A Social Network Analysis on the Impact of Zip Codes on a Simulated COVID-19 Transmission Outbreak

Jay Mantuhac

SOCIOL 280

Professor David Schaefer

10 December 2021

Introduction

COVID-19 is an airborne infectious disease, caused by the SARS-CoV-2 virus, that primarily attacks the respiratory system, which can result in different symptoms, including coughing, loss of taste, pneumonia, and potentially, death. Since its discovery in 2019, the virus has resulted in a worldwide pandemic, resulting in the occurrence of 263 million cases and 5.2 million deaths worldwide by December 2021 (*WHO Coronavirus (COVID-19) Dashboard*, n.d.). The virus is transmitted through airborne droplets as a result of close human to human contact, and as such, a multitude of different environmental factors can facilitate the primary transmission of the virus, including high population density, and poor infrastructural conditions, which can result in a lack of access to high quality ventilation (Azuma et al., 2020). Throughout the entirety of the COVID-19 pandemic, however, social factors, including race and ethnicity and socioeconomic status, play a role in differentiating which individuals get infected as a result of these environmental factors. The intersection of such social factors results in a disparity between COVID-19 incidence and prevalence, in which certain groups experience an increased likelihood of COVID-19 transmission, infection, morbidity, and mortality.

Due to the complex nature in which social factors interact with each other, one key demographic factor that can be looked at in lieu of looking at the individual effects of a given social factor is the zip code of a given neighborhood. Key demographic factors, including race/ethnicity and socioeconomic status, play a role as to which individuals live in which zip codes, and so, there is a correlation between where people live and the quality of health and life they may experience in a given geographic area. When comparing the zip code of one neighborhood to another, the interaction between these social factors is so powerful that zip codes can actually act as a more powerful and robust predictor of one's life expectancy compared to an individual’s genetic makeup, such that the expected life expectancy of one zip code can be 20 to 30 years greater compared to the life expectancy of another zip code just miles away (Graham, 2016)**.**

The transmission of COVID-19, as a result, can greatly differ based on the zip codes present in a given geographical area. For instance, Hanson et al. (2020) looked at how different COVID-19 transmission dynamics are by zip code in the state of Indiana from January to March 2020. The study found the median demographic percentage for all the zip codes in the state of Indiana, then compared median COVID-19 infection rates (per 10,000 persons) between zip codes that exceed the median demographic percentage and zip codes that are below the median demographic percentage. Among a variety of key demographics, the median COVID-19 infection rate was 2 times greater among zip codes that exceeded the median demographic compared to zip codes that were below the median demographic. In particular, having an increased percentage of African American residents and increased density of individuals in a given geographic area had the greatest correlation with COVID-19 infection (Hanson et al., 2020), both of which are characteristics that are consistent with factors that are associated with greatly increased COVID-19 transmission and infection.

Social network analysis provides a series of tools and methodologies in modelling COVID-19 transmission on an individual level, and as a result, can play a huge role in looking at the impact that one’s zip code has on fueling transmission of the virus for a given network of individuals. The concept of homophily, for instance, is defined as like-minded individuals, or individuals with similar attributes, being more likely to interact with each other. Given network data, we can assess the impact of homophily on a given characteristic for each individual, including zip code. Furthermore, Exponential Random Graph Models (ERGMs) can also be used to describe the impact that zip codes and other characteristics have on predicting the likelihood that a tie exists between 2 individuals. In other words, ERGMs are analogous to logistic regression models, and in the context of COVID-19 transmission, can be used to predict the likelihood of infection from one individual to another given an individual's zip code, along with other characteristics..

This project has 3 purposes. The first is to assess whether or not there is homophily based on zip code in a network of individual COVID-19 transmission for a given network. The second is to use ERGMs to predict the likelihood of COVID-19 transmission for that same network given the zip code of an individual as well as other different characteristics. Furthermore, we also wanted to use the developed ERGM to test if other demographic factors (primarily race/ethnicity and gender) are better drivers for COVID-19 transmission than zip code.

In addition, exploratory analysis of this network revealed that a majority of nodes present in this network were from the single zip code 30338, and so we hypothesize that the COVID-19 outbreak may have been driven by individuals living in this specific zip code.

Materials and Methods

The data for this project was downloaded through MicrobeTrace, a Javascript-based software application, developed by the Centers for Disease Control (CDC), that is used to integrate, visualize, and conduct data analysis on network-based and bioinformatics data (Campbell et al., 2021). The focus of MicrobeTrace is to get rid of some of the challenges that come with data integration, specifically when looking at molecular data, in order to increase efficiency in gathering insights from such data in an effort to increase the speed of implementing public health responses. Through this, MicrobeTrace can be utilized to visualize network transmission data for COVID-19.

Specifically, the dataset used for this project was a sample dataset generated by the MicrobeTrace in order to provide individuals with the ability to demonstrate the capabilities of MicrobeTrace, and as a result, does not reflect an actual COVID-19 outbreak. This dataset simulates a small COVID-19 outbreak (consisting of 263 individuals in the network) among 11 different zip codes out of the US State of Georgia, with infections occurring between January and March 2020. Given that the dataset is modeling the transmission of COVID-19, we chose to keep the network a directed network in order to showcase the direction of transmissibility between individuals.

The dataset consists of 2 different csv files. The first was a csv file that detailed individual-level nodal attributes, which include demographic characteristics, including the individual's ID number, age, zip code, race and ethnicity, date of symptom onset, date of symptoms resolution, date of positive specimen collection, whether or not they developed pneumonia and acute respiratory distress syndrome (both categorized as binary variables), whether or not the individual died from their infection, and whether or not the individual experienced a given symptom of COVID-19. The second csv file was an edgelist, consisting of 2 columns, that detailed which individual infected which other individual.

After converting the edgelist to a sociomatrix, a binary n x n matrix for n number of nodes present in the network, we generated initial visualizations of the network, which primarily included visualizing the network based on gender, ethnicity, and zip code. In addition, we also generated network level statistics using the sna package, including the number of nodes and edges present in the network, outdegree and indegree distributions, and a dyad census.

After getting these whole network descriptive statistics, we conducted a homophily analysis by using the network package in order to calculate the baseline E-I* index by zip code, which is the ratio between the number of "internal" ties within each zip code and the number of "external" ties between zip codes. The E-I* Index is a measure of baseline homophily present in the network without any adjustments made to control for other network characteristics. In order to adjust for the 11 zip codes present in the network, we also calculated an odds ratio to assess the odds of observing an "internal" tie over the odds of observing an "external" tie.

Furthermore, using the coda and statnet packages, we then generated ERGMs to understand the individual level impacts as to what drives the homophily observed in the descriptive homophily analysis. We generated multiple models and used the Akaike Information Criterion (AIC) value of each model in order to compare which model was best fit for the observed network data. The best fitting model was then used in order to extract odds ratios for observing a tie between 2 nodes for each of the covariates present in the model.

All programming, data wrangling, and statistical analyses were conducted using R Version 4.0.5 ("Shake and Throw"), along with respective R packages described in this section.

Results

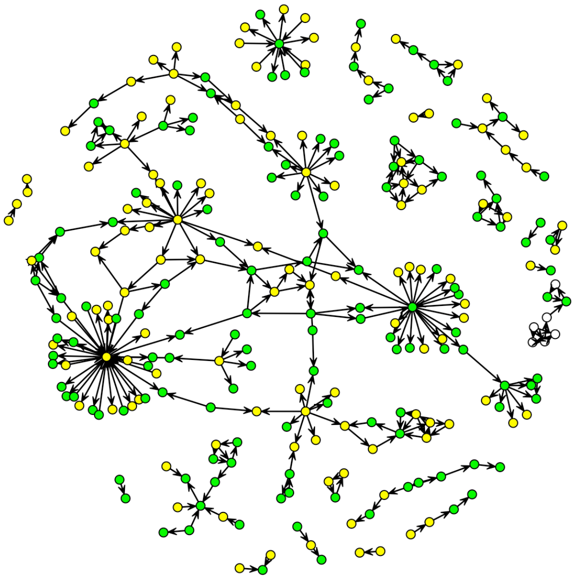


Figure 1. Sociomatrix Colored by Gender

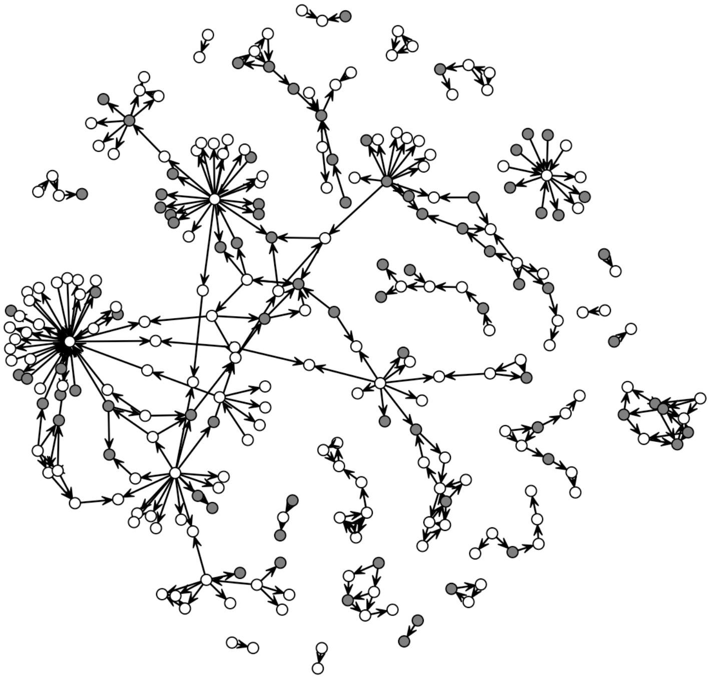


Figure 2. Sociomatrix Colored by Race/Ethnicity

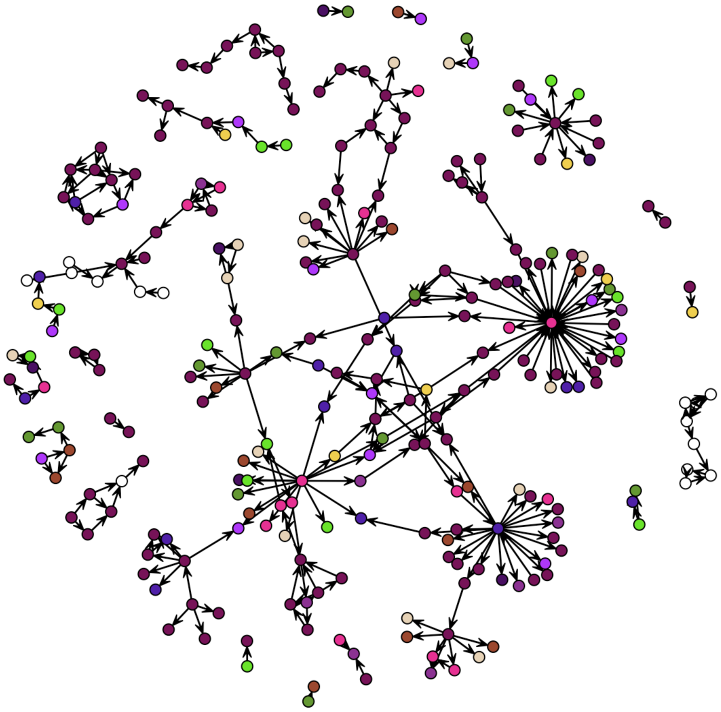


Figure 3. Sociomatrix Colored by Zip Code

After filtering for individuals who do not show up in the network , the final network consists of 263 individuals. From the data points that are present in the original date, there are 119 females and 138 males (Figure 1), 85 Black individuals and 170 White individuals (Figure 2), and a total of 11 different zip codes represented in our network (Figure 3).

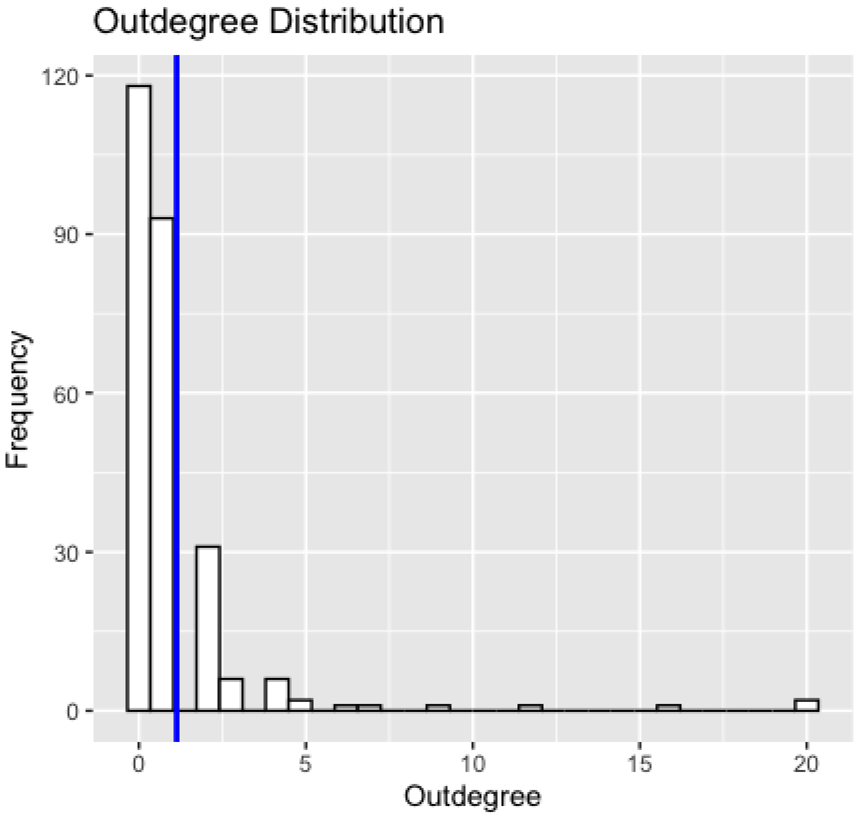


Figure 4. Whole Network Outdegree Distribution, with marked mean outdegree

In the entire network, the outdegree distribution is heavily right skewed, with a mean outdegree value of 1.12 (Figure 4). The mean degree outdegree value can be interpreted as the R_0 value in epidemiology, which, given an individual infected by a disease, demonstrates the number of people that individual is expected to infect (Achaiah et al., 2020). Thus, for the whole network, the estimated R_0 is equal to 1.12.

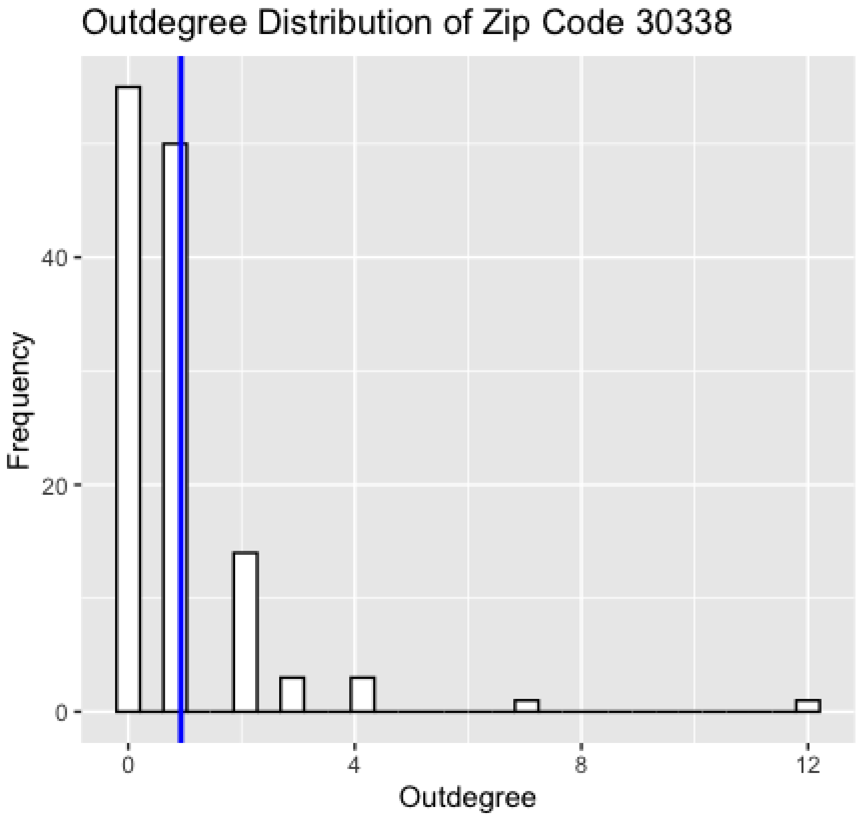


Figure 5. Outdegree Distribution of Zip Code 30338, which consist of the most nodes (n = 127) out of all zip codes in the dataset, with marked mean outdegree

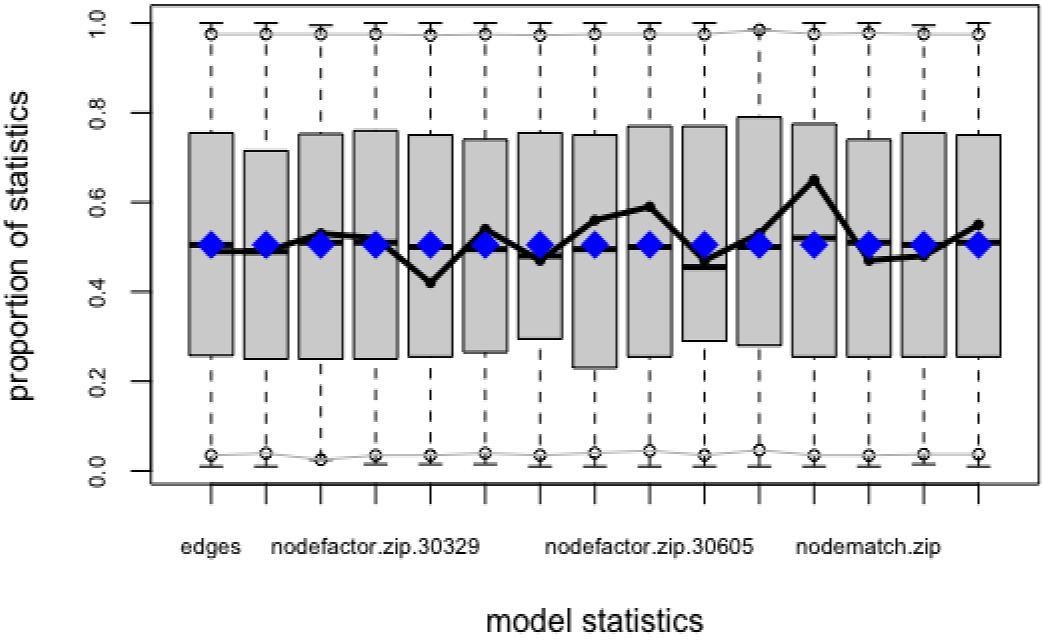
One zip code, 30338, was of particular interest given that most individuals in the dataset belonged to this zip code. For this zip code specifically, the mean outdegree value is 0.93, which corresponds to an R_0 value of 0.93 (Figure 5).

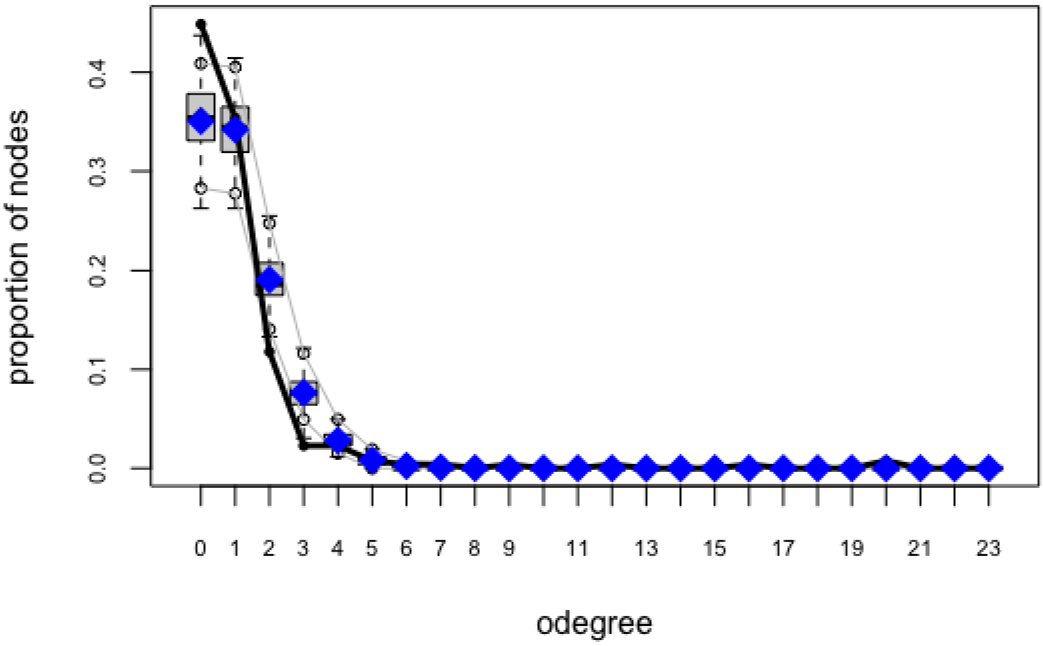
Moving into the homophily analysis, at baseline, the E-I* index was found to be 0.26. In addition, without any adjustment, the baseline odds of observing a homophilous tie is 1.71 times the odds of observing a heterophilous tie (ɑ = 1.71).

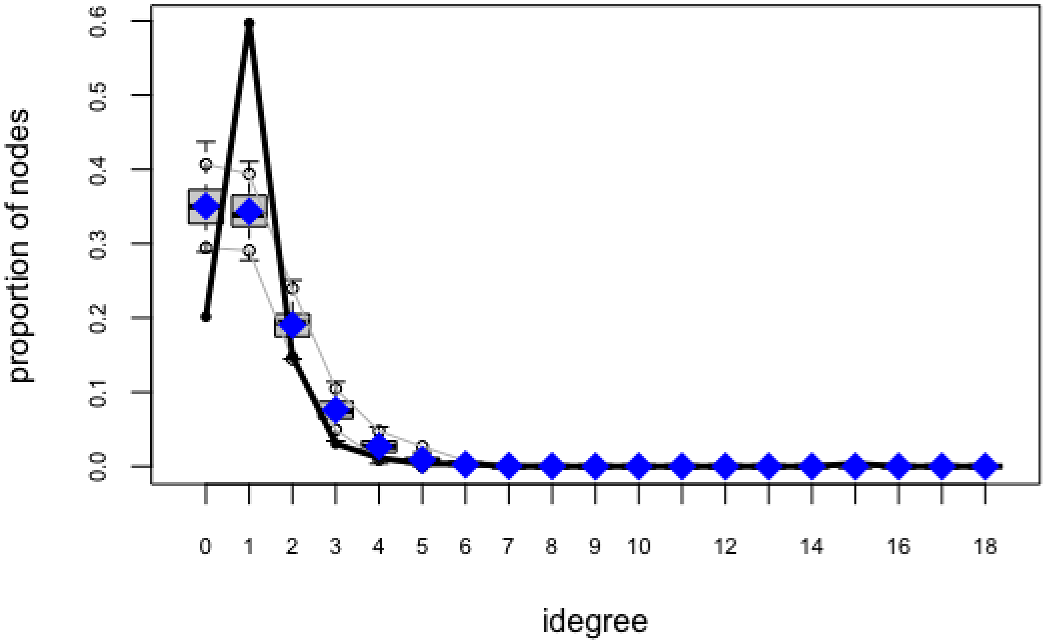
	Estimate	P-value
Edges	**-5.55**	**0.0000**
Zip code 30306	**-1.63**	**0.0094**
Zip code 30324	**0.90**	**0.0018**
Zip code 30329	-0.50	0.2125
Zip code 30338	-0.26	0.3385
Zip code 30340	-0.72	0.0924
Zip code 30341	**-1.20**	**0.0192**
Zip code 30602	-0.30	0.5265
Zip code 30605	-0.72	0.0929
Zip code 30606	**-2.09**	**0.0422**
Zip code 35064	-0.62	0.2276
Zip code 37116	**0.69**	**0.0215**
Homophily - Zip Code	**0.60**	**0.000**
Homophily - Gender	0.16	0.1735
Homophily - Race	0.05	0.6694

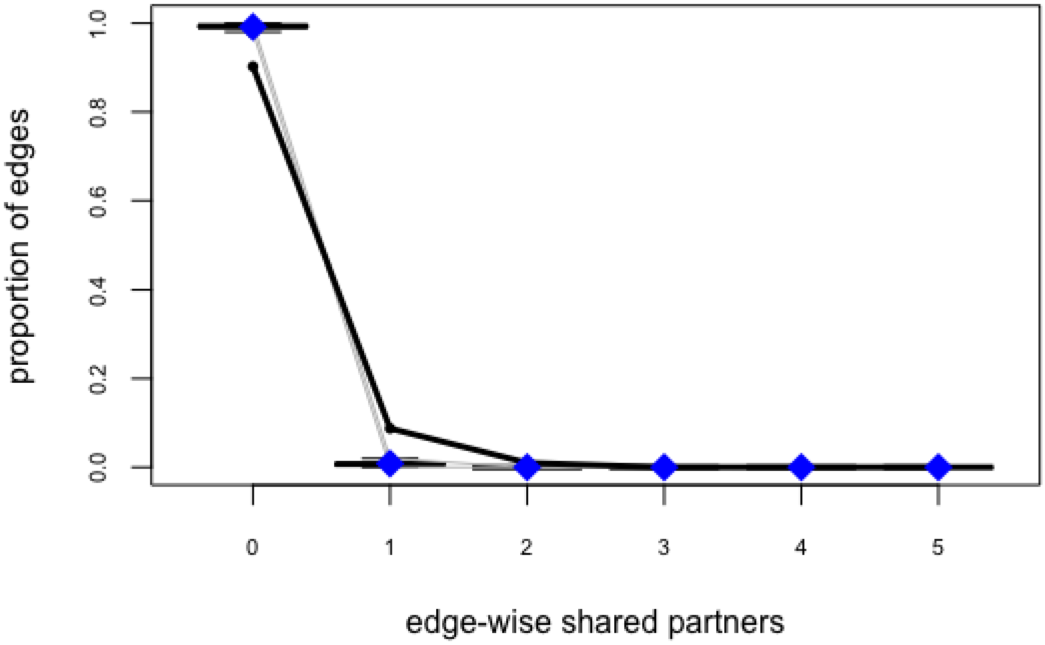
Table 1. Comprehensive ERGM Model Output, with statistically significant covariates bolded

The comprehensive ERGM model was constructed using the nodeofactor arguments to assess for the effect of each zip code on influencing outdegree (i.e. the proxy for our R_0 value) and the nodematch argument to assess for homophily on a given characteristic. From our comprehensive ERGM model, the most significant predictors of observing a tie between 2 nodes, which corresponds to observing a COVID-19 infection between any 2 individuals, are the number of edges, zip codes 30306, 30324, 30341, 30606, and 37116, and homophily driven by zip code (Table 1).


proportion of statistics
0.0
0.2
0.4
0.6
0.8
1.0
edges
nodefactor.zip.30329
nodefactor.zip.30605
nodematch.zip
model statistics


proportion of nodes
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
7
8
9
11
13
15
17
19
21
23
odegree


proportion of nodes
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
7
8
9
10
12
14
16
18
idegree


proportion of edges
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
edge-wise shared partners

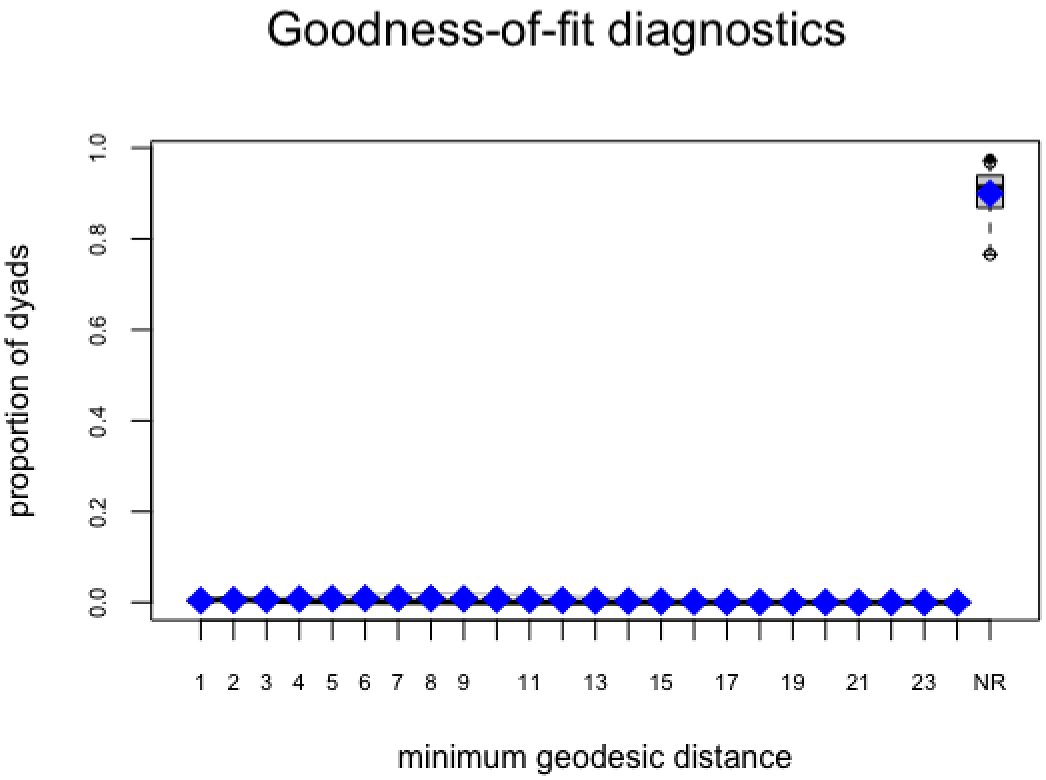


Figure 6. Comprehensive Model Goodness of Fit Plots

Overall, the model demonstrates a decent fit for the network data that we have, given that among all our network statistics, the distribution of simulated values (out of 100 simulations) matches with the observed statistics (Figure 6)

Discussion

When looking at the structure of the entire network, one of the biggest observations found in the sociogram colored by the zip code is the presence of a high number of maroon colored nodes (n = 127) present, which correspond to the single zip code 30338. This observation ultimately led into the question of whether or not an outbreak behind this single zip code is responsible for driving COVID-19 transmission in this network, or whether or not zip code acted as the main driving force for this simulated COVID-19 outbreak. Within zip code 30338, the mean outdegree of 0.93 (acting as the R_0 proxy), indicates that the infection should have died down in this zip code. Although visually, that may not appear to be the case, as a good number of individuals transmitting the disease were from that zip code, most participants who were from that single zip code had an outdegree value of 0, indicating that these individuals were infected by COVID-19, but ultimately did not pass the infection onto any other individuals in the network. As a result, this initial result reveals the COVID-19 transmission within this network may not have actually been driven by this single zip code alone.

One important piece of information to note, however, is that the R_0 value is not an innate trait of the infectious disease alone. Rather, the R_0 value can be influenced by a wide variety of external factors, including population density, the number of people mobile in a given geographic area, and the number of people who are immune to the disease, whether through natural immunity or through vaccination. As a result, the R_0 value can vary depending on the location being assessed. At the time when this outbreak would have taken place (which is between January and March 2020), the original variant of COVID-19 would have been the primary variant circulating among the population. Although estimates for the R_0 of this original strain do vary, given that COVID-19 was still considered a “novel” virus at the time, the World Health Organization (WHO) initially estimated the R_0 value of the original COVID-19 variant to be between 1.4 and 2.4 (Achaiah et al., 2020). Our estimate of the original R_0 value for the entire network ($R_0 = 1.12$) is fairly consistent with the estimate achieved by the WHO.

With that being said, however, we do have evidence that the network demonstrates homophily by zip code. Baseline measures, primarily the E-I* index and the derived odds ratio for observing a homophilous tie over a heterophilous tie ($\alpha = 1.71$) indicate that homophily is present without any adjustment.

The comprehensive ERGM developed for this project does provide adjustment in odds ratios based on zip code, gender, and race and ethnicity, as these are the main demographics that are captured by the dataset. Insights from the ERGM reveal that, similar to what exploratory analysis revealed, the COVID-19 outbreak in this network may not have been influenced by the zip code 30338, as originally hypothesized. A lack of statistical significance for observing a tie in this zip code reveals that we don't quite have evidence that living in the zip code 30338 contributes to increased odds of COVID-19 infection (i.e. observing a directed tie from one individual to another). However, other zip codes in the dataset (30306, 30324, 30341, 30606, and 37116) observed statistically significant values. These results reveal that we have evidence that living in zip codes 30306, 30341, and 30606 act as a protective factor against COVID-19 transmission, while living in zip codes 30324 and 37116 act as risk factor for COVID-19 transmission.

In addition, the statistically significant results from the ERGM also reveal that there is homophily present by zip code and that this homophily also contributes to increased odds of transmitting COVID-19. We also tested if homophily was present based on gender and race/ethnicity, although the lack of statistical significance, despite both being estimated risk factors, suggests that that may not be the case.

Limitations

There are several limitations to this project that need to be noted. The first is that due to a lack of individual-level COVID-19 transmission network data that is publicly available, which would have been at incredibly high risk of violating patient privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA), the dataset used to conduct this project is dummy data generated for the purposes of demonstrating the capabilities of MicrobeTrace to users and stakeholders. As such, this project does not actually reflect any actual COVID-19 outbreak in the US state of Georgia. Although results of this project cannot be used to inform real-life practices and public health policies, the methods used in this project can be applied to analyzing real life COVID-19 data, especially with data generated from COVID-19 contact tracing efforts, as well as other related human-human transmitted infectious diseases.

Furthermore, the overall project scope is limited to COVID-19 transmission occurring between January and March 2020, which, at the time of writing this paper, would have occurred over a year and a half ago. Since that time frame, new research on COVID-19, the development of new COVID-19 variants, and the development of various COVID-19 vaccines would have drastically changed how COVID-19 is transmitted within a given network. As a result, the results gained from this paper would be considered long outdated by the time this paper is submitted. In fact, at the time of writing this paper, the Delta variant of COVID-19 is considered the dominant strain in circulation, which has an estimated R_0 value of 5.08 based on a review of 5 different studies (Liu & Rocklöv, 2021), while very recent news on the rise of the Omicron variant, which is estimated to be more transmissible than the Delta variant, have not yet given enough time for studies to estimate the R_0 value for this variant.

Another limitation to note is the limited number of demographic variables that are captured in the provided dataset. The dataset captures very basic demographic variables without capturing other risk factors for COVID-19 transmission, including the number of people in an individual's household, whether or not the individual is considered an essential worker, and the variant of COVID-19 that the individual had contracted. All of these factors have the potential to increase COVID-19 transmission, such that the R_0 value can be inflated, especially among individuals who may experience or possess a combination of risk factors for COVID-19 transmission. Since such factors were not captured, the external validity of this project is extremely limited. In other words, we cannot generalize the results of this project to make inferences about how COVID-19 is transmitted in a real life community, given that the variables captured in the dataset do not necessarily tell the whole story about how social, cultural, and economic factors interact with each other to facilitate the spread of COVID-19. While we did find that there is a homophilous relationship based on zip code, there are many other factors, besides gender and race/ethnicity, that underlie the observed homophily. It is unfortunate that given the dataset limitations, we could not capture more of those social factors that drive this homophily by zip code.

Finally, the big limitation around the development of the comprehensive ERGM is the fact that all of the covariates present in the model are dyad-independent covariates, and as a result, this model assumes that episodes of COVID-19 transmission from one individual to another are completely independent. Given that, at minimum, there needs to be close proximity between actors in order to actually transmit the disease, the assumption of dyad-independence is easily violated. As a result of this limitation, effect estimates in this ERGM were calculated using Maximum Likelihood methods. Given the lack of triads in the overall structure of the network, we justified that there was no need to include any dyad-dependent variables in our ERGM. However, this decision does come with the limitation that the estimates in the model cannot be calculated using Markov Chain Monte Carlo (MCMC) techniques, which, ultimately, prevents us from evaluating the ERGM for degeneracy. As a result, the ERGM was evaluated using only goodness of fit diagnostic techniques, which tend to be a subjective set of means of assessing whether or not the comprehensive ERGM truly fits the network that we have.

Directions for Future Research

While looking at zip code is a good first step in understanding how social factors contribute to the transmission of COVID-19, understanding the role that individual social factors play and how these social factors interact with each other to facilitate COVID-19 transmission (as well as any other infectious disease in general) are important to ensure effective means of containing the virus through public health interventions and policy. As such, one way of incorporating social network analysis principles into studying the impact and interactions of social determinants of health can include the development of ERGMs for predicting the likelihood of disease transmission. One important direction for such methods is to incorporate interactions between social factors of interest into these ERGMs and to assess the impact that such interactions may have on infectious disease transmission. Furthermore, since individual nodes and dyads are dependent on location, another direction for these models is to also incorporate spatial methods in order to provide better insight into the relationship between location and COVID-19 transmission

Although public health looks at health outcomes on a population level, social network analysis provides a set of tools for assessing the impacts (and complex interactions) of social factors on interactions between individuals. As such, social network analysis is an incredibly powerful tool that should be incorporated into public health research in order to study not just infectious disease transmission, but other health outcomes of interest as well. This project demonstrates just one of the ways of doing so, as the results that come from this kind of research can further solidify public health's advocacy in developing interventions and policies for improving health outcomes in society.

References

Achaiah, N. C., Subbarajasetty, S. B., & Shetty, R. M. (2020). R0 and Re of COVID-19: Can We Predict When the Pandemic Outbreak will be Contained? *Indian Journal of Critical Care Medicine : Peer-Reviewed, Official Publication of Indian Society of Critical Care Medicine*, *24*(11), 1125–1127. https://doi.org/10.5005/jp-journals-10071-23649

Azuma, K., Yanagi, U., Kagi, N., Kim, H., Ogata, M., & Hayashi, M. (2020). Environmental factors involved in SARS-CoV-2 transmission: Effect and role of indoor environmental quality in the strategy for COVID-19 infection control. *Environmental Health and Preventive Medicine*, *25*(1), 66. https://doi.org/10.1186/s12199-020-00904-2

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), e1009300. https://doi.org/10.1371/journal.pcbi.1009300

Graham, G. N. (2016). Why Your ZIP Code Matters More Than Your Genetic Code: Promoting Healthy Outcomes from Mother to Child. *Breastfeeding Medicine*. https://doi.org/10.1089/bfm.2016.0113

Hanson, A. E., Hains, D. S., Schwaderer, A. L., & Starr, M. C. (2020). Variation in COVID-19 Diagnosis by Zip Code and Race and Ethnicity in Indiana. *Frontiers in Public Health*, *8*, 593861. https://doi.org/10.3389/fpubh.2020.593861

Liu, Y., & Rocklöv, J. (2021). The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *Journal of Travel Medicine*, *28*(7), taab124. https://doi.org/10.1093/jtm/taab124

WHO Coronavirus (COVID-19) Dashboard. (n.d.). Retrieved December 4, 2021, from https://covid19.who.int