 Dec 7;18(Winter):1g. PMID: 33633517; PMCID: PMC7883357.

2. Campbell, E. M., Boyles, A., Shankar, A., Kim, J., Knyazev, S., Cintron, R., & Switzer, W. M. (2021). MicrobeTrace: Retooling molecular epidemiology for Rapid Public Health response. *PLOS Computational Biology*, *17*(9). https://doi.org/10.1371/journal.pcbi.1009300

3. CDCgov. (n.d.). *MicrobeTrace*. GitHub. Retrieved from https://github.com/CDCgov/MicrobeTrace/wiki.

4. Centers for Disease Control and Prevention. (2021, August 5). *Appendices*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/appendix.html.