Network Analysis of SFU Course Registrations

William Ruth\textsuperscript{1} and Richard Lockhart\textsuperscript{2}

\textsuperscript{1}Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC Canada, wruth@sfu.ca
\textsuperscript{2}Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC Canada

April 28, 2021

Abstract

We investigate the dynamics of disease infection via shared classes at Simon Fraser University, a medium-sized school in Western Canada. Specifically, we use simulation to examine the impact of keeping classes above a certain size online, while those below that size are allowed to meet in person. We use simple models for infection and recovery, as well as multiple regimes for infectiousness and class size threshold. Graph theoretic properties of the student-class enrollment network are also computed, and one such statistic is used to model an important output of our simulations.

Keywords: Disease modelling, network analysis, stochastic simulation.

1 Introduction

1.1 Past Work

Statistical and mathematical models are powerful tools for studying the SARS-CoV-2 pandemic. Many authors have developed sophisticated models to predict the spread of the disease, which have influenced policy and, ultimately, saved lives (Vespignani et al. 2020).
The problem of disease modelling is large and multifaceted. Here, we focus exclusively on transmission within a the context of a university. Specifically, we investigate the effect of moving certain classes online, to see whether limited in-person instruction can be maintained while preventing a major outbreak. Our data come from Simon Fraser University in Burnaby, Canada, but the framework can be easily applied to other institutions.

Giving a complete overview of the SARS-CoV-2 modelling literature here would be impossible. We provide only a brief summary of some closely related work. Models can be broadly classified into two categories: individual level, and differential equation based (Brauer et al., 2019). Individual level models investigate the effects of individual agents’ actions on population level outcomes, whereas the differential equation technique involves directly modelling population level phenomena. We work entirely within the individual level model framework. See Estrada (2020) for an overview of differential equation models for SARS-CoV-2 spread. See Kiss et al. (2017) for a thorough overview of modelling and analysis of disease spread on networks.

Many individual level models have a compartmental structure, such as SIR (susceptible, infectious, removed) or SEIR (susceptible, exposed, infectious, removed) (Brauer, 2008; Deardon et al., 2010), where individuals are assigned to a category based on their disease status, and the researcher models how individuals move between categories. While much work has been done on modelling community transmission (see, e.g., BCCDC, 2021; Chang et al., 2020; Rădulescu et al., 2020; Tuïte et al., 2020), some authors have instead directed their efforts toward understanding outbreaks on university campuses (Gressman and Peck, 2020; Kharkwal et al., 2020; Frazier, 2020; Bahl et al., 2020; Borowiak et al., 2020; Ambatipudi et al., 2021; Weeden and Cornwell, 2020). Gressman and Peck (2020) simulated social dynamics within a university and the corresponding infection rates. They examined the effects of various interventions, including mask wearing, remote instruction,
and random testing; with particular attention paid to the test’s false-positive rate (i.e. specificity). Borowiak et al. (2020) studied the effects of different strategies for grouping students in dorms and classes. Bahl et al. (2020) developed a detailed model of how students and faculty interact on a small university campus. In a seminar presentation, Frazier (2020) discussed both individual-level and differential equation models for disease spread at a large campus; paying particular attention to universal testing schemes and strategies for contact tracing. Kharkwal et al. (2020) developed a detailed framework for simulating infections, which integrates models of various phenomena related to the disease into a single simulation. Ambatipudi et al. (2021) developed a framework for assessing risk of infection over the course of a semester based on room crowding and air circulation. Weeden and Cornwell (2020) took a different approach to investigating enrollment at Cornell University. Instead of studying disease transmission directly, they measured numerous graph-theoretic properties of the enrollment network. Their focus was on measuring the connectedness of the network.

1.2 Our Contribution

We received data on enrollments at Simon Fraser University (SFU), a medium-sized school located just outside of Vancouver, Canada. At SFU the term from September to December is called fall, that from January to April is spring, and the May to August term is summer. There is no winter in Vancouver. Our data cover three terms: spring and fall 2019, and spring 2020. The dataset for each term contains 110,000-120,000 entries, where each entry corresponds to a specific course taken by a specific student. We also have records of on which days each class meets, but not at what time. Our dataset does not include any distance learning courses, co-op courses (i.e. work experience), or courses that do not meet at one of SFU’s main campuses. A number of classes in the dataset do not have any
meeting days.

Along with many other universities, SFU adopted a near-total lockdown policy in response to the SARS-CoV-2 virus. Although this lockdown has dramatically reduced the possibility of on-campus transmission, it has also adversely impacted students’ learning. This and other policies for handling COVID-19 have already been implemented, evaluated, and updated, but it is still worthwhile to evaluate alternative strategies in preparation for any potential future pandemics. One such alternative is to allow some classes to meet in person, provided that this can be done without much additional risk of infection. This approach has the clear advantage of allowing some instruction to happen in-person, but also has the potential for more infections. Particularly catastrophic would be an outbreak on campus, where a large proportion of the student body becomes infected.

The goal of our study is to investigate the likelihood of such an outbreak on SFU’s main campus. Although there are countless ways in which students can infect each other on and off campus, we focus specifically on disease transmission through classes. As such, we chose to omit all non-meeting classes. Ideally, we would have investigated these courses further. However, for privacy reasons we are unable to identify the names of any of these courses, and are thus unable to learn any more about them. Removing these courses will undoubtedly have changed the structure of the enrollment network at SFU, but not in a way that impacts person-to-person contact and thus disease transmission (ostensibly, there is no in-person interaction in a course with no meeting days). We treat labs and tutorials as distinct classes with no inherent connection to the main course with which they are affiliated (other than overlapping enrollment), since each meeting, be it lecture or tutorial, is a separate opportunity for disease spread.

The enrollment network at SFU contains a number of isolated groups of students. That is, groups of students who share classes with each other, but not with anyone outside the
group. In graph theory, these groups are called connected components of the network (see Section 3.2 for more detail). Since the only avenue of disease transmission that we study is via shared classes, and there are no shared classes between components, we should focus on one connected component at a time. It turns out that each term’s network is dominated by a single large component, so we keep only this main component and omit all the small ones. Note that the inclusion or exclusion of tutorials has no effect on which students belong to the same connected component (Alice and Bob share a tutorial if and only if they also share the class to which that tutorial belongs).

We model each term separately. At the start of a term, we infect one randomly chosen individual. We then track how the disease spreads through classes over 90 days (roughly the duration of the in-class portion of a term at SFU), with several different regimes for how likely the disease is to be spread through a single person-to-person contact. We also consider several strategies for moving large classes online to slow the spread of the infection. Multiple simulations are run under each regime, then various numerical and graphical summaries are reported. We find that moving large classes online does slow infection spread. This improvement is seen across all infectiousness regimes, although higher infectiousness requires more aggressive movement toward online instruction.

It is important to note that none of our simulation results should be taken literally. There are many factors that influence how a disease might spread across a university campus, and we can’t hope to model all of them. As such, our findings are meant to be interpreted qualitatively; suggesting trends across variables, rather than as a tool to set specific policy numbers.

In the spirit of Weeden and Cornwell (2020), we also measure graphical summary statistics of the three terms’ enrollment networks. We then compare some of these graphical summaries to the output of our simulation.
Computation is done using the R programming language \cite{R core team 2020}, with network summary calculations making extensive use of the igraph package \cite{Csárdi and Nepusz 2006}.

2 Simulation Study

In order to investigate the relationship between network structure and disease transmission, we carry out a simulation study. We use a susceptible-infectious-removed (SIR) model for the behaviour of SARS-CoV-2. See \cite{Brauer 2008} for a detailed overview of SIR and other disease models.

2.1 Infection Dynamics

The mechanisms by which individuals pass between susceptible (S), infectious (I) and removed (R) is the focus of our modelling. Each student is enrolled in a small number of classes, and we only allow a student to transmit or receive the disease from those with whom they share a class. At the start of a term, we infect one student chosen at random, who then begins to infect their classmates. We model the probability that the disease passes between two students in the same class as inversely proportional to the square root of the class size, with the same proportionality constant across all classes. This means that an infectious student will tend to infect a greater number of students in a larger class, but that the disease is more likely to spread along particular S-I pairs in smaller classes. We chose this profile to reflect the idea that the maximum distance between students is less in smaller classes, but that one person will come into close contact with more of their peers in larger classes (e.g. entering and leaving a large lecture hall). In particular, transmission between a specific S-I pair is more likely to occur in a tutorial that they share than in the corresponding lecture. After some amount of time, an infectious student passes to the
removed group, whereupon they no longer interact with the infection (other than by being included in class sizes).

The dataset we received from SFU gives the days on which each class meets, but not at what time or for how long. As such, we discretize time to the scale of a single day, and consider whether a class does or does not meet on a particular day. On any day that a class does meet, we generate an iid Bernoulli random variable with success probability proportional to $1/\sqrt{n}$, where $n$ is the size of this class. We infect each $S$ individual for which at least one of the associated Bernoulli trials is a success, and otherwise leave their susceptible status unchanged. This process is then repeated for every class that meets on a particular day, and iterated over all days in the term.

For ease of computation, we model the duration of SARS-CoV-2 infection using a geometric random variable. Medical literature suggests that individuals stop being infectious within 10 days of showing symptoms [US CDC 2020], and that the incubation time averages around 5 days [Lauer et al. 2020]. Since the geometric distribution is supported on the positive integers, we cannot constrain recovery times to be strictly within a certain interval. Instead, we require that 95% of infectious individuals recover within 15 days. This constraint gives us a success probability for our geometric distribution of 0.18.

While the description of our simulation sounds computationally demanding, there are multiple opportunities to reduce the amount of work required. First, $S$ student $j$ becomes infected in class $k$ if any of the $I$ students in this class have a successful Bernoulli trial for student $j$. Let $I^{(k)}$ be the number of infectious students in class $k$, and $\theta$ be the proportionality constant in the infectiousness probability (i.e. probability of infection along a single S-I pair is $\theta/\sqrt{n}$). The probability of any successes among the $I^{(k)}$ trials is $p^{(k)} := 1 - (1 - \theta/\sqrt{n^{(k)}})^{I^{(k)}}$, where $n^{(k)}$ is the number of students enrolled in class $k$. Furthermore, we can remove dependence on which $S$ student is being discussed, and instead directly
simulate how many $S$ students are infected in class $k$ using a binomial random variable with $n$ equal to the number of $S$ students in the class, $S^{(k)}$, and success probability equal to $p^{(k)}$. Next, we randomly select from among these $S$ students which ones will become infectious. These simplifications reduce the process of choosing which students are infected in class $k$ from $O(I^{(k)}S^{(k)})$ to $O(1)$.

A similar simplification is possible with the random recovery times. Naively, we would have to generate an observation from a geometric random variable every time a new student becomes infectious, then decrement this random variable at every time step until the value reaches zero. Given that thousands of students may be infectious at any given time step, we can simplify this process considerably by making the following observations about the geometric distribution [Weisstein, 2020]. First, the geometric distribution is memoryless. That is, conditional on still being infectious at time $t$, the distribution of recovery time for a student is the same as the distribution of recovery time when that student was originally infected. While this assumption is clearly not true for real-world disease recovery, the computational savings provided by the memoryless property are seen to outweigh this drawback. In fact, the geometric distribution’s continuous memoryless analogue, the exponential distribution, is popularly used to model waiting times in continuous problems because of how memorylessness eases computation (see, e.g., Almutiry et al., 2020; Kiss et al., 2017). Next, we observe that the geometric distribution models the number of independent and identically distributed Bernoulli trials before the first success. Combining this interpretation in terms of Bernoulli trials with the memoryless property, we see that whether an infectious student recovers at a particular time $t$ can be checked simply by generating a single Bernoulli trial and recovering the student if and only if the trial is successful. This also means that we can identify who, among a group of infectious students, recovers at time $t$ by generating an observation from a binomial random variable, with $n$
equal to the size of the group, and randomly recovering a number of students from the group equal to the outcome of our binomial observation.

### 2.2 Study Design

In our simulation study, we investigate two decision variables. First is the infectiousness of the disease, represented by the infectiousness parameter. We do not have a dataset that would allow us to estimate this parameter empirically, so we instead consider several regimes, with low, medium and high infectiousness corresponding to parameters values of 0.2, 0.6 and 1 respectively.

The second variable in our study is the maximum permissible class size. SFU moved all classes online for the latter portion of spring term (Simon Fraser University 2020a) and all of Summer term (Simon Fraser University 2020b). A small number of courses that require in-person instruction took place on campus for fall term, but most instruction remained online (Simon Fraser University 2020c). We endeavour to provide guidance for administrators, who have to decide which classes to allow on-campus, by investigating the effect of various class-size restrictions on the spread of the pandemic within SFU. Specifically, we simulate disease spread when only classes below a certain percentile are included in the network. Candidate maximum class size quantiles are 50%, 75%, 90% and 95%. For comparison, we also run the simulation with all classes included (i.e. 100%). Table 1 gives the explicit sizes of classes corresponding to the thresholds listed above for each term.

---

1 Accurately estimating the infectiousness parameter directly from data seems like an extremely difficult task, given the challenges inherent in collecting accurate data on epidemics. It is not clear that even the popular $R_0$ value would be of much help (Delamater et al., 2019).

2 As discussed in Section 1.2 we treat tutorials and labs independently of their main course. Here, this means that we might remove a large course from the network while retaining that course’s lab sections. In practice, this amounts to keeping the lecture component of a course online while having students come to campus for in-person labs.
| Threshold | Spring-2019 | Fall-2019 | Spring-2020 |
|-----------|-------------|-----------|-------------|
| 50%       | 18          | 18        | 18          |
| 75%       | 26          | 27        | 27          |
| 90%       | 59          | 63        | 62          |
| 95%       | 102         | 105       | 104         |
| 100%      | 467         | 481       | 429         |

Table 1: Sizes of classes corresponding to each threshold proportion across the three terms.

It is important to note that our goal with this simulation study is not to identify a ‘magic number’ for the maximum allowable class size. Our study contains many simplifications which limit the generalizability of any conclusions we draw (see Section 5.1). Rather, we seek to investigate the sensitivity of disease spread to maximum class size. There is a qualitative difference between a specialized course with 20 students, and an introductory calculus course with 400 students, which does not apply when comparing a class with 80 students to one with 85. We focus on this qualitative difference when varying maximum class size.

For each parameter combination, we run the simulation 100 times. Each simulation run is initialized by moving a single randomly chosen student to the infectious group before iterating through the procedure described in Section 2.1. The response variable that we measure is the trajectory of the infection. That is, the number of students that are susceptible \(S\), infectious \(I\) and removed \(R\) at each time point. We focus our presentation on the infectious trajectory since this group is the primary driver of the infection, and the sizes of the other groups can be approximately inferred from one group’s size (the sum of the three group sizes is constant and there is no direct movement between \(S\) and \(R\)).

To summarize, our simulation study includes two covariates: infectiousness rate, which varies between low, medium, and high, as well as maximum class size percentile, which varies between 50%, 75%, 90%, 95%, and 100%. The response variable of our simulation
| Threshold | Develop Rate |
|-----------|--------------|
| 50%       | 3% (1.7%)    |
| 75%       | 11% (3.1%)   |
| 90%       | 29% (4.5%)   |
| 95%       | 29% (4.5%)   |
| 100%      | 48% (5.0%)   |

Table 2: Proportions of simulation runs where the infection develops into an active outbreak (i.e. infects at least 10 people) for various limits on the proportion of classes allowed to meet in person. Standard errors are given in parentheses.

is the trajectory of the infection, specifically the trajectory of infectious cases.

### 2.3 Results

For simplicity of presentation, we only show results for spring term of 2019 and the high infectiousness regime. For the rest of our results, see the supplemental material for this article [Ruth and Lockhart, 2021].

A striking feature of our simulations is that only a small fraction of infections develop into full-blown outbreaks, with the others stalling in their early stages. Table 2 gives the proportion of simulations that develop into full outbreaks, where we define an outbreak as an infection with at least 10 total people becoming infectious. Call the proportion of infections that do not stall the develop rate of a simulation regime.

Figure 1 gives some sample trajectories for various class size thresholds. Each curve shows the number of students who are infectious on each day of term under a single simulation run. We only display developed infections, and for readability include at most 5 trajectories in each plot (although the actual number of trajectories may be fewer if there are not at least 5 developed infections). The plot for fully in-person instruction (omitted) looks qualitatively like that of 95%, but with approximately twice as many peak infections. Figure 2 gives the cumulative number of students who have become infectious.
(a) 50% in-person

(b) 75% in-person

(c) 90% in-person

(d) 95% in-person

Figure 1: Sample trajectories. Each subfigure correspond to which proportion of classes are included. Each individual curve displays the number of students that are infectious on a particular day in a single simulation. Only infections that develop into outbreaks are shown.

In order to investigate all simulations with active outbreaks for a particular parameter setting simultaneously, we compute the average number of infectious students at each time point under each infectiousness/class size regime. Note that the average is taken only over infections that develop into outbreaks. This gives an ‘average trajectory’ for each parameter combination in each term. We also include ±1 standard error bands to show uncertainty in the mean (also computed only on infections that don’t stall). Figure 3 gives
Figure 2: Sample cumulative trajectories. Each subfigure correspond to which proportion of classes are included. Each curve represents the proportion of students that have been infected on a particular day in a single simulation. Only infections that develop into outbreaks are shown.
some average trajectories. See the supplemental material [Ruth and Lockhart, 2021] for plots with the other parameter combinations and terms.

2.4 On Community Transmission

One might take issue with our choice to start with only 1 infectious individual in each term, especially given the presence of community transmission in many countries worldwide. We counter that our framework remains useful as a tool to study the effect of extra-campus outbreaks. The single infectious student at the beginning of each simulation can instead
be seen as someone who has contracted the disease from a source outside their classes at any point during the semester. The early part of an infection trajectory then represents the effect of this newly infected student over the remaining part of term. For example, if a student is infected by someone they live with (off-campus) on day 30 of class, then the first 60 days in Figures 13 represent the estimated effect of this student over the remaining 60 days of that term.

Our interpretation of mid-term infections using partial trajectories lends an important consequence to the rates at which infections develop into active outbreaks given in Table 2 (i.e. the develop rates). These rates represent the estimated probability that an out-of-class infection leads to an active outbreak. It is crucial to keep these proportions low since, at time of writing, reported prevalence of SARS-CoV-2 in the healthcare region to which SFU belongs is around 0.27%. This means that we can expect some amount of off-campus transmission to occur, and we want to avoid any of these cases leading to an active on-campus outbreak.

3 Network Summaries

Weeden and Cornwell (2020) (hereafter WC) studied registration data for students at Cornell University in the fall of 2019 seeking understanding of the potential for the SARS-CoV-2 virus to spread from student to student by classroom contact. Their work considers three groups of students at Cornell. One group consists of undergraduate students registered in the College of Arts and Sciences (CAS). The second group is all undergraduate students while the largest group adds in almost all graduate students.

Much like in our work, each of the three data sets used by WC is a list of records, with one record for each individual enrollment in a single course. From this data one can construct two networks: a two mode network linking students to their courses and a single
mode network of students, created by regarding two students as being connected if they
took the same class in the same term. An important difference between the Cornell dataset
and ours is that their dataset lacks any information about when or how often classes meet.
This makes their dataset conducive to network analysis, but less helpful for a detailed
simulation study as in our Section 2.

Interest centres on the connectedness of students. The idea is that students who share
a class would come together in reasonably close proximity several times a week; this would
be expected to increase transmission of SARS-CoV-2. These students then have other
classes in common with other students to whom they could potentially pass the virus.
Connectedness is measured using statistics designed for networks (aka graphs).

We perform a similar analysis on the network of students at Simon Fraser University
(SFU). Details of our dataset are described in Section 1.2. While WC had access to infor-
mation at multiple organizational levels (university-wide, undergraduate only and single
faculty), our dataset contains no such stratification. We reproduce some of the various
tables and summary statistics of WC, then discuss the comparison. In the interest of com-
paring quantities which are most similar, we compare our results only to those obtained
by WC on their university-wide network.

Our main findings are:

1. Compared to Cornell, students at SFU take fewer courses in larger classes.

2. The student body is very interconnected at both institutions. That is, for almost all
pairs of students there is a short chain of students in common courses through which
the disease could pass to get from one to the other. This similarity is despite the
differences in class sizes and numbers of classes taken.

3. Many of the summary statistics used by Weeden and Cornwell (2020) don’t take
Table 3: Characteristics of students at Cornell University and at SFU. Cornell data are from Weeden and Cornwell (2020, Table 1). SD is standard deviation.

|                                | Cornell Fall 2019 | Spring 2019 | SFU Fall 2019 | Spring 2020 |
|--------------------------------|-------------------|-------------|---------------|-------------|
| # students (n)                 | 22,051            | 24,071      | 25,089        | 23,836      |
| Courses per student            |                   |             |               |             |
| Mean (SD)                      | 5.4 (1.8)         | 4.5 (2.0)   | 4.7 (2.1)     | 4.5 (2.0)   |
| Median                         | 5                 | 4           | 5             | 4           |
| Co-enrolled students           |                   |             |               |             |
| Mean (SD)                      | 612 (425)         | 309 (241)   | 325 (252)     | 294 (217)   |
| Median                         | 542               | 258         | 266           | 256         |

account of the number of short routes between students; instead they focus on the existence of any short route between two students. If the disease can spread from one student to another by many different routes then transmission from one student to the other will be much more likely than if the only way to get the disease from one to another is by following one of a small number of possible chains. We suggest below that one might compute, as another summary to be considered, the number of paths that lead from one student to another in $k$ steps, for small values of $k$.

### 3.1 General Summary Statistics

General summary statistics for the data are in Tables 3 and 4. These tables combine data for the fall of 2019 for the full network at Cornell, taken from Table 1 in Weeden and Cornwell (2020), with the network at SFU for spring 2019, fall 2019, and spring 2020. Table 3 describes the students in these groups and Table 4 describes the courses.

The numbers of students at SFU are similar to, if slightly higher than, the number at Cornell. We have fewer courses per student and substantially lower numbers of co-enrolled
Table 4: Characteristics of courses at Cornell University and at SFU. Cornell data is from Weeden and Cornwell (2020, Table 1). SD is standard deviation.

|                     | Cornell Fall 2019 | Cornell Spring 2019 | SFU Fall 2019 | SFU Spring 2020 |
|---------------------|-------------------|---------------------|---------------|-----------------|
| # courses (m)       | 6,072             | 3,546               | 3,789         | 3,552           |
| Mean enrollment (SD)| 19.5 (42.9)       | 30.6 (40.1)         | 31.0 (41.0)   | 30.4 (38.4)     |
| Median enrollment   | 8                 | 18                  | 18            | 18              |
| 90th percentile     | 45                | 62                  | 64            | 65              |
| # w/ 100-199 students | 110              | 158                 | 157           | 165             |
| # w/ 200+ students  | 64                | 41                  | 51            | 39              |

students. For clarity, co-enrollment numbers are computed as follows. For each student $i$, count the number of students sharing a class with $i$; count each other student at most once. This is exactly the total of the $i$th row of the matrix $M$. Note, however, that if $i$ and $j$ share several classes together and $i$ becomes infectious then the disease is more likely to be transmitted to $j$ than if they shared only one class together. The numbers recorded in the table are the mean and median of these co-enrollment counts over all students $i$.

Cornell has considerably more courses than SFU, but SFU’s courses tend to be larger. The only exception to this size difference is in the largest of courses (those with 200+ students), of which Cornell consistently has the most.

We now construct the formal networks. Specifically, the records in our dataset are used to create a bi-adjacency matrix, which we call $A$, linking classes and students: 1 row for each unique student and 1 column for each unique class. If student $i$ took class $j$ then there would be a 1 in the $i,j$th entry in the matrix; if not the entry is 0.

The network described by a bi-adjacency matrix can also be represented as a two-mode network (also called a bipartite graph). A two-mode network is a graph where the nodes
can be divided into two groups, and edges only exist between nodes in different groups. In our case, the two groups are classes and students, and edges only exist between a student and a course that student is enrolled in (not between students or between courses).

3.2 The 2-mode network

In Table 5 we provide various summary statistics for the bipartite graphs of the Cornell data set and the three SFU terms. Cornell numbers are drawn from Table 2 of (Weeden and Cornwell, 2020). The bipartite or two-mode graph for spring 2019 at SFU could have a maximum of $n \times m = 24,071 \times 3,546 = 85,355,766$ edges. In fact, there are only 108,554 edges: an edge density of 108,554/85,355,766 = 0.0013. Other terms at SFU are similar, and all densities at SFU are higher than at Cornell.

The formula we use is not the only choice for density. If the two mode structure is ignored there are a total of $n + m = 27,617$ nodes in the graph and the maximum number of undirected edges between any two distinct nodes would be $27,617 \times 27,616/2$. WC used this second normalization in their work; we have adjusted their results to match our definition of density. We provide 1 more digit in our calculations than did WC because the densities are so small.

A popular concept in network analysis, and one which is relevant for thinking about chains of disease transmission, is connectedness. Consider any pair of nodes in the 2-mode graph (students, classes or one of each), and ask whether it is possible to find a chain of

\[ \text{As an aside, these bipartite graph densities tend to be low, and to be lower in schools with lots of students. Consider a network of } n \text{ students and } m \text{ courses. Suppose the average course load for a student is } \bar{L} \text{ courses and the average class size of } \bar{C}. \text{ Then the number of edges in the graph is } n\bar{L} = m\bar{C} \text{ because these are two different ways of counting the number of rows in our original data set. The number of possible edges in the bipartite graph is } nm \text{—this would arise if every student took every course! So the actual edge density is } \frac{n\bar{L}}{nm} = \frac{\bar{L}}{m} = \frac{m\bar{C}}{nm} = \frac{\bar{C}}{n}. \text{ That is, if class sizes are the same at two institutions then the one with more students will have a lower edge density. Or you might note that } \bar{L} \text{ is likely to be in the range 3 to 5 no matter what the institutional size is. So those with higher numbers of courses will likely have lower densities.} \]
Table 5: Social Network Measures for the 2-mode (student-to-course) graph. Data from part of Table 2 in [Weeden and Cornwell (2020)] with SFU added. Network density is given as $l/(nm)$, which compares the number of edges to the maximum possible for a bipartite graph.

|                      | Cornell Fall 2019 | Cornell Spring 2019 | SFU Fall 2019 | SFU Spring 2020 |
|----------------------|-------------------|---------------------|--------------|-----------------|
| # students ($n$)     | 22,051            | 24,071              | 25,089       | 23,836          |
| # courses ($m$)      | 6,072             | 3,546               | 3,789        | 3,552           |
| # edges ($l$)        | 118,314           | 108,554             | 117,588      | 107,902         |
| Network density      | 0.0009            | 0.0013              | 0.0012       | 0.0013          |
| Largest component    |                   |                     |              |                 |
| Proportion: students | 0.991             | 0.971               | 0.966        | 0.969           |
| Proportion: courses  | 0.976             | 0.970               | 0.969        | 0.968           |

nodes starting with $i$ and ending with $j$ so that each two consecutive nodes in the chain are linked. If so then we say $i$ and $j$ are in the same component of $M$. Note that such a chain would necessarily consist of an alternating sequence of students and classes. Conceptually, the students in this chain form a sequence of potential transmissions and the classes in this chain are vulnerable to a chain reaction of outbreaks. For the SFU network in spring 2019, we find 149 components of which one is very large. The other 148 components represent various special cases, but the results reported here focus on the single large component.

Also recorded in Table 5 are the proportions of students and courses in the largest component. That is, the number of courses and students in the largest component divided by the corresponding counts in the entire network. These numbers are very close to 1; almost all students and courses are connected to each other.
3.3 The 1-mode network

The bi-adjacency matrix can be used to build a network of students. This network is represented by an adjacency matrix with 1 row and 1 column for each student. The matrix has a 1 in row \( i \) and column \( j \) if students \( i \) and \( j \) are in some class together; the other entries are 0. For definiteness, the diagonal elements, where \( i = j \), are also set to 0; you cannot transmit the disease to yourself. The graph theory jargon is that our student network / graph has no ‘loops’.

This adjacency matrix is computed in three steps. First, multiply the bi-adjacency matrix above by its transpose. The resulting matrix (denoted \( AA^\top \)) is symmetric; the \( i, j \)th entry counts the number of courses these two students, \( i \) and \( j \), attend together. This matrix is then replaced by a matrix of the same size which is 1 wherever the corresponding entry in \( AA^\top \) is at least 1, and 0 otherwise. Finally, set the diagonal elements to 0 and call the resulting matrix \( M \). This matrix corresponds to a ‘one-mode network’ or a graph.

It is somewhat more natural to discuss connectedness of the one-mode network. For any two students, say \( i \) and \( j \), if we imagine that \( i \) is infectious, then we ask whether it is possible to find a chain of students starting with \( i \) and ending with \( j \) so that each consecutive pair of students in the chain are in a class together. If so then \( i \) and \( j \) are in the same component of \( M \). This is equivalent to asking whether students \( i \) and \( j \) are connected in the two-mode network (the chains in the one-mode network come from those in the two node network by eliminating the courses that connect two students). We define the length of a chain to be 1 less than the number of students in the chain. That is, a chain linking two students via a third has length 2, and a chain linking a student to themself with no other links has length 0. When discussing lengths, it is common to refer to chains as paths.

Table 6 records summary statistics for the network of students; as indicated above,
Table 6: Social Network Measures for the projected 1-mode (student-to-student) graph. Data from Table 2 in Weeden and Cornwell (2020) with SFU added. SFU data reflects only the largest component of the network. Path counts are not given by Weeden and Cornwell.

|                              | Cornell                  |          | SFU                  |          |          |          |
|------------------------------|--------------------------|----------|----------------------|----------|----------|----------|
|                              | Fall 2019                | Spring 19| Fall 2019            | Spring 20 | Fall 2019| Spring 20 |
| # unique edges (l)           | 5,832,358                | 3,706,679| 4,064,905            | 3,500,917|          |          |
| Network density              | 0.024                    | 0.013    | 0.013                | 0.013    |          |          |
| Average geodesic             | 2.466                    | 2.779    | 2.730                | 2.722    |          |          |
| Network diameter             | 10                       | 16       | 16                   | 15       |          |          |
| Proportion reachable in k steps |                         |          |                      |          |          |          |
| k = 1                        | 0.024                    | 0.014    | 0.014                | 0.013    |          |          |
| k = 2                        | 0.594                    | 0.449    | 0.460                | 0.445    |          |          |
| k = 3                        | 0.921                    | 0.894    | 0.903                | 0.906    |          |          |
| k = 4                        | 0.966                    | 0.938    | 0.947                | 0.951    |          |          |
| k-step path counts           |                          |          |                      |          |          |          |
| k = 1                        | —                        | 7.49 × 10^6 | 8.22 × 10^6 | 7.00 × 10^6|          |          |
| k = 2                        | —                        | 3.70 × 10^9 | 4.25 × 10^9 | 3.18 × 10^9|          |          |
| k = 3                        | —                        | 2.09 × 10^{12} | 2.52 × 10^{12} | 1.63 × 10^{12} |          |          |
| k = 4                        | —                        | 1.26 × 10^{15} | 1.58 × 10^{15} | 8.86 × 10^{14} |          |          |
statistics for SFU were computed only on the largest connected component. A few terms in this table warrant definition. First, a ‘geodesic’ is any path of shortest length between two nodes. The distance between two nodes is then the length of a geodesic (in our networks there are typically many geodesics between any two nodes). Two students who share a class are at distance 1. If they do not share a class but each share a class with the same third student their distance is two. The average geodesic is obtained by averaging the lengths of the geodesics between every pair of students. The network diameter is the length of the largest geodesic.

If a pair of students share a class, we say that they are reachable in 1-step (as in WC). If two students either share a class or both share classes with a common third student, we say that the original pair is reachable in 2-steps. More generally, a pair of students is called reachable in \( k \) steps if there is a path of length at most \( k \) which links these two students. Table 6 gives the proportion of student pairs which are reachable in \( k \) steps, for \( k = 1, \ldots, 4 \).

Finally, we give the number of paths of length \( k \) between any pair of students, for \( k = 1, \ldots, 4 \). Although this statistic is not given by WC, we feel that it is especially relevant to disease transmission. Any path of length 1 (i.e. any pair of students who share a class) is a (relatively) easy path for the disease to be transmitted. Any path of length 2 (i.e. any pair of students who share a class with a common third student) is a path for disease spread, albeit one along which transmission is more difficult than a 1-step path. This reduced probability of transmission is offset somewhat by the vastly larger numbers of paths of higher order. There are typically around three orders of magnitude more paths of length \( k + 1 \) in our networks than paths of length \( k \).

Some connectivity properties are invariant to inclusion or exclusion of tutorials and labs, collectively referred to as small sections. Under the assumption that every student
registered in the main course is also registered in at least one of the smaller sections, the presence or absence of short paths connecting students in unaffected by whether we include the small section or not. In particular, in Table 6, this invariance property holds for the average geodesic length, network diameter, and proportion reachable in $k$ steps. However, the number of $k$-step paths is not invariant to the presence of small sections. Students who share a small section of some course represent 2 paths for disease transmission, whereas students who share only the main course represent only 1 path.

There are a few summaries in which the differences between SFU and Cornell seem potentially important to us. First, the average number of co-enrolled students per student is quite a bit larger at Cornell. Table 1 of WC reports an average co-enrollment of 612. It appears, however, from a calculation in the bottom three lines of page 228 that the number which should be compared to the SFU numbers we report is 529; the difference may arise from counting the same pair of students co-enrolled in two classes as two co-enrollments and so on. Since the average class sizes are so much smaller at Cornell and since this ordering is maintained for median and 90th percentiles it appears that the extra co-enrollments may be due to some extremely large classes.

Network densities in the student to student network show the same pattern. However, recall that mean co-enrollment is network density times number of students (see Footnote 3 in Section 3.2). Since the number of students is fairly similar between the two schools, network densities just reproduce co-enrollment figures.

SFU network diameters are much larger than at Cornell. We think, however, that these diameters are much affected by things like some small graduate programs which are connected to the bulk of graduate students only by chains involving small numbers of graduate students taking single graduate courses outside their department or perhaps senior undergraduates taking a graduate course in their discipline. Patterns of cross-listing
graduate and undergraduate courses may differ between the two schools. Thus we think these differences in diameters are not likely to play an important role in whether or not an infection spreads widely.

4 Relationship Between Network Summaries and Simulation Results

As discussed in Section 2.4, we are particularly interested in understanding the development rates of the infection under various class size thresholds and infectiousness rates. Specifically, we believe that the connectedness of the student network is relevant to the rate at which infections develop into outbreaks.

One might imagine measuring network connectedness using geodesics (see Section 3.3), as it is reasonable to imagine shorter paths between students leading to more infections. One limitation of using geodesics to measure connectedness is that this method ignores the number of short paths. For example, a large class represents many opportunities for disease spread, but the tutorials for this large class introduce even more opportunities. Since including or omitting tutorials has no effect on whether students are connected, a connectedness measure based on geodesics will miss the extra potential for infection generated by including tutorials. Furthermore, as we remove large classes from the network, the number of long geodesics may increase, since removing all paths of length 2 between two students changes their geodesic distance to 3, thereby increasing the number of length 3 geodesics.

In light of the above concerns about geodesics, we measure network connectedness using the number of paths of various lengths. To measure the association between development rates and connectedness, we perform a small logistic regression analysis. We focus specifically
on the number of $k$-step paths between students in the network for small values of $k$ ($k = 1, 2, 3, 4$). We measure the number of $k$-step paths and the develop rates at numerous class size thresholds (smallest 50%, 55%,..., 100%). Table 6 in Section 3.3 gives the number of $k$-step paths for the full network. Table 2 in Section 2.3 gives develop rates for some class size thresholds in spring 2019. See the supplemental material Ruth and Lockhart (2021) for path counts on the reduced networks, as well as develop rates for other terms and other threshold values.

Given that the number of $k$-step paths often varies over multiple orders of magnitude, we apply a log-transform to the path counts before modelling.

We use logistic regression to model the relationship between develop rates and log-number of $k$-step paths for $k = 1, 2, 3, 4$. Using stepwise selection with AIC (Hastie et al., 2009), we find that only paths of length $1 - 3$ are relevant for predicting develop rates. Figure 4 gives observed and fitted stall rates under the high infectiousness regime in spring term 2019. Note that the X-axis, class size proportion threshold, is not the input to our model, but is instead monotonically related to the number of $k$-step paths for each value of $k$.

See the supplemental material for plots of other terms (Ruth and Lockhart, 2021).

5 Discussion

The results of our simulation show that moving at least some classes online is critical. A return to on-campus instruction with even 75% of classes in-person could lead to catastrophic numbers of infections.

Notice that infections tend to grow rapidly, then start to shrink and eventually die out before the end of term. Furthermore, many infections stall before developing into a full outbreak. This prevalence of short-lived infections is, at least in part, due to the large
Figure 4: Observed and fitted develop rates using numbers of $k$-step paths for $k = 1, 2, 3$. 
number of small classes in the network. Specifically, approximately 4% of classes have one student enrolled and around 17% of classes have at most ten students enrolled. These small classes offer fewer person-to-person contacts, thus having fewer expected infections per day. The most extreme form of this small-class effect is in one-person classes, in which the infection is unable to spread to other students (although the infection can still spread via an infectious student’s other classes).

5.1 Limitations

There are many limitations to our study which restrict the generalizability of its conclusions. Firstly is the source of the data. Our network is constructed using only data from a single university, SFU. Both the enrollment network and any corresponding summary statistics can differ between universities, as seen with Cornell and SFU in Section 3. While we believe that our methodology is relevant to any potential transmission network, our conclusions can only be applied to the network in our study (i.e. students connected by classes at SFU). Other institutions wishing to replicate our study using their own network data can find our code on Github, which will be made available once we obtain permission to release our data.

We now discuss some limitations which apply to the network for which we have data. We do not mean for this to be an exhaustive list, but rather to illustrate some of the challenges involved with modelling disease spread on a real population. To start, our network only links students through shared classes. As is clear from a cursory inspection of any university campus, classes are not the only way in which students interact. It is conceivable that we could incorporate data on living arrangements for students in residence, but no dataset could account for all the ways in which students meet for coffee, or stand near each other outside a classroom, or on a bus... In short, we cannot account for all the
ways in which a disease can spread throughout the student population, so instead accept that we must limit our study (and therefore its conclusions) to the effect of transmission through shared courses.

Another limitation is the implicit assumption that every student who is enrolled in a class attends every meeting of that class. This assumption is clearly not true. In fact, there may be systematic bias toward lower attendance for classes at less popular times (e.g. the earliest classes at SFU start at 8:30 am). SFU does not keep records of class attendance, so the data required to account for attendance in our model does not exist. Some work has been done to study rates of class attendance (Devadoss and Foltz 1996; van Blerkom 1992), but incorporating these models into our study is beyond the scope of this paper.

The last limitation we discuss relates to class scheduling. At SFU, classes meet at the same times each week, typically for one or more hours on one or more days. We were only able to obtain data for the day(s) on which a class meets. This prevents us from accounting for the amount of time actually spent in a room with classmates. Given more detailed information, we could develop a model which more closely reflects real-world behaviour, but data privacy concerns limit the specificity of the data we are able to access.

5.2 Extensions

In the previous section, we discussed some inherent limitations to our study based on the dataset we were provided. Here we briefly mention some ways our model could be expanded to incorporate other aspects of disease transmission.

As was discussed in Section 5.1, our model does not account for the possibility of infection outside of classes. While it would be impossible to fully model student behaviour, one might introduce a random number of infections at each time step. The addition of random infections from outside the disease model is referred to as a spark term and is
discussed by Deardon et al. (2010). These additional infections would represent out-of-
class interactions that take place on-campus, as well as the possibility of contracting the
disease somewhere off-campus. Random infections could be assigned uniformly across
the susceptible population, or a separate model could be developed to describe students’
heterogeneous risks of transmission outside classes.

There are many ways in which we could make our disease model more sophisticated.
We use the susceptible-infectious-removed (SIR) model for disease transmission, but there
are many other such compartment models (Brauer 2008). Of particular interest for us
might be the susceptible-exposed-infectious-removed (SEIR) model, in which individuals
undergo a latency period where they have been infected with the disease but are not yet
able to infect others. There is also concern about whether it is possible to be reinfected
with SARS-CoV-2 after recovery, which we could describe using a susceptible-infectious-
removed-susceptible (SIRS) model. While some research has been done on reinfection of
SARS-CoV-2 and related viruses (Deng et al. 2020; Gudbjartsson et al. 2020; Edridge
et al. 2020), it is not yet clear how prominent this problem is in humans (nor is it clear
how to model duration of the removed period).

There are many possible control measures to limit further spread by infectious individ-
uals. Examples include testing, quarantining and contact tracing. These methods could
all conceivably be incorporated into a model for disease transmission (see, e.g., Gressman
and Peck 2020). These would all be important for devising a comprehensive strategy for
returning to on-campus instruction. We chose instead to focus exclusively on a single as-
pect of controlling disease transmission, class size. One could also conceivably model mask
usage and physical distancing by decreasing the infectiousness rate; a feature that we have
included in our model. However, a more realistic description of mask usage would need to
account for incomplete and imperfect usage.
Acknowledgements

The authors would like to thank Mallory Gilmore for her thoughtful feedback on an early version of this manuscript, as well as to the staff at SFU who provided us with the data.

References

Waleed Almutiry, Vineetha Warriyar K. V., and Rob Deardon. Continuous time individual-level models of infectious disease: a package EpiILMCT, 2020.

Mythri Ambatipudi, Paola Carrillo Gonzalez, Kazi Tasnim, Jordan T. Daigle, Taisa Kulyk, Nicholas Jeffreys, Nishant Sule, Rafael Trevino, Emily M. He, David J. Mooney, and Esther Koh. Risk quantification for SARS-CoV-2 infection through airborne transmission in university settings. preprint, Occupational and Environmental Health, April 2021. URL http://medrxiv.org/lookup/doi/10.1101/2021.03.31.21254731

Riti Bahl, Nicole Eikmeier, Alexandra Fraser, Matthew Junge, Felicia Keesing, Kukai Nakahata, and Lily Z. Wang. Modeling COVID-19 spread in small colleges, 2020.

BCCDC. Epidemiology & modelling presentations. http://www.bccdc.ca/health-info/diseases-conditions/covid-19/modelling-projections January 2021.

Molly Borowiak, Fayfay Ning, Justin Pei, Sarah Zhao, Hwai-Ray Tung, and Rick Durrett. Controlling the spread of covid-19 on college campuses, 2020.

Fred Brauer. Compartmental models in epidemiology. In Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors, Mathematical Epidemiology, chapter 2, pages 19–79. Springer-Verlag Berlin Heidelberg, 2008.
Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical models in epidemiology*. Springer, 2019.

Sheryl L. Chang, Nathan Harding, Cameron Zachreson, Oliver M. Cliff, and Mikhail Prokopenko. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature Communications*, 11(5710), 2020.

Gábor Csárdi and Tamás Nepusz. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

Rob Deardon, Stephen P. Brooks, Bryan T. Grenfell, Matthew J. Tildesley, Nicholas J Savill, Darren J. Shaw, and Mark E. J. Woolhouse. Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, 20:239–261, 2010.

Paul L. Delamater, Erica J. Street, Timothy F. Leslie, Y. Tony Yang, and Kathryn H. Jacobsen. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases*, 25(1):1–4, 2019. doi: https://dx.doi.org/10.3201/eid2501.171901.

Wei Deng, Linlin Bao, Jiangning Liu, Chong Xiao, Jiayi Liu, Jing Xue, Qi Lv, Feifei Qi, Hong Gao, Pin Yu, et al. Primary exposure to sars-cov-2 protects against reinfection in rhesus macaques. *Science*, 369(6505):818–823, 2020.

Stephen Devadoss and John Foltz. Evaluation of factors influencing student class attendance and performance. *American Journal of Agricultural Economics*, 78(3):499–507, 1996.

Arthur WD Edridge, Joanna Kaczorowska, Alexis CR Hoste, Margreet Bakker, Michelle Klein, Katherine Loens, Maarten F Jebbink, Amy Matser, Cormac M Kinsella, Paloma
Rueda, et al. Seasonal coronavirus protective immunity is short-lasting. *Nature Medicine*, pages 1–3, 2020.

Ernesto Estrada. COVID-19 and SARS-CoV-2. Modeling the present, looking at the future. *Physics Reports*, 869(January):1–51, 2020.

Peter Frazier. Reopening cornell during the covid-19 pandemic, 2020. URL https://datascience.columbia.edu/event/data-for-good-peter-frazier-cornell-university/.

Philip T. Gressman and Jennifer R. Peck. Simulating COVID-19 in a university environment, 2020.

Daniel F Gudbjartsson, Gudmundur L Norddahl, Pall Melsted, Kristbjorg Gunnarsdottir, Hilma Holm, Elias Eythorsson, Asgeir O Arnthorsson, Dadi Helgason, Kristbjorg Bjarnadottir, Ragnar F Ingvarsson, et al. Humoral immune response to sars-cov-2 in iceland. *New England Journal of Medicine*, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2009.

Himanshu Kharkwal, Dakota Olson, Jiali Huang, Abhiraj Mohan, Ankur Mani, and Jaideep Srivastava. University operations during a pandemic: A flexible decision analysis toolkit, 2020.

István Z. Kiss, Joel C. Miller, and Péter L. Simon. *Mathematics of epidemics on networks*. Springer, 2017.

Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation
period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Rădulescu, Cassandra Williams, and Kieran Cavanagh. Management strategies in a SEIR-type model of COVID 19 community spread. *Scientific Reports*, 10(21256), 2020.

William Ruth and Richard Lockhart. Supplemental material for network analysis of SFU course registrations. 2021.

Simon Fraser University. In-person instructions cancelled. https://www.sfu.ca/students/support/updates/in-person-instructions-cancelled.html March 2020a.

Simon Fraser University. Remote methods for summer term. https://www.sfu.ca/students/support/updates/remote-methods-for-summer-term.html March 2020b.

Simon Fraser University. Fall 2020 courses with in-person components. https://www.sfu.ca/students/support/academic-policies-and-procedures/fall-2020-courses-in-person.html 2020c.

Ashleigh R Tuite, David N. Fisman, and Amy L. Greer. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ*, 192(19), 2020.

US CDC. Duration of isolation and precautions for adults with covid-19. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html 2020. Accessed: 2020-10-01.
Malcolm L. van Blerkom. Class attendance in undergraduate courses. *The Journal of Psychology*, 126(5):487–494, 1992.

Alessandro Vespignani, Huaiyu Tian, Christopher Dye, James O. Lloyd-Smith, Rosalind M. Eggo, Munik Shrestha, Samuel V. Scarpino, Bernardo Gutierrez, Moritz U.G. Kraemer, Joseph Wu, Kathy Leung, and Gabriel M. Leung. Modelling COVID-19. *Nature Reviews Physics*, 2:279–281, 2020. ISSN 25225820. doi: 10.1038/s42254-020-0178-4. URL [http://dx.doi.org/10.1038/s42254-020-0178-4](http://dx.doi.org/10.1038/s42254-020-0178-4)

Kim A. Weeden and Benjamin Cornwell. The small-world network of college classes: Implications for epidemic spread on a university campus. *Sociological Science*, 7:222–241, 2020.

Eric W. Weisstein. Geometric distribution. From MathWorld—A Wolfram Web Resource. [https://mathworld.wolfram.com/GeometricDistribution.html](https://mathworld.wolfram.com/GeometricDistribution.html) 2020. Last visited on 2020/07/31.