The Small-World Network of College Classes: Implications for Epidemic Spread on a University Campus

Kim A. Weeden, Benjamin Cornwell

Cornell University

Abstract: To slow the spread of the novel coronavirus, many universities shifted to online instruction and now face the question of whether and how to resume in-person instruction. This article uses transcript data from a medium-sized American university to describe three enrollment networks that connect students through classes and in the process create social conditions for the spread of infectious disease: a university-wide network, an undergraduate-only network, and a liberal arts college network. All three networks are "small worlds" characterized by high clustering, short average path lengths, and multiple independent paths connecting students. Students from different majors cluster together, but gateway courses and distributional requirements create cross-major integration. Connectivity declines when large courses of 100 students or more are removed from the network, as might be the case if some courses are taught online, but moderately sized courses must also be removed before less than half of student-pairs are connected in three steps and less than two-thirds in four steps. In all simulations, most students are connected through multiple independent paths. Hybrid models of instruction can reduce but not eliminate the potential for epidemic spread through the small worlds of course enrollments.

Keywords: COVID-19; social network analysis; social epidemiology; higher education; small worlds

O N March 6, 2020, the University of Washington became the first university in the United States to shift to online instruction in response to SARS-CoV-2, the novel coronavirus, and COVID-19, the disease it can cause. A number of institutions soon followed suit, and by late March, most universities had either suspended classes or shifted to online instruction. By early May, many institutions had cancelled face-to-face summer sessions and were considering whether and how face-to-face instruction could resume again.

The epidemiological justification for canceling face-to-face instruction is that infected students can spread a virus to other students in the classroom, who can then infect students in their other classes. In network terms, face-to-face instruction creates a bipartite, or "two-mode," network (see Breiger 1974; Borgatti and Everett 1997) in which students are connected through their classes and classes are connected through their students. To be sure, co-enrollment in a class with someone who is carrying the virus does not necessarily mean exposure to the virus—sick students may not attend classes, or they may be seated some distance away. Similarly, exposure does not necessarily mean infection and will depend on the characteristics of the virus. Additionally, enrollment in the same class does not capture all possible sources of contact between students, a point to which we return in the discussion. Even so, co-enrollment networks represent a major source of social structure in

college students' day-to-day lives, and the features of these networks can affect how a virus spreads.

The fields of social epidemiology and mathematical biology offer ample evidence that social networks contribute to distribution and diffusion of risky health behaviors and disease (Dickison, Havlin, and Stanley 2012; Hill et al. 2010; Luke and Harris 2007; Miller and Kiss 2014; Morris 2004; Valente 2010). Interpersonal contacts in everyday settings can contribute to propagated outbreaks of disease through direct contact—for example, when an infected person sheds the virus through respiratory droplets or aerosols in their breath. Social distancing can mitigate such transmission (Caley, Philip, and McCracken 2008; Cooley et al. 2016; Kelso, Milne, and Kelly 2009; Poletti, Ajelli, and Merler 2012; Valdez, Macri, and Braunstein 2012). Community spread can also occur via indirect contact, as when an infected person sheds the virus onto a surface that becomes contaminated.

In this context, the structure of the social network prior to social distancing can provide insight into factors that increase the likelihood of both direct and indirect exposure to a virus. These factors include the number of contacts people have, whether people are clustered in a single component or are concentrated in subgroups that are connected through only a few bridges or hubs, and the length of contact chains that connect people to each other indirectly. A two-mode network can also provide insight into indirect contact that occurs via exposure in common forums such as stores, restaurants, clubs, organized group events, and, as we examine here, classrooms (Binson et al. 2001; Cornwell and Schneider 2017; Feld 1981; Frost 2007; Laumann et al. 2004; Niekamp et al. 2013; Oster et al. 2013).

Prior research on the social networks of students tends to focus on questions about how a particular network structure comes into being (e.g., racially homophilous friendship networks) or how a particular network structure causally affects an outcome such as student grades or college enrollment decisions (see Biancani and McFarland 2013 for a review). Some research has explored the "class size paradox," which shows that even though average class sizes tend to be relatively small, most students get exposure to a much larger number of other students because of the presence of a few large courses (Feld and Grofman 1977). This has major implications for the prospect of epidemic spread of disease on campus via direct contact alone. But, little work has been done to describe the overall structure of networks among students. In a notable exception, Israel and colleagues (2020) used transcript data from the University of Michigan to identify specific courses and students that have high degree centrality, meaning they act as "connectors," or hubs, across the network. Our analysis complements this important effort in a different context by focusing on how the social organization of universities and classes into majors affects network structure and the attributes of this structure that are most relevant to urgent social epidemiological and policy questions.

We use complete transcript data from Cornell University to describe and visualize the structure of three two-mode enrollment networks during a typical semester: the university-wide network that includes all undergraduates, graduate students, professional masters students, and continuing education students; the network of undergraduate students; and the network of undergraduate students and courses in Cornell's liberal arts college. In each case, we examine the clustering of students

by their fields of study. We also simulate enrollment networks that remove courses from the observed data according to their size, as might be the case in hybrid models of instruction in which some courses are taught online and others in person. These analyses allow us to assess how large courses contribute to the "small worlds" of course enrollment networks and, analogously, how removing these courses might reduce the factors that influence whether course enrollments potentiate epidemic spread of a virus.

It is not our goal to provide an epidemiological model predicting the spread of COVID-19, for which much more information would be needed on the attributes of the virus, the physical distancing behavior of students and instructors within classes, the immune response of individuals within the network, and other parameters. Instead, we have three goals. First, we describe the basic features of co-enrollment networks on a college campus that can affect the risk of a viral spread. Our hope is that this will provide a starting point for network epidemiologists who do wish to estimate a predictive model under different assumptions about R0 and other parameters. Second, we advance existing empirical knowledge of the social structure of higher education, using one university as a case study (see also Israel et al. 2020). Third, and perhaps most critically, we provide relevant evidence to university leaders who must balance the benefits of face-to-face instruction against the potential risk it entails, and to the faculty, staff, parents, and students who are trying to understand these decisions.

Data

Our transcript data are from Cornell University, a residential university in Ithaca, New York, that enrolls approximately 15,000 undergraduate students, 6,200 graduate students, and 2,700 professional students. Each college at Cornell admits its own students and sets its own graduation requirements. Although there is no university-wide curriculum, most colleges allow students to take courses outside their college of enrollment, and some colleges outsource their introductory or required gateway courses to larger colleges. Integration also occurs at the course level through "cross-listing," wherein a course taught in its originating department (the "parent" course) also has course numbers in other departments or colleges ("child" courses).

The data cover all undergraduate, graduate, and professional masters students who enrolled in courses in the fall of 2019, excluding those who spent the semester in a study abroad program. We exclude courses that are not taught on the main Ithaca campus and, analogously, students who took all of their courses away from Ithaca. We also exclude 12 courses that are based in Ithaca but taught through asynchronous distance learning as well as all courses taught through Cornell's executive education and joint MBA programs.[1] Conversely, we include all low-credit courses (e.g., physical education, some art or music courses for nonmajors). We combine cross-listed courses with their parent courses, given parent and child courses meet at the same time and place and with the same instructor. We also combine co-meeting courses, which are most often seminars offered to both upper division undergraduates and early career graduate students.

We treat all sections, meaning independently taught iterations of the same course, as separate courses. Sections are most often found in two types of courses: large gateway courses, where for pedagogical reasons departments would rather split up larger courses; and independent study or research courses, in which students sign up for a section corresponding to the faculty member supervising their study. The latter creates a large number of one- and two-student classes and contributes to a right-skewed distribution of classes by size (see Table 1) but accurately reflects the student experience on this campus. We do not, however, treat mandatory weekly discussion sections that are attached to larger lectures as separate courses.

Table 1 provides descriptive statistics at the student and course level for the three analytic samples. The university-wide data set includes 22,051 students and 6,072 courses. The undergraduate-only data include 14,811 students and 4,209 courses, including co-meeting courses with graduate students. The liberal arts college data include 4,434 undergraduate students and 1,652 courses taught in the College of Arts and Sciences (CAS) and the cross-college undergraduate program in biology. Students in the liberal arts college can also take courses offered by other colleges, although with some restrictions imposed by minimum in-college credit hours embedded in the graduation requirements. Because non-CAS courses are excluded from the liberal arts analytic sample, the average number of courses per student in Table 1 (column 3) is lower for the CAS students than it is for all undergraduates (column 2).

We differentiate students by their field of study. To keep the network graphs visually manageable, we collapse majors and graduate fields into six categories: humanities, fine arts, performing arts, and design; social sciences; science, technology, engineering, and math (STEM); multidisciplinary and mixed; undeclared; and law and business. The humanities, social sciences, and STEM categories include undergraduates with two declared majors in the same field: for example, a student double-majoring in biology and chemistry is coded as STEM. The "multidisciplinary and mixed" category includes students in "design-your-own" majors and students in majors or graduate fields that explicitly require courses in at least two of the broad fields of STEM, social sciences, and humanities. It also includes dual-major students whose majors do not fall in the same field: for example, biology and English. Most students in the "undeclared" category are first- and second-year undergraduates in the liberal arts college and one other smaller college, although some are in a continuing education program. Students in Cornell's other undergraduate colleges are typically admitted directly into a major or, in the case of engineering, can be safely coded as STEM even if they haven't declared a specialty.

We use the transcript data to construct for each of the three analytic samples an affiliation matrix, $\mathbf{A}$, which is a binary, two-mode matrix (see Borgatti and Everett 1997; Wasserman and Faust 1994). $\mathbf{A}$ contains information about a set $\{n_1, n_2,...,n_g\}$ of g students who are arrayed down the rows of the matrix and their ties to a set $\{m_1, m_2,...,m_h\}$ of h courses/sections, which are arrayed across the columns of the matrix. $\mathbf{A}$ contains $g \times h$ total cells. The cells in $\mathbf{A}$ contain a set $\{e1, e2,...,el\}$ of l lines between these two classes of nodes. A given cell in $\mathbf{A}$—for example, $<n_1, m_1>$—indicates whether student n_1 was enrolled in course m_1 (0 = "no," 1 = "yes").

Table 1: Characteristics of students and courses in the university, undergraduate, and liberal arts college networks

	University	Undergraduate	Liberal Arts College
Students			
Number of students (n)	22,051	14,811	4,434
Majors (%)			
Humanities, Arts, and Design	7.4	7.5	14
Social Sciences	12.5	15.1	28.2
STEM	50.6	50.4	47.7
Multidisciplinary or Mixed	6.0	7.8	4.6
Undeclared	2.2	1.7	5.2
Business and Law	21.3	17.6	0.3
Mean courses per student	5.4	5.6	3.7
	(1.8)	(1.2)	(1.4)
Median courses per student	5	6	4
Mean co-enrolled students	612.1	697.3	161.6
	(425.2)	(426.9)	(124.2)
Median co-enrolled students	542	639	
Courses			
Number of courses (m)	6,072	4,209	1,652
Mean size	19.5	19.7	9.8
	(42.9)	(45.4)	(18.4)
Median size	8	9	5
90th percentile	45	42	16
Number with 100–199 students	110	74	14
Number with 200+ students	64	58	1

Notes: Data are from student transcripts from Cornell University, Fall 2019. Course enrollments in undergraduate and liberal arts columns only include undergraduates in the university or the liberal arts college, respectively. Liberal arts college students may also take courses outside the college; these courses are not counted in their average or median number of courses. Standard deviations are presented in parentheses.

Many structural features of network **A** have implications for direct or indirect epidemic spread of disease. One is the network's overall level of cohesion, which can be measured in several ways. We calculate two-mode network density, which is the number of observed ties divided by the total possible ties, and the number of ties in the largest component of **A**, meaning students/courses that are at least indirectly connected to each other through at least one path.

This represents a minimal threshold for connectedness and is a fairly weak indication of potential for virus transmission. We therefore also examine the extent to which the course network connects students through multiple potential transmission paths. We calculate the proportion of students who are members of a two-mode bi-component—that is, students who are connected through two or more independent paths (Cornwell and Burchard 2019). Students who are connected in this way remain connected even when any randomly selected (e.g., large) course moves online or when any randomly selected (e.g., well-connected) student is absent from the network. This helps to quantify the number of short but indirect pathways through which a virus might propagate between students.

We are also interested in the possibility that the positions of some actors in the network, because of how it is wired, might play a disproportionately large

role in transmitting between other pairs of actors. To assess this, we calculate the betweenness centralization of the overall bipartite network. Betweenness centrality refers to the extent to which a given node sits on the shortest (geodesic) paths that link all of the other pairs of actors in the network (Freeman 1977, 1979). A node with high betweenness centrality serves as a common gateway between nodes. Betweenness centralization quantifies the extent to which particular nodes are more central than other nodes in the network. In the two-mode context (see Borgatti and Everett 1997), it provides a sense of how both courses and students constitute particularly influential vectors of transmission between each other, irrespective of to which vertex set a given node belongs.

We create a one-mode projection of $\mathbf{A}$ (derived by multiplying $\mathbf{A}$ by its transpose, $\mathbf{A}^{T}$), $\mathbf{A}_P$, which yields a valued matrix that reflects the number of courses in which each student is co-enrolled with each other student (Breiger 1974). We dichotomize $\mathbf{A}_P$ and then use the resulting binary network to analyze the structure of (potential) student-to-student contact via shared courses. We derive several measures from the projected network. The clustering coefficient, which measures transitivity (Holland and Leinhardt 1971; see also Watts and Strogatz 1998 for a closely related measure), indicates the extent to which students who are enrolled in courses with a common third party also tend to take courses with each other. Such clustering provides opportunities for reinforced contact transmission.

We also examine the average geodesic distance between each pair of students via their courses. A distance of 2, for example, means that while a given pair of students is not enrolled in any courses together, the two students are enrolled in different courses with some third student in common. Finally, we calculate k-step reach between student-pairs, which indicates how many steps exist between them in the projected student-to-student network (e.g., distance of $k = 1$ means that the students are in a class together).

The Small Worlds of Course Enrollment Networks

University Network

The two-mode network for the university, including all graduate and undergraduate students, is depicted in Figure 1. This network includes 28,123 nodes (6,072 courses and 22,051 students) and 118,314 edges (i.e., course enrollments). Light gray circles represent courses, colored circles represent students, and course enrollments are indicated with light gray lines that link students to their courses. Nodes are arranged in the two-dimensional space using the Fruchterman-Reingold algorithm in Pajek64 5.08 (see Batagelj and Mrvar 2018) such that students are positioned close to the courses in which they are enrolled and, by extension, other students who are enrolled in the same courses.

The university enrollment network is highly structured by discipline and student level. Students in the social sciences (dark blue circles), and STEM (orange circles) occupy identifiable regions in the network, although the regions shade into each other.[2] The law and business students (light blue circles) are split into two clusters: the cluster in the lower left part of Figure 1 is primarily undergraduates,
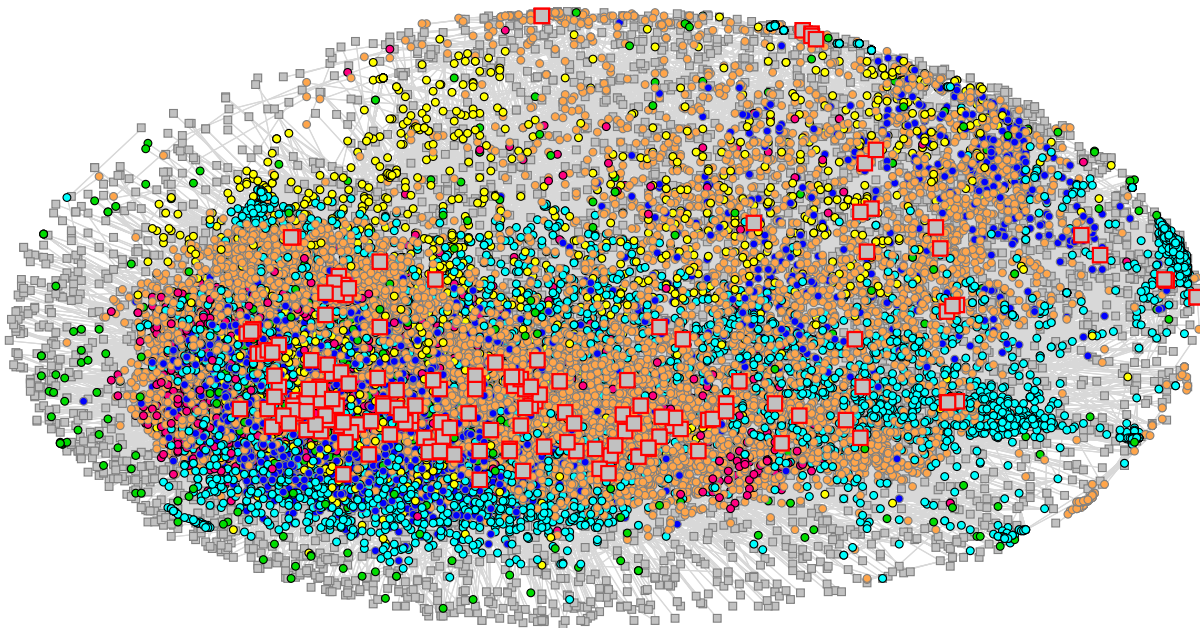
Figure 1: University network. *Notes*: This network layout was arranged using the Fruchterman-Reingold algorithm in Pajek64 5.08 (Batagelj and Mrvar 2018), and nodes were colored, sized, and shaped using Netdraw (Borgatti 2002). Light gray squares represent courses, and larger gray squares with red borders indicate courses with 100 students or more enrolled. Students are represented by small circles with colors identifying their major(s): yellow = humanities, arts, and design; dark blue = social sciences; orange = STEM; red = multidisciplinary/mixed; green = undeclared; light blue = business and law. Students' enrollment in particular courses is indicated with light gray lines. This diagram excludes 338 nodes that were not connected to this main component.

and the cluster running in diagonal from upper left to lower right is primarily graduate and professional masters students. More generally, undergraduate students are clustered in the lower left part of the network diagram, where students and courses are closer together. Multidisciplinary and mixed-major students (red) are scattered throughout Figure 1 but tend to occupy the space near the social sciences and STEM. Humanities, arts, and design students (yellow) are also interspersed around the diagram but are more often found in the upper left region of the diagram (suggesting a law and humanities cluster) and toward the periphery, where classes are smaller. Students who have not declared a major (green) tend to be found near other undergraduates in areas where the STEM, social sciences, and multidisciplinary fields are integrated.

The university course enrollment network has low overall density but considerable clustering (see Table 2, column 1). The density of the student-to-student projection of the two-mode network is 0.024, meaning that a given student shares a class with 2.4 percent of other Cornell students, on average. The average student has the potential to share a classroom with about 529 different students by the time they have completed one round of courses on their schedule (5,832,358*2 bi-directed arcs["edges"]/22,051 students = 529). This does not mean that the average student

Table 2: Characteristics of university, undergraduate, and liberal arts college course enrollment networks.

Social Network Measures	University	Undergraduate	Liberal Arts College
2-Mode (Student-to-Course) Network			
Number of students (n)	22,051	14,811	4,434
Number of courses (m)	6,072	4,209	1,652
Number of edges (l)	118,314	83,024	16,232
Network density	0.000	0.001	0.001
Proportion of students in largest component	0.991	0.999	0.998
Proportion of courses in largest component	0.976	0.991	0.993
Proportion of students in largest bi-component	0.945	0.998	0.926
Proportion of courses in largest bi-component	0.730	0.778	0.736
Betweenness centralization	0.098	0.096	0.115
Projected 1-Mode (Student-to-Student) Network			
Number of unique edges (l)	5,832,358	4,582,808	332,643
Network density	0.024	0.042	0.034
Clustering coefficient (transitivity/closure)	0.480	0.460	0.503
Average geodesic distance	2.466	2.060	2.281
Network diameter (largest observed distance)	10	4	5
Proportion reachable in k steps			
$k = 1$	0.024	0.042	0.034
$k = 2$	0.594	0.896	0.686
$k = 3$	0.921	0.998	0.994
$k = 4$	0.966	0.998	0.997

Notes: Based on transcript data from Cornell University in Fall 2019.

will share a classroom with 529 other classmates, given attendance is rarely 100 percent. On the other hand, most classes meet more than one time per week, giving each student multiple opportunities to be in the same room as co-enrolled students.

The weighted clustering coefficient for the projected one-mode student-to-student network is 0.480. This means that when student A has at least one class with student B, and student B has at least one class with student C, it is also frequently the case that student A has a class with student C. In a randomly wired Erdős-Rényi network of the same size and density as this network, one would expect to see a clustering coefficient of 0.024 (equal to the overall density of the network); the clustering in the observed network is more than an order of magnitude greater. This is not surprising, given that students commonly move through their major and distribution requirements with sets (cohorts) of fellow students. Despite high clustering, the entire network constitutes close to a single component, with only 1.2 percent of nodes (courses and students) outside the main component.

Most students in the full enrollment network can reach each other through very few steps. The average student-to-student geodesic distance in the projected one-mode network from the main component of the two-mode network is 2.5. (Geodesic distances are the shortest paths between a given pair of students.) This means that students are, on average, 1.5 courses removed from each other. Although it is unlikely that any two randomly chosen students would be enrolled in the same course, it is likely that they would be enrolled in different courses that both include the same third party. By definition, the percentage of pairs of students who are connected in one step is the same as the measure of network density: 2.4 percent.

Going one step beyond, to $k = 2$, the majority (59.4 percent) of pairs of students can reach each other via their connections to one shared classmate. This percentage increases to 92.1 percent at three steps and to 96.6 percent at four steps.

High clustering and short average path lengths mean that the student-to-student network created through course enrollment is a small-world network (Watts and Strogatz 1998). Even so, some students are many steps from each other. The network has a diameter of 10, meaning there are 10 steps between the most distant pair of students; visually, this is also the distance between the opposite sides of Figure 1. In some sense, it is more remarkable that a third-year law school student is indirectly connected to a first-year student majoring in fiber science and apparel design than that they are connected by up to 10 steps. As we show in the next section, this relatively large network diameter is driven by the inclusion of professional and graduate students in the university network.

Students in the university network are also typically connected to each other via multiple pathways that are independent of each other, what Moody and White (2003) call "embedded ties." Fully 94.5 percent of the students are in the largest bi-component of the two-mode network, which is the substructure that would require the removal of at least two nodes (course or student) before it breaks apart— that is, before any pair of nodes becomes unreachable. Put differently, nearly all students are connected through multiple independent pathways, and there is no single course and no single student whose removal from the substructure would eliminate the potential for mutual indirect exposure between any pair of students.

Undergraduate and Liberal Arts College Networks

The subnetwork of undergraduates is depicted in Figure 2, which uses the same coloring scheme as Figure 1 to identify students of different majors and courses with 100 or more undergraduates. The undergraduate network has a "stadium" structure, with students arrayed around a set of common large courses in the middle cluster. These large courses are often lower-division courses used by students in various majors to meet distribution requirements. Some large courses also appear in field-specific areas of the graph, which tend to be gateway courses to a particular major. More advanced and major-specific courses appear around the periphery of the network.

The clustering of students by field of study is, if anything, more apparent in the undergraduate network than in the university network. In Figure 2, there are distinct regions with high concentrations of students in STEM (orange), social sciences (dark blue), and law and business (light blue; at the undergraduate level, these are all business-related majors). The multidisciplinary or mixed majors (red) are concentrated in the space between the sciences and social sciences, whereas the humanities, arts, and design (yellow) and undeclared (green) majors are dispersed throughout the network.

The undergraduate network exhibits the same "small world" attributes as the university network but in exaggerated form (see Table 2, column 2). In the two-mode undergraduate network, nearly all (99.9 percent) of the students are in the main component. The student-to-student network projected from this two-mode
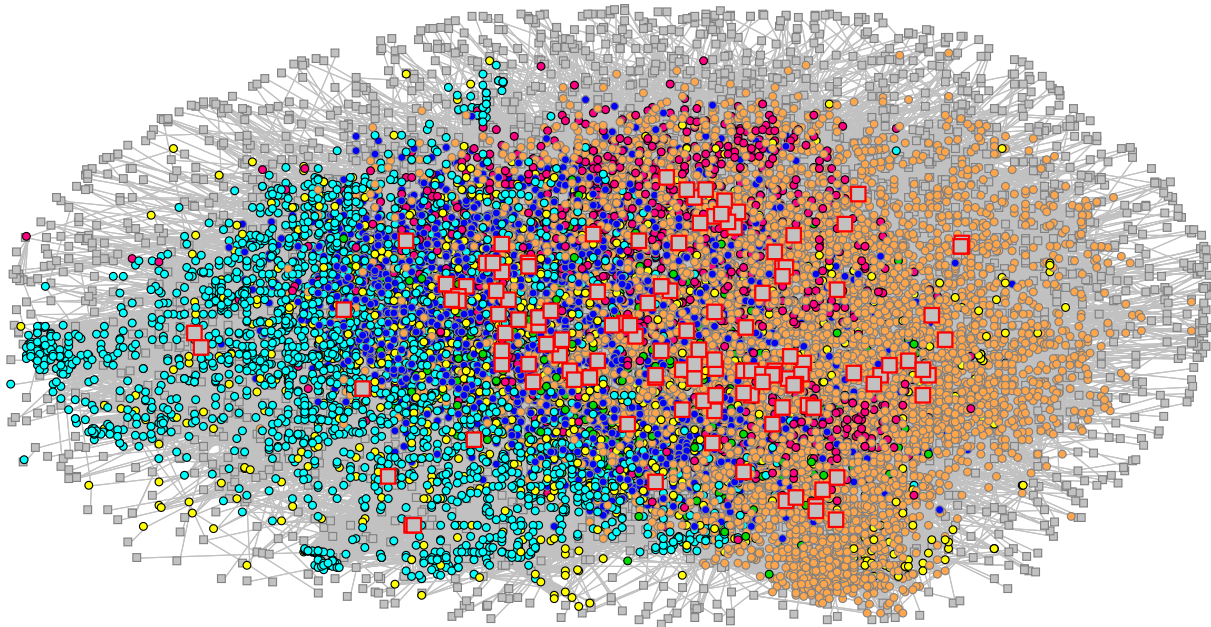
Figure 2: Undergraduate network. *Notes*: This network layout was arranged using the Fruchterman-Reingold algorithm in Pajek64 5.08 (Batagelj and Mrvar 2018), and nodes were colored, sized, and shaped using Netdraw (Borgatti 2002). We moved some extended course pendants closer to the main structure manually and resized the diagram to amplify student positions but did not move any student nodes. Light gray squares represent courses, and larger gray squares with red borders indicate courses with 100 students or more enrolled. Students are represented by small circles with colors identifying their major(s): yellow = humanities, arts, and design; dark blue = social sciences; orange = STEM; red = multidisciplinary/mixed; green = undeclared; light blue = business and law. Students' enrollment in particular courses is indicated with light gray lines. This diagram excludes 53 nodes that were not connected to this main component.

network shows slightly higher density (0.042) but similar clustering (0.460) as the projected student-to-student network that includes graduate and professional students. In one cycle of their schedule, the average undergraduate student will be co-enrolled with about 619 (4,582,808*2 bi-directed arcs/14,811 students = 619) other undergraduates. The number of students the average undergraduate actually encounters will, as in the university network, depend both on attendance patterns and on the number of times per week a given course meets.

In the undergraduate network, the average geodesic path length connecting pairs of students is 2.1, which is slightly shorter than the average path length in the university network (2.5). On average, any randomly chosen pair of undergraduates will be connected to each other by just one step (i.e., a third student). The longest path (network diameter) in the undergraduate network is substantially shorter than that in the university network: four steps instead of 10. Only a small percentage (4.2 percent) of pairs of undergraduate students take the same class, but 89.6 percent of student-pairs take a class with a third student who connects them indirectly, and 99.8 percent of student-pairs are connected after three steps.

Figure 3 depicts the subset of undergraduate students enrolled in courses offered by CAS, the liberal arts college. This network has significantly fewer large courses, depicted as squares with red borders, than the other networks. This is due both to the nature of the disciplines and pedagogies within the college (e.g., humanities courses are often taught as small seminars) and to the exclusion of non-CAS students in CAS courses from the analytic sample. These large courses tend to be interposed between the social science (dark blue) and STEM (orange) regions.

Field-specific clusters of students are evident in the liberal arts network but not as distinctly as in the university and undergraduate networks. This is what one might expect in a liberal arts setting in which emphasis is on cross-disciplinary exploration. Humanities and arts students (yellow) and multidisciplinary and mixed majors (red) are scattered throughout the liberal arts network.[3] Students who have not declared majors (green), who by definition are first- and second-year students, tend to be toward the center of Figure 3 and adjacent to larger, more general, and lower-level courses.

The liberal arts college network, like the other two networks, effectively constitutes a single component, with 99.8 percent of students in the main component (see Table 2, column 3). With a weighted clustering coefficient of 0.503, clustering in the projected student-to-student network of liberal arts college students is greater than in the other two networks. A given student in the liberal arts college will share a course with about 150 unique CAS students through one cycle of course, assuming perfect attendance.

In the liberal arts college, the average path length (2.3), longest path length (5), and share of student-pairs who can reach each other through k steps are consistent with the other two networks. Although only 3.4 percent of student-pairs are connected directly by being in the same class, 68.6 percent are connected indirectly in two steps and 99.4 percent by three steps. Furthermore, liberal arts students are connected via many alternative pathways: the largest bi-component contains 92.6 percent of the students. In short, the liberal arts college network, like the university and undergraduate networks, constitutes a small world in which students are not only connected to each other with short path lengths but also through multiple paths.

Large Courses and the Viability of Hybrid Models

In each of the three networks depicted in Figures 1 to 3, large courses serve as hubs, tying together students from different majors and network regions. In these data, 174 courses enrolled at least 100 students, and the largest course enrolled more than 900 students. It only takes one large class on a student's schedule to increase their potential exposure to other students exponentially, even if actual exposure in the sense of sharing physical proximity is affected by both room size and attendance (see discussion). Because many large courses tend to fulfill distribution requirements for graduation, they draw students from different disciplinary backgrounds and at different stages of advancement through their degrees, thus enhancing their role as hubs in the enrollment network.
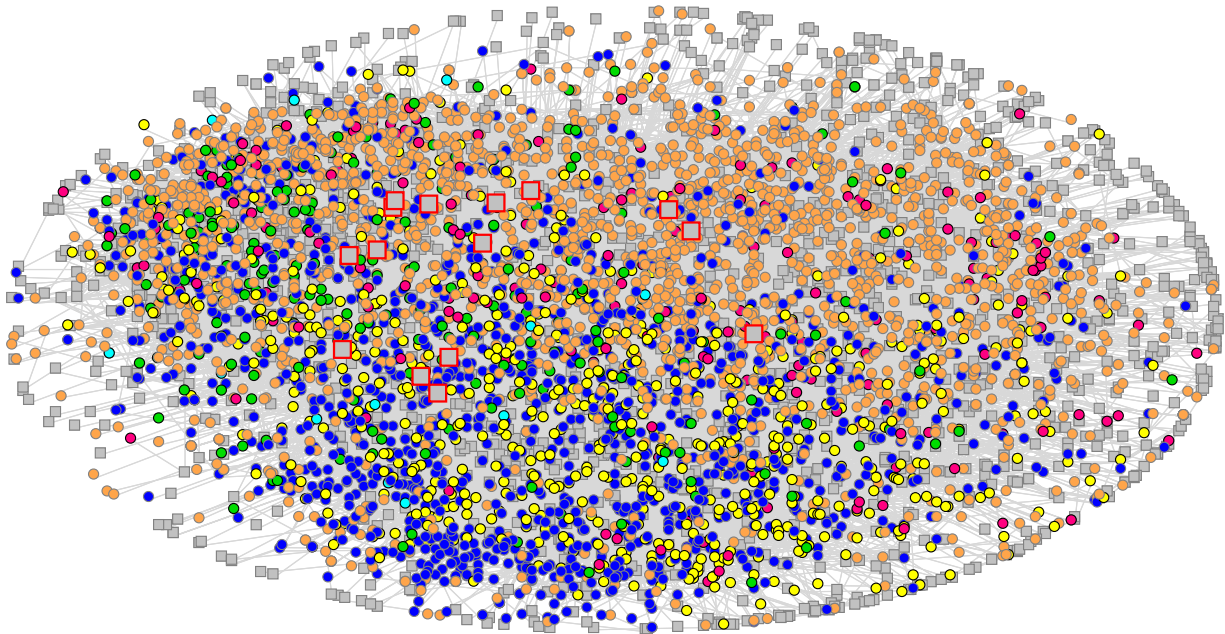
Figure 3: Liberal arts network. *Notes*: This network layout was arranged using the Fruchterman-Reingold algorithm in Pajek64 5.08 (Batagelj and Mrvar 2018) and nodes were colored, sized, and shaped using Net-draw (Borgatti 2002). Light gray squares represent courses; larger gray squares with red borders are courses of ≥ 100 students. Students are represented by small circles with colors identifying their major(s): yellow = humanities, arts, and design; dark blue = social sciences; orange = STEM; red = multidisciplinary/mixed; green = undeclared; light blue = business/law. Students' enrollment in particular courses is indicated with light gray lines. This diagram excludes nine nodes that were not connected to this main component.

To what extent do these large courses affect path lengths, the share of students who can be "reached" in a given number of steps, and the share of students and courses connected through multiple pathways? To assess this, we simulate networks in which we remove courses, as might occur in a hybrid mode of instruction in which some courses are taught online and others face to face (see Table 3). Each simulation assumes that student enrollment behaviors do not change—that is, the same students who would have enrolled in a large course offered through face-to-face instruction would also enroll in it if offered online. The simulations differ in the size of the courses we remove, where this refers to the counts of students in the relevant analytic sample. For example, courses of 100 or more in the liberal arts college refer to courses with 100 or more liberal arts students, even if students from other colleges might also be enrolled. The university network results (see panel A of Table 3) are cleaner in the sense that all enrollments are "counted," so we focus more attention on them in our discussion.

Table 3 shows that the lower the course size threshold in the simulated network, the greater the impact on connectivity within the network. In the university network, the percentage of students in the contact network, meaning students who are connected to *any* other students, declines from 96.8 percent when only courses with 100 students or more are removed from the network to 79.4 percent when

Table 3: Simulated enrollment networks in hybrid models with some courses

Size of courses removed from network	% of students in contact network	% of students in largest component	% of students in largest bi-component	Betweenness centralization	Average path length between students[a]	Proportion of all student-pairs reachable in k steps				
						$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Panel A: University (n = 22,051 students)										
None (observed)	100.0	99.1	94.5	0.098	2.47	0.024	0.593	0.921	0.966	0.971
≥100	96.8	94.9	83.8	0.012	2.94	0.005	0.181	0.775	0.891	0.899
≥75	95.1	93.2	79.9	0.013	3.06	0.004	0.123	0.704	0.854	0.867
≥50	90.8	87.5	68.0	0.008	3.28	0.002	0.061	0.510	0.747	0.765
≥40	86.5	83.0	60.0	0.011	3.46	0.002	0.039	0.365	0.656	0.687
≥30	79.4	76.3	48.6	0.011	3.75	0.001	0.021	0.207	0.504	0.576
Panel B: Undergraduate (n = 14,811 students)										
None (observed)	100.0	99.9	99.8	0.096	2.06	0.042	0.896	0.998	0.998	NA[b]
≥100	97.9	97.6	90.0	0.014	2.68	0.008	0.307	0.941	0.953	0.953
≥75	96.5	96.2	85.2	0.011	2.82	0.006	0.202	0.887	0.925	0.925
≥50	92.7	92.1	73.5	0.010	3.05	0.003	0.100	0.705	0.846	0.848
≥40	89.9	88.7	66.6	0.010	3.20	0.003	0.068	0.566	0.780	0.786
≥30	84.3	83.4	57.1	0.007	3.43	0.002	0.041	0.375	0.673	0.695
Panel C: Liberal arts college (n = 4,434 students)										
None (observed)	100.0	98.6	92.6	0.115	2.28	0.034	0.686	0.994	0.997	0.997
≥100	98.8	96.9	86.7	0.044	2.54	0.018	0.448	0.956	0.971	0.971
≥75	97.2	97.2	81.7	0.036	2.69	0.013	0.317	0.902	0.939	0.939
≥50	94.6	94.0	73.5	0.028	2.91	0.008	0.178	0.778	0.882	0.884
≥40	92.3	91.5	68.1	0.028	3.07	0.007	0.122	0.658	0.832	0.837
≥30	88.6	87.3	60.9	0.030	3.26	0.004	0.079	0.499	0.748	0.762

Notes: Based on transcript data from Cornell University in Fall 2019. [a] Limited to student-pairs that can reach each other through a path. The distance between student-pairs that are unreachable is undefined. [b] Not applicable. All student-pairs are connected in fewer than five steps.

courses with 30 students or more are removed. The share of students in the main component of the network likewise declines from 94.9 percent at a 100-student threshold to 76.3 percent at a 30-student threshold. For context, the median course at Cornell enrolls eight students, and the 90th percentile course enrolls 45 students (see Table 1); courses of 100 students and 30 students fall at approximately the 97th and 84th percentiles, respectively.

Table 3 also shows the reduction in connectivity of the students in the simulations using lower enrollment thresholds. The average path length connecting students in the contact network increases from 2.5 in the observed data to 2.9 in the 100-student threshold model and 3.75 in the 30-student threshold model. The percentage of student-pairs who are enrolled in courses ($k = 1$) is low in all scenarios. The percentage of student-pairs who do not share any classes together but who share at least one class with a third student in common ($k = 2$) declines from 59.3 percent in the observed data to 18.1 percent using the 100-student threshold and to less than 10 percent in simulations using thresholds of 50 students or fewer. These percentages increase dramatically at three steps out under all simulated thresholds: at the 50-student threshold, roughly half (51.0 percent) of student-pairs can reach each other, and at the 30-student threshold, just more than one-fifth (20.7 percent) can. By four steps out, the majority of student-pairs can reach each other in all scenarios.

Betweenness centralization also declines as course enrollment thresholds become more aggressive (see Table 3). However, the largest drop occurs between the observed network (0.098) and the simulated network in which the largest courses are removed (0.012; see Table 3). To some extent, this reflects the normalizing effect of centrality measures when variation in node degree (in this case, course size) is suppressed. However, the much higher centralization in the observed network also reflects the outsized role of very large courses in creating short paths between students who might otherwise be only distantly connected.

We also assess the robustness of connections between students under different hybrid model scenarios, meaning the extent to which multiple independent paths connect the same pair of students. In the observed data, the largest bi-component in the university network contains 94.5 percent of the students in the network. This drops to 83.8 percent when courses of 100 students or more are removed from face-to-face instruction, and to 48.6 percent when courses of 30 students or more are removed (see Table 3). Removing large or moderate-sized courses from face-to-face instruction still leaves most students connected to each other through multiple independent pathways.

We find a similar pattern of results for simulations of hybrid instructional models fit to the undergraduate and liberal arts college networks (see Table 3, panels B and C). In both sets of simulations, the lower the enrollment threshold at which a course is removed from the network, the more elongated the path lengths and the lower the percentage of students who share a class together ($k = 1$) or are connected by one intervening step ($k = 2$). A threshold of 30 students reduces the percentage of student-pairs connected by three steps to 37.5 percent in the undergraduate network and to 49.9 percent in the liberal arts network. By four steps, even the 30-student

threshold leaves 67.3 percent of pairs of undergraduates and 74.8 percent of pairs of liberal arts undergraduates connected.

In both networks, indirect connectedness is also quite robust in the simulations. For example, 57.1 percent of undergraduates remain connected within large bi-components even under the threshold of 30 students. This percentage is slightly higher when we restrict the analysis to the liberal arts college. Finally, we see the same pattern of betweenness centralization as in the full university network: the nodes that play the biggest role in linking together the network through short paths are the largest courses. However, the level of betweenness centralization at a 30-student threshold (0.007) is half that of the level of betweenness centralization at a 100-student threshold (0.014). This shows that midsized courses contribute in a nontrivial way to the presence of short transmission paths in the network.

Discussion

Course enrollments expose students to large numbers of other students throughout one cycle of a schedule, and they create multiple short chains of connection between students that can potentiate the spread of a virus through college campuses. The enrollment networks that connect students to classes, and to each other, are classic "small worlds" characterized by short path lengths and high clustering. Although all three of the networks we examined exhibit these "small world" attributes, connectivity is greater in the undergraduate and liberal arts college networks than in the network including graduate students.

Connectivity in these networks remains high when one excludes the very largest courses from the network, and modest under lower enrollment thresholds. For example, when courses with 100 students or more are removed from the university network, the percentage of student-pairs that can reach each other in three steps declines from 92.1 percent to 77.5 percent; when courses of 30 students or more are removed, this percentage declines to 20.7 percent. However, even under the lower enrollment threshold, the majority (50.4 percent) of students are connected in four steps. Other analyses show that the number of alternative pathways by which students are (indirectly) connected remains high even when large and moderately sized courses become virtual, especially in the undergraduate-only and liberal arts college networks.

At some level, it is inevitable that course enrollment networks will be "small worlds." High clustering (transitivity) and short path lengths among students in a single-mode network projected from a two-mode network is expected wherever: (1) the distribution of courses by size is broad and right-skewed, with the median course enrolling fewer students than the mean course (see Table 1); (2) the distribution of the number of courses each student takes is narrower and approximates a Poisson distribution; and (3) student-pairs take a least two courses together (a "four-cycle") or three students create a closed triad through three shared courses (a "six-cycle"), perhaps because they are in the same major (Vasques Filho and O'Neale 2020). One implication is that to maximally reduce the potential for the transmission of a virus among students through face-to-face instruction, universities would need to offer enough small courses that the mean course size is approximately equal

to the mean number of courses a student takes.[4] This would require teaching a very large share of courses online or, alternatively, "breaking up" larger face-to-face courses into many smaller sections.[5]

There are several important limitations of this study. First, and most obviously, we cannot assess whether our results generalize to other educational institutions. Even universities of similar size differ in the number and type of majors that are offered, the extent to which majors and courses are concentrated in a single college or multiple colleges, and a college's autonomy in setting curricular requirements. All of these may affect specific patterns of course enrollments. Still, we believe the "small world" character of each of the three networks in the Cornell data is likely to generalize insofar as most universities have similar distributions of course sizes (i.e., broad and right-skewed), similar distributions of numbers of courses per student, and similar network substructures created by students in the same major taking the same courses (Vasques Filho and O'Neale 2020).

Second, course enrollment networks understate the social and physical connections among courses and students, especially on residential campuses but even on nonresidential or low-residential campuses. They do not capture connections between courses that are created by sharing an instructor or a common classroom space, which, depending on the virus, could affect transmission through contaminated surfaces. They do not capture the incidental contacts that occur between students in hallways, on quadrangles between class periods, in libraries, or in the commercial areas that surround most colleges or universities. And, crucially, course enrollment networks do not capture the many ways that students, particularly on a residential campus, are connected through friends, parties, athletics, co-curricular and extra-curricular activities, and living situations. Given the multiplex ways that students come in contact with each other outside the classroom, the results of this article are a conservative depiction of the extent to which college campuses facilitate contact among students.

At the same time, our analysis likely overstates the density of the networks through which a virus might be transmitted through shared enrollments. Two students who are enrolled in the same large lecture course may never come in close physical proximity to each other. Similarly, classes, particularly large ones, rarely achieve full attendance. Because multiple pathways connect pairs of students, low attendance or large physical distances separating students in any single class may not have an appreciable effect on a given student's risk of exposure. Even so, future work should consider factors such as square footage of seating within classrooms and attendance rates to fine-tune estimates of the features of course enrollment networks relevant to viral spread.

Finally, the network we have analyzed here is a static, aggregate representation of the overall course enrollment record. It does not capture the dynamics of movement between courses, and it ignores the sequence in which students attend each of their classes. This temporal sequencing may be consequential for epidemic spread. For example, if larger classes tend to meet earlier in the week and smaller classes and discussion sections later in the week, an infection that arrives with a student after a weekend away from campus may spread more quickly. Classes that meet three times a week offer more chances of repeated exposure than those that meet

only once a week and more opportunities for all students to be in contact with each other despite one-time absences. Future research should undertake a more detailed analysis of these temporal dynamics and their implications for the potential of an epidemic spread.

Initial reactions to universities' decisions to shift to online instruction were mixed, with some observers lauding it as a necessary step in the fight against the diffusion of COVID-19 and others criticizing it as an overreaction. Despite the limitations of our study, our results suggest that the former view is closer to the mark. The very same "small world" networks among students that in normal times create an intellectually and socially vibrant campus experience can also increase the risk of an epidemic spread of a highly infectious disease. Although hybrid models of instruction, particularly those with more aggressive course enrollment thresholds, can elongate the paths connecting students to each other indirectly and reduce clustering, they do not "break apart" the small worlds of course enrollment networks. These models can reduce but not eliminate the risk of epidemic spread of a virus through the networks created by college classes.

Data Availability

The data for this article were used with permission from Cornell University. A simplified, de-identified version of the data will be posted on SocArXiv.

Ethical Compliance

We have complied with all relevant ethical regulations. The study is exempt from IRB review.

Notes

1 Cornell's business school, which is located in Ithaca, collaborates with off-campus entities (Cornell Tech; Cornell Weill; Queen's University in Kingston, Ontario; Tsinghua University in Beijing) to offer executive education and joint MBA programs. Most courses in these programs are offered in the partners' cities or through distance learning, and are excluded by our "no non-Ithaca courses" decision rule. For consistency, we also exclude the handful of executive education and joint MBA courses offered in Ithaca. This removes 256 student*course records out of a total of more than 118,000 records.

2 Network diagrams in a color scheme that is accessible to readers with some forms of color-blindness are provided in Appendix A of the online supplement.

3 There are only 14 law and business students (light blue circles) in the liberal arts college. Law is not offered as an undergraduate major, and students interested in pre-business programs typically enter or transfer to other colleges.

4 We are grateful to Demival Vasques Filho for pointing out this implication of Vasques Filho and O'Neale (2020) for our project via personal communication.

5 In theory, one could also think about identifying particular students who are highly connected to each other and removing these students from face-to-face instructional

network. Given the inequities this would entail, this doesn't seem feasible as a university policy.

References

Batagelj, Vladimir, and Andrej Mrvar. 2018. *Exploratory Social Network Analysis with Pajek*. 3rd ed. New York, NY: Cambridge University Press. https://doi.org/10.1017/9781108565691.

Biancani, Susan, and Daniel A. McFarland. 2013. "Social Networks in Higher Education." Pp. 151–214 in *Higher Education: Handbook of Theory and Research*, edited by M. B. Paulsen. Dordrecht, the Netherlands: Springer. https://doi.org/10.1007/978-94-007-5836-0_4.

Binson, Diane, William J. Woods, Lance Pollack, Jay Paul, Ron Stall, and Joseph A. Catania. 2001. "Differential HIV Risk in Bathhouses and Public Cruising Areas." *American Journal of Public Health* 91:1482–86. https://doi.org/10.2105/ajph.91.9.1482.

Borgatti, Stephen P. 2002. *NetDraw: Graph Visualization Software*. Harvard, MA: Analytic Technologies.

Borgatti, Stephen P., and Martin G. Everett. 1997. "Network Analysis of 2-Mode Data." *Social Networks* 19:243–69. https://doi.org/10.1016/S0378-8733(96)00301-2.

Breiger, Ronald L. 1974. "The Duality of Persons and Groups." *Social Forces* 53:181–90. https://doi.org/10.2307/2576011.

Caley, Peter, David J. Philip, and Kevin McCracken. 2008. "Quantifying Social Distancing Arising from Pandemic Influenza." *Journal of the Royal Society Interface* 5:631–39. https://doi.org/10.1098/rsif.2007.1197.

Cooley, Philip C., Sarah M. Bartsch, Shawn T. Brown, William D. Wheaton, Diane K. Wagener, and Bruce Y. Lee. 2016. "Weekends as Social Distancing and Their Effect on the Spread of Influenza." *Computational and Mathematical Organization Theory* 22:71–87. https://doi.org/10.1007/s10588-015-9198-5.

Cornwell, Benjamin, and Jake Burchard. 2019. "Cohesion and Embeddedness in Two-Mode Networks." *The Journal of Mathematical Sociology* 43:179–94. https://doi.org/10.1080/0022250x.2019.1606806.

Cornwell, Benjamin, and John A. Schneider. 2017. "Social Venue Range and Referral Chain Impact: Implications for the Sampling of Hidden Communities." *PloS One* 12:e0181494. https://doi.org/10.1371/journal.pone.0181494.

Dickison, Mark, S. Havlin, and H. E. Stanley. 2012. "Epidemics on Interconnected Networks." *Physical Review E* 85:066109. https://doi.org/10.1103/physreve.85.066109.

Feld, Scott L. 1981. "The Focused Organization of Social Ties." *American Journal of Sociology* 86:1015–35.

Feld, Scott L., and Bernard Grofman. 1977. "Variation in Class Size, the Class Size Paradox, and Some Consequences for Students." *Research in Higher Education* 6:215–22. https://doi.org/10.1007/bf00991287.

Freeman, Linton C. 1977. "A Set of Measures of Centrality Based on Betweenness." *Sociometry* 40:35–41. https://doi.org/10.2307/3033543.

Freeman, Linton C. 1979. "Centrality in Social Networks: Conceptual Clarification." *Social Networks* 1:215–39 https://doi.org/10.1016/0378-8733(78)90021-7.

Frost, Simon D. W. 2007. "Using Sexual Affiliation Networks to Describe the Sexual Structure of a Population." *Sexually Transmitted Infections* 83:i37–i42. https://doi.org/10.1136/sti.2006.023580.

Hill, Alison L., David G. Rand, Martin A. Nowak, and Nicholas A. Christakis. 2010. "Infectious Disease Modeling of Social Contagion in Networks." *PLoS Computational Biology* 6:e1000968. https://doi.org/10.1371/journal.pcbi.1000968.

Holland, Paul W., and Samuel Leinhardt. 1971. "Transitivity in Structural Models of Small Groups." *Comparative Group Studies* 2:107–24. https://doi.org/10.1177/104649647100200201.

Israel, Uriah, Benjamin P. Koester, and Timothy A. McKay. 2020. "Campus Connections: Student and Course Networks in Higher Education." *Innovative Higher Education* 45:135–51. https://doi.org/10.1007/s10755-019-09497-3.

Kelso, Joel K., George J. Milne, and Heath Kelly. 2009. "Simulation Suggests that Rapid Activation of Social Distancing Can Arrest Epidemic Development Due to a Novel Strain of Influenza." *BMC Public Health* 9:117. https://doi.org/10.1186/1471-2458-9-117.

Laumann, Edward O., Stephen Ellingson, Jenna Mahay, Anthony Paik, and Yousik Youm, eds. 2004. *The Sexual Organization of the City.* Chicago, IL: University of Chicago Press.

Luke, Douglas A., and Jenine K. Harris. 2007. "Network Analysis in Public Health: History, Methods, and Applications." *Annual Review of Public Health* 28:69–93. https://doi.org/10.1146/annurev.publhealth.28.021406.144132.

Miller, Joel C., and Istvan Z. Kiss. 2014. "Epidemic Spread in Networks: Existing Methods and Current Challenges." *Mathematical Modeling of Natural Phenomena* 9:4–42. https://doi.org/10.1051/mmnp/20149202.

Moody, James, and Douglas R. White. 2003. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." *American Sociological Review* 68:103–27. https://doi.org/10.2307/3088904.

Morris, Martina, ed. 2004. *Network Epidemiology: A Handbook for Survey Design and Data Collection.* Oxford, United Kingdom: Oxford University Press.

Niekamp, Anne-Marie, Liesbeth A. G. Mercken, Christian J. P. A. Hoebe, and Nicole H. T. M. Dukers-Muijers. 2013. "A Sexual Affiliation Network of Swingers, Heterosexuals Practicing Risk Behaviours that Potentiate the Spread of Sexually Transmitted Infections: A Two-Mode Approach." *Social Networks* 35:223–36. https://doi.org/10.1016/j.socnet.2013.02.006.

Oster, Alexandra M., Cyprian Wejnert, Leandro A. Mena, Kim Elmore, Holly Fisher, and James D. Heffelfinger. 2013. "Network Analysis among HIV-Infected Young Black Men Who Have Sex with Men Demonstrates High Connectedness around Few Venues." *Sexually Transmitted Diseases* 40:206–12. https://doi.org/10.1097/olq.0b013e3182840373.

Poletti, Piero, Marco Ajelli, and Stefano Merler. 2012. "Risk Perception and Effectiveness of Uncoordinated Behavioral Responses in an Emerging Epidemic." *Mathematical Biosciences* 238:80–89. https://doi.org/10.1016/j.mbs.2012.04.003.

Valdez, L. D., P. A. Macri, and L. A. Braunstein. 2012. "Intermittent Social Distancing Strategy for Epidemic Control." *Physical Review E* 85:036108. https://doi.org/10.1103/physreve.85.036108.

Valente, Thomas W. 2010. *Social Networks and Health: Models, Methods, and Applications.* Oxford, United Kingdom: Oxford University Press.

Vasques Filho, Demival, and Dion R. J. O'Neale. 2020. "Transitivity and Degree Assortativity Explained: The Bipartite Structure of Social Networks." *Physical Review E* 101:052305. https://doi.org/10.1103/physreve.101.052305.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications.* New York, NY: Cambridge University Press.

Watts, Duncan J., and Steven Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393:440–42. https://doi.org/10.1038/30918.

Kim A. Weeden: Department of Sociology, Cornell University. E-mail: kw74@cornell.edu.

Benjamin Cornwell: Department of Sociology, Cornell University. E-mail: btc49@cornell.edu.