Statistics and Its Interface Volume 14 (2021) 29–32

The need to incorporate communities in compartmental models

Michael J. Kane* and Owais Gilani

Tian et al. provide a framework for assessing population-level interventions of disease outbreaks through the construction of counterfactuals in a large-scale, natural experiment assessing the efficacy of mild, but early interventions compared to delayed interventions. The technique is applied to the recent SARS-CoV-2 outbreak with the population of Shenzhen, China acting as the mild-but-early treatment group and a combination of several US counties resembling Shenzhen but enacting a delayed intervention acting as the control. To help further the development of this framework and identify an avenue for further enhancement, we focus on the use and potential limitations of compartmental models. In particular, compartmental models make assumptions about the communicability of a disease that may not perform well when they are used for large areas with multiple communities where movement is restricted. To illustrate this phenomena, we provide a simulation of a directed percolation (outbreak) process on a simple stochastic block model with two blocks. The simulations show that when transmissibility between two communities is severely restricted an outbreak in two communities resembles a primary and secondary outbreak potentially causing policy and decision makers to mistake effective intervention strategies with non-compliance or inefficacy of an intervention.

AMS 2000 subject classifications: Primary 37M05; secondary 62P10.
Keywords and phrases: Stochastic block model, SIHR compartmental model, Directed percolation.

1. THE NEED FOR POPULATION-SCALE NATURAL EXPERIMENTS

Epidemiologists have been developing improved methods and tools for modeling disease outbreaks at the population level in order to better understand the spread of communicable diseases, to predict their progression, and to evaluate interventions devised to control their spread. Disease outbreaks like the HIV/AIDS pandemic, the 2009 swine flu pandemic, and influenza have made this need apparent for decades. However, the 2020 SARS-CoV-2 pandemic has created a new sense of urgency for development by surpassing other recent outbreaks in terms of the combination of its communicability and its lethality.

Given the nature and scope of pandemics, it is generally difficult to conduct controlled experiments to test the efficacy of various interventions. Therefore, scientists often rely on natural experiments to evaluate population-scale interventions. In their analysis, Tian et al. [7] propose one such natural experiment using the synthetic control method to assess the effect of delays in imposing mild interventions on the spread of SARS-CoV-2 disease (COVID-19).

The city of Shenzhen in Guangdong province, China, implemented early (within four days of first confirmed case) but mild interventions, namely requiring residents to wear face masks, 14 days isolation for overseas travelers, cancellation of public gatherings, and delayed reopening of schools. To study how effective this intervention was, and to estimate the relationship between number of days between the implementation of such an intervention and the excess number of cases of COVID-19, Tian et al. [7] constructed a synthetic control population using a weighted mixture of 68 counties in the United States such that this population was as similar to Shenzhen as possible along known factors influencing COVID-19 spread. They estimated these weights by combining population density and latitude of the US counties with primary components derived from principle components analysis of the confirmed cases of COVID-19 in these counties. For these two populations, they used the SIHR (susceptible, infected, hospitalized, recovered) compartmental model to compare the trajectories of cumulative confirmed cases over a 16 day period during which Shenzhen had intervention policies in place but "synthetic" Shenzhen did not. Results showed that after six days, the impact of the intervention started becoming visible, with the number of confirmed cases diverging between actual and synthetic Shenzhen after this period. They also concluded that increasing the number of days in delaying an intervention resulted in increased COVID-19 cases (2.32 times more cases after a 4-day delay, and 2.51 times for a 5-day delay).

Quantifying the true impact of any intervention to control the spread of a pandemic in the absence of controlled experiments is challenging. The paper provides a useful framework for policy-makers and scientists to evaluate a mild intervention to curb the spread of COVID-19. Results provide evidence for the need to implement an intervention, even if it is rather mild, as soon as possible soon after an outbreak is suspected. There are some challenges that were

*Corresponding author.

not specifically addressed in this particular analysis. The COVID-19 testing regime in China and the US was very different, particularly in the early days and months of the pandemic. Using confirmed cases to fit the models likely results in an underestimate of the actual impact of the intervention since the actual cases in the US during that time were likely higher. Additionally, the analysis employs the commonly used compartmental SIHR model. While these models are very useful in studying disease spread in fairly homogeneous populations in terms of the mixing of individuals within the population, they can be found wanting when this assumption is not true. Below, we propose an avenue of development that fits into the framework provided by Tian et al. [7] and potentially enhances their results by considering a more localized approach where populations are viewed as a mixture of distinct communities based on the patterns of communication of individuals within and between communities.

2. COMPARTMENTAL MODELS AND THE CONNECTION ASSUMPTION

Compartmental models generally make the assumption that the number of new infections (the force of the infection) is a constant times the current number of infections. Inherent in this assumption is the idea that individuals in a population are homogeneous in the way that they interact with each other, i.e., an individual is equally likely to infect any other susceptible individual in the population. While this assumption generally holds true when observing smaller populations, it may be violated for populations organized into communities where members of the same community are more likely to communicate a disease than two individuals in separate communities. In this case, the number of contacts may be different between individuals in different communities. This situation is more realistic for larger populations or urban centers which tend to be more segregated along various socio-demographic characteristics.

To show how this can affect outbreak dynamics, we provide a simulation of the directed percolation over a stochastic block model [4] with two blocks where the total expected number of edges is kept constant but the inter- and intra-connection probabilities vary. To study this effect, we simulate the outbreak process described in Algorithm 1 on a class of stochastic block models (SBM) to examine the effect on the number of cases over time when the total number of expected edges is kept constant but the expected number of edges between blocks varies.

Let $\mathcal{G}_2(P, \vec{n})$ be a stochastic block model with two blocks. Let P be a symmetric 2×2 matrix where the top-left entry gives the probability that two vertices in block 1 are connected, the bottom-right entry gives the probability that two vertices in block 2 are connected, and the off-diagonal gives the probability that a vertex from block 1 is connected to a vertex from block 2. Let $\vec{n} = [n_1, n_2]$ be a vector indicating

```
Let g be the graph;
Set the state of all vertices of g, V(g) to be "susceptible";
while not all states of V(g) are "removed" do
    if No states of V(g) are "infected" and there are still
      "susceptible" then
        Pick a v from V(g) where the state is susceptible;
        Set the state of v to "infected";
    else
        if Any vertices v_h with state "hospitalize" then
            Move each member of v_h to state "removed"
              with probability p;
        end
        if Any vertices v_i with state "infected" then
            Move each member of v_i to state
              "hospitalized" with probability p;
        end
        if Any vertices v_s with state "susceptible" then
            Pick a vertex v_t at random from the vertices
              in state "infected" with neighboring vertices
              with state "susceptible";
            Set all of the neighbors of v_t in v_s to infected;
        end
    end
end
```

Algorithm 1: A directed percolation on a graph g

the size of block 1 and block 2 respectively. For simplicity in this experiment $n = n_1 = n_2$ and

$$P = \begin{vmatrix} c_1/n & c_2/n \\ c_2/n & c_1/n \end{vmatrix}$$

where c_1 and c_2 are constant values. From this construction, we can see that $\mathcal{G}_2(P, \vec{n})$ can be thought of as a mixture of Erdős-Rényi (ER) graphs [2].

It is well known from Janson et al. [5] and others that when $c_1 \geq 1$ then there is a high probability that there will be a single, large community with all other components having size $O(\log n)$. The presented algorithm is a variation on that of Achlioptas and Molloy [1], which was shown to have asymptotic convergence to a differential equation, which can be solved in closed form for an ER graph with n nodes and connection probability c/n. The bottom-center visualization in Figure 1 shows the behavior of this process and we *conjecture* that, under the correct parameterization for an ER graph or SBM with a single block, the specified directed percolation converges to some set of compartmental models. Like the compartmental model, the directed percolation on an ER graph assumes that the rate of vertices transitioning from susceptible to infected is a function of the number infected, the number susceptible, and the number of contacts per person at any time.

Let $G_1 \sim \mathcal{G}_1(c/(2n), 2n)$ be a sample from an ER graph. Then the expected number of edges is

$$\mathbb{P}E(G_1) = \binom{2n}{2} \frac{c}{2n}$$

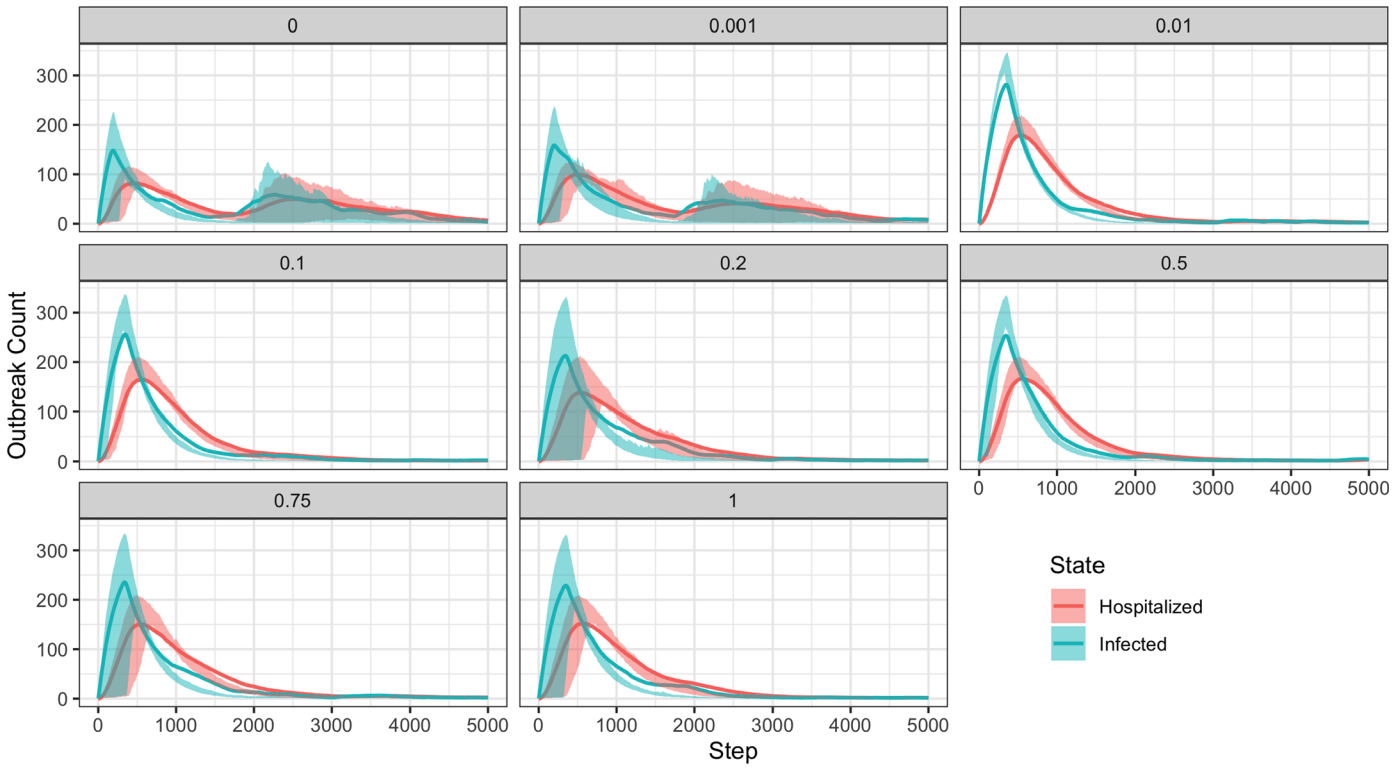

Figure 1. Outbreak simulation results (mean and inter-quartile range) on a stochastic block model with populations of size 1000 over 100 runs. The number above each plot represents the value of α with $\alpha = 1$ corresponding to unconnected blocks and $\alpha = 0$ corresponding to equality in the the intra- and inter-block connection probabilities.

$$\begin{aligned} &= \frac{(2n-1)c}{2} \\ &\approx cn \end{aligned} \tag{1}$$

For $G_2 \sim \mathcal{G}_2(P, \vec{n})$ the expected number of edges is the expected number of edges within the block and the expected number of edges across blocks:

$$\begin{aligned} \mathbb{P}E(G_2) &= (n-1)c_1 + nc_2 \\ &\approx n(c1 + c2) \end{aligned} \tag{2}$$

Equations (1) and (2) show that setting the sum c_1 and c_2 in the stochastic block model to c in the ER graph, with $0 \leq c_1, c_2 \leq 1$ will yield two graphs with approximately the same number of edges, in expectation. We can therefore examine the effect of varying the block connection probability compared to the ER graph for the specified directed percolation by setting

$$\begin{aligned} c_1 &= (1-\alpha)c \\ c_2 &= \alpha c \end{aligned}$$

and allowing α to vary between 0 and 1. In this setup, α controls the level of mixing between the two blocks/communities: when α is zero, the inter-block connection probability is zero and the two blocks will not be connected; when α is one, the inter-block connection probability is the same as the intra-block connection probability resulting in the ER graph. The variables c, c_1, and c_2 can be interpreted as the expected number of individuals anyone will infect in the ER graph, the number of individuals anyone will infect in the same block, and the number of individuals anyone will infect in the other block at time zero, respectively. Thus, c can be related directly to the basic reproduction number R_0.

The simulation was performed for various values of α with c fixed at a value of 3.28, which is the mean estimate of R_0 from Liu et al. [6], in order to show the potential misspecification of compartmental models assuming constant connection probability in a 2-block graph and 1000 vertices. One hundred graphs were generated for each value of α and the directed percolation was run over each graph, keeping track of the number of "susceptible", "infected", "hospitalized", and "removed" vertices, taking the states from Tian et al. [7].

Figure 1 shows the infected and hospitalized counts for the simulation with values of α varying from zero to one. The graph provides two immediate insights. First, the process is robust for a large range of inter-connection probabilities. This implies that the compartmental models likely provide accurate estimates where the population is separated into

communities and the communities are well-connected. Second, when this probability is very small (α close to zero), a secondary outbreak appears to occur. This happens when the process must visit a large portion of a block before a vertex connected to both blocks is labeled "infected." Together these results imply that community-connectedness can be a driver for the appearance of a secondary outbreak, but only when connections between communities are sparse.

3. THE OPPORTUNITY OF INTEGRATING MOBILITY COMMUNITIES

The simulations show that when transmissibility between two communities is severely restricted, an outbreak in the two communities resembles a primary and secondary outbreak, potentially causing policy and decision makers to mistake effective "lock-down" strategies with non-compliance or inefficacy of an intervention. This issue presumably becomes magnified as the number of communities becomes larger and the rate at which an outbreak occurs within a community is much faster than the rate at which the outbreak spreads through the set of communities in a specified area. To address this potential mis-specification, we hypothesize that mobility data could be used to identify distinct spatial areas in which people tend to live and work (also known as "activity spaces" [3]). Human movement both within and between these spatial areas could be used in the proposed algorithm, or one resembling it, to create an agent-based model that more accurately estimates population-level outbreaks by taking into account community-level movement patterns.

ACKNOWLEDGEMENTS

This work was partially supported by the National Science Foundation (NSF) Grant Human Networks and Data Science - Infrastructure (HNDS-I), award numbers 2024335 and 2024233.

Received 10 October 2020

REFERENCES

[1] Achlioptas, D. and Molloy, M. (1997). The analysis of a list-coloring algorithm on a random graph. In *Proceedings 38th Annual Symposium on Foundations of Computer Science* 204–212.
[2] Erdös, P. and Rényi, A. (1959). On random graphs Publ. *Math. debrecen* **6** 290–297. MR0120167
[3] Golledge, R. G. (1997). *Spatial behavior: A geographic perspective*. Guilford Press.
[4] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088
[5] Janson, S., Knuth, D. E., Łuczak, T. and Pittel, B. (1993). The birth of the giant component. *Random Structures & Algorithms* **4** 233–358. MR1220220
[6] Liu, Y., Gayle, A. A., Wilder-Smith, A. and Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine* **27**.
[7] Tian, T., Luo, W., Jiang, Y., Chen, M., Pan, W., Zhao, J., Yang, S., Zhang, H. and Wang, X. (2020). The timing and effectiveness of implementing mild interventions of COVID-19 in large industrial cities. *Available at SSRN 3619819*.

Michael J. Kane
New Haven, CT 06510
USA
E-mail address: michael.kane@yale.edu

Owais Gilani
Lewisburg, PA 17837
USA
E-mail address: owais.gilani@bucknell.edu