

Modeling COVID-19 screening strategies at NYUAD: Report 1

Xuliang Guo, Oscar Wan, Shafer Smith, Francesco Paparella

July 2020

Abstract

We build a model describing the propagation of COVID-19 within a small, close-knit population of susceptible individuals, which can be taken as representative of the NYUAD campus. The model is used to assess the NYUAD COVID-19 screening strategy, evaluating how many potential infected individuals there might be within the community at the moment of the first detection. The results may be used to improve the screening strategy, compatibly with given constraints in terms of number of tested individuals, and frequency of the tests.

1 Introduction and motivation

During Summer 2020 New York University Abu Dhabi (NYUAD) began a COVID-19 screening initiative with the goal of monitoring “the NYUAD community for a potential spread of COVID-19, estimate the prevalence of asymptomatic cases and validate more efficient sample collection and detection methods” [6]. The initiative is currently in the pilot phase, and will scale up in the Fall as the classes will resume.

We are developing a model aimed at assessing the effectiveness of the COVID-19 screening protocol. The present first report illustrates the overall modeling strategy; documents a first, very simple version of the model; and gives some preliminary estimates of the number of infected individuals that may exist within the community at the time when the first COVID-19 case were detected as a function of the screening strategies.

2 The model

Most epidemiological models (e.g. the celebrated S.I.R. model) are based on ordinary differential equations or difference equations, where individuals are clumped together in *compartments* (e.g. Susceptible, Infected, Recovered) [3]. The model, then, gives the time evolution of the size of each compartment relative to the entire population being modeled. While this approach has many merits, we find it unsuitable for modeling small populations, where stochastic effects and the way the population is structured may have profound effects on the outcome of an outbreak.

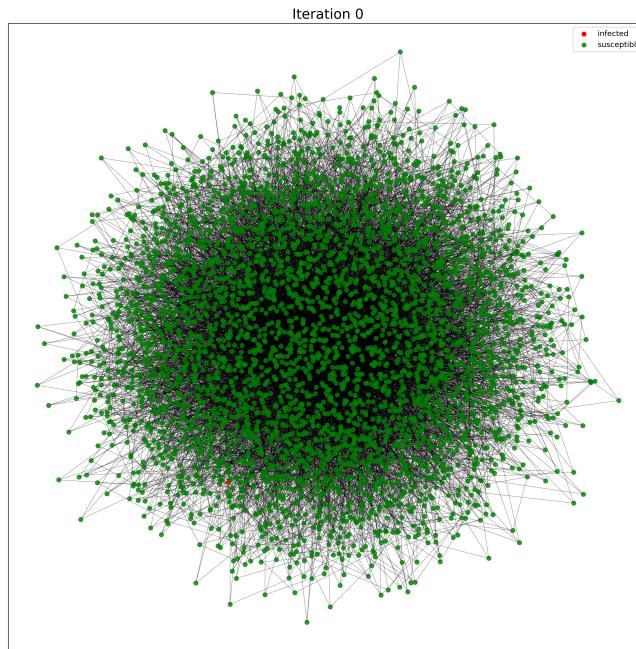


Figure 1: A Barabási–Albert network with 4000 nodes and $m = 3$ minimum edges per node. Note the large number of nodes with only three edges at the periphery. The center contains a handful of nodes having more than 100 edges.

Considering the limited number of individuals that compose the NYUAD community (we use the indicative number $N = 4000$) a modeling framework where individuals are explicitly represented becomes feasible. In our model each individual is a *node* joined to other individuals by *edges*. Pairs joined by an edge are assumed to come into some form of contact that may lead to contagion if one in the pair is infected. Pairs not joined by an edge represent individuals who never undergo any form to mutual interaction that can potentially lead to contagion. Thus, for example, a student who shops at the convenience store and the store cashier would be joined by an edge, but a faculty teaching remotely would not be joined

by an edge to a student who only sees her online. The set of all the nodes and of all the edges constitutes the *network* within which the disease can spread.

It should be obvious that reconstructing the real NYUAD network is impossible, short of a titanic and ethically dubious surveillance program. Guessing an approximation to the real NYUAD network is an exercise where uncertainty quickly escalates to unbearable levels as one attempts to achieve realism. Therefore, rather than focusing on a single arrangement of the network, we shall consider (over the course of several reports) many different kinds of arrangements, with increasing complexity, exploring how the results change depending on the network properties.

In this report, the 4000 nodes of the simulated NYUAD network are attached together in a so-called *scale-free* network, built using the preferential attachment algorithm of Barabási and Albert [1]. The defining property of such a network is that a very small number of nodes has a very high number of edges, and progressively higher number of nodes have a progressively smaller number of edges. More precisely, in a Barabási–Albert network the probability p_k that a given node has k edges, or *degree distribution*, is $p_k \approx k^{-3}$. Once the number of nodes is fixed, the only free parameter is the minimum number m of edges that a node can have. In this report we use $m = 3$, which is representative of the size of a small household. A visual representation of a Barabási–Albert network with $N = 4000$ and $m = 3$ is given in Figure 1. Although no human network should be expected to be exactly a Barabási–Albert network, evidence exists that in some academic communities the degree distribution of close-distance interactions among individuals follows a power-law [8].

Mimicking the traditional S.I.R. model, each node of the network is assumed to be in one of the following states: Susceptible, Infected, Removed. For simplicity, we have not (yet) included an incubation period, therefore any infected node is also assumed to be infectious (that is, able to propagate the disease). The spread of the disease is modeled as follows [3, 4, 5]: *i*) if a node is infected, each day any susceptible node connected with an edge to the infected one has a probability b to turn to the “infected” status; *ii*) any infected node turns to the “removed” status after $D = 5$ days. “Removed” nodes should be intended to represent individuals that either have spontaneously recovered from the disease, or have manifested symptoms and have been put in quarantine and are thus unable to further spread the contagion. Therefore, “infected” nodes must be intended as asymptomatic or pauci-symptomatic individuals that are unaware of having contracted the disease.

Lacking any direct information on the transmission probability, at this stage we feel that it is best to keep the value b constant across the whole network, and explore how the outcome changes for varying values of b .

3 Results

The model is initialized by setting all nodes to the “susceptible” state, except one, chosen with uniform random probability, which is set to the “infected” state. In the following we will first show how an outbreak is expected to run its course in the absence of any external measures; then we will assess how many cases should be expected to be present, on average, in the network when the first positive is detected; finally we will explore how that number changes as a function of the number of connections k of the first infected (patient zero).

3.1 Simulations of an outbreak

We first present the results of simulations where no tests are performed and the outbreak is left running its course. In Figure 2, as a function of time, we show the number of susceptible, infected, and removed individuals averaged over 100 re-runs (i.e. *realizations*) of the same simulation (thick lines). In this stochastic process, means need not be representative of individual realizations. Therefore, we also show the number of infected and the number of total cases (i.e. infected + removed) in each individual realization as thin lines (susceptible and removed are omitted to avoid visual clutter).

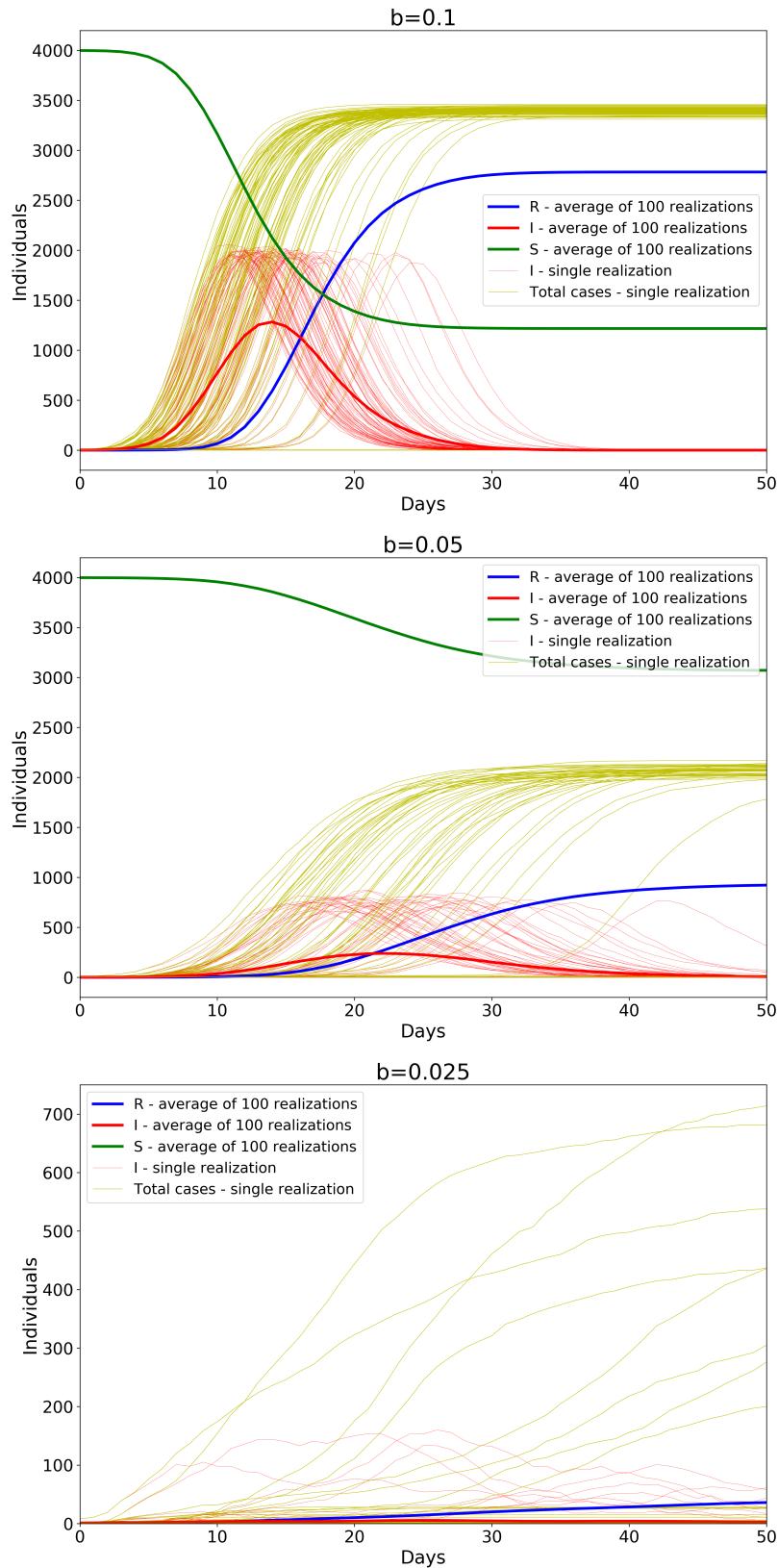


Figure 2: Thick lines show the number of susceptible (green), infected (red), and removed (blue) individuals averaged over 100 realizations of the same simulation. Thin red lines and the yellow red lines show, respectively the number of infected and the total number of cases (infected + removed) of each realization. The three groups of simulations differ only for the value of the transmission probability b . Note that in the lower panel ($b = 0.025$) the extent of the vertical axis is severely reduced with respect to the other two cases, and the curve of the susceptible is not visible.

It is evident that in some realizations the outbreak doesn't really occur: the initially infected node transmits the disease to at most one or two additional nodes before they all transit to the “removed” status: the contagion does not propagate. Depending on the transmission probability, if the disease reaches more than a handful of nodes, then it doesn't stop until it reaches a fairly-well defined maximum, which is a function of the transmission probability (about 2000 nodes when $b = 0.1$, about 750 nodes for $b = 0.05$ and roughly 100 nodes for $b = 0.025$). The timing of the outbreak is also very variable. In some realizations the outbreak involves only a few nodes for several days, or even a few weeks, before the beginning of the phase of rapid, exponential-like growth. As a consequence, the interval of time between the first infection and the peak of the outbreak is highly variable. For $b = 0.025$ only about 10% of the realizations lead to an outbreak, which is small in magnitude and highly erratic in its time evolution. The time variability between the onset and the peak is highly enhanced at such low value of b . It is important to stress that all this variability among individual realizations is due to chance alone, and cannot be attributed to any structural difference between realizations with the same value of b . As we shall illustrate below, an important factor affecting the rapidity of the spread of the disease is the number of connections of the first infected node.

The probability of an outbreak drops rapidly with b . For $b \simeq 0.013$ or smaller the probability of an outbreak becomes negligible. Below this threshold one should expect at most the presence of a few isolated cases whose presence, however, never snowballs into a massive contagion.

The table below reports the doubling time at the onset of the *average* outbreak for several values of b , as well as an estimate of the parameter R_0 obtained as the ratio between the overall disease transmissibility T and the percolation threshold T_c [4, 7] (see Appendix for details).

	$b = 0.025$	$b = 0.05$	$b = 0.075$	$b = 0.1$
Doubling Time	4.03 days	2.08 days	1.33 days	1.01 days
$R_0 = \frac{T}{T_c}$	1.85	3.52	5.03	6.65

3.2 The effectiveness of screening tests

We have designed a set of simulations aimed at assessing the effectiveness of the on-campus COVID-19 screening protocol. The simulations select a pool of Q nodes, representing the volunteers participating in the screening, and test the status of those nodes every P days. For simplicity, our virtual COVID-19 tests are considered to be perfect: the chance of false test results is zero. As for the simulations of the previous subsection, each run begins with all nodes being susceptible, except a randomly chosen one, which is infected. The simulation stops when, upon testing, any of the Q volunteers is found in the “infected” or “removed” status. Otherwise the simulation stops when the number of infected in the whole network drops to zero. Upon stopping the status of all the nodes in the network is recorded.

Figure 3 shows the number of infected and recovered (total cases) present in the network at the time of the first positive test result. This function is shown for testing periods P ranging from 2 to 28 days, and for $b = 0.1, 0.75, 0.5, 0.25$.

It is immediately evident that the total number of cases at the time of first detection is only weakly dependent on the size Q of the testing pool. Even with Q as low as 200 nodes the total number of cases is not substantially higher than with a testing pool of 1200 nodes. Only for very small sizes of the testing pool and relatively short testing intervals the results depend markedly on Q .

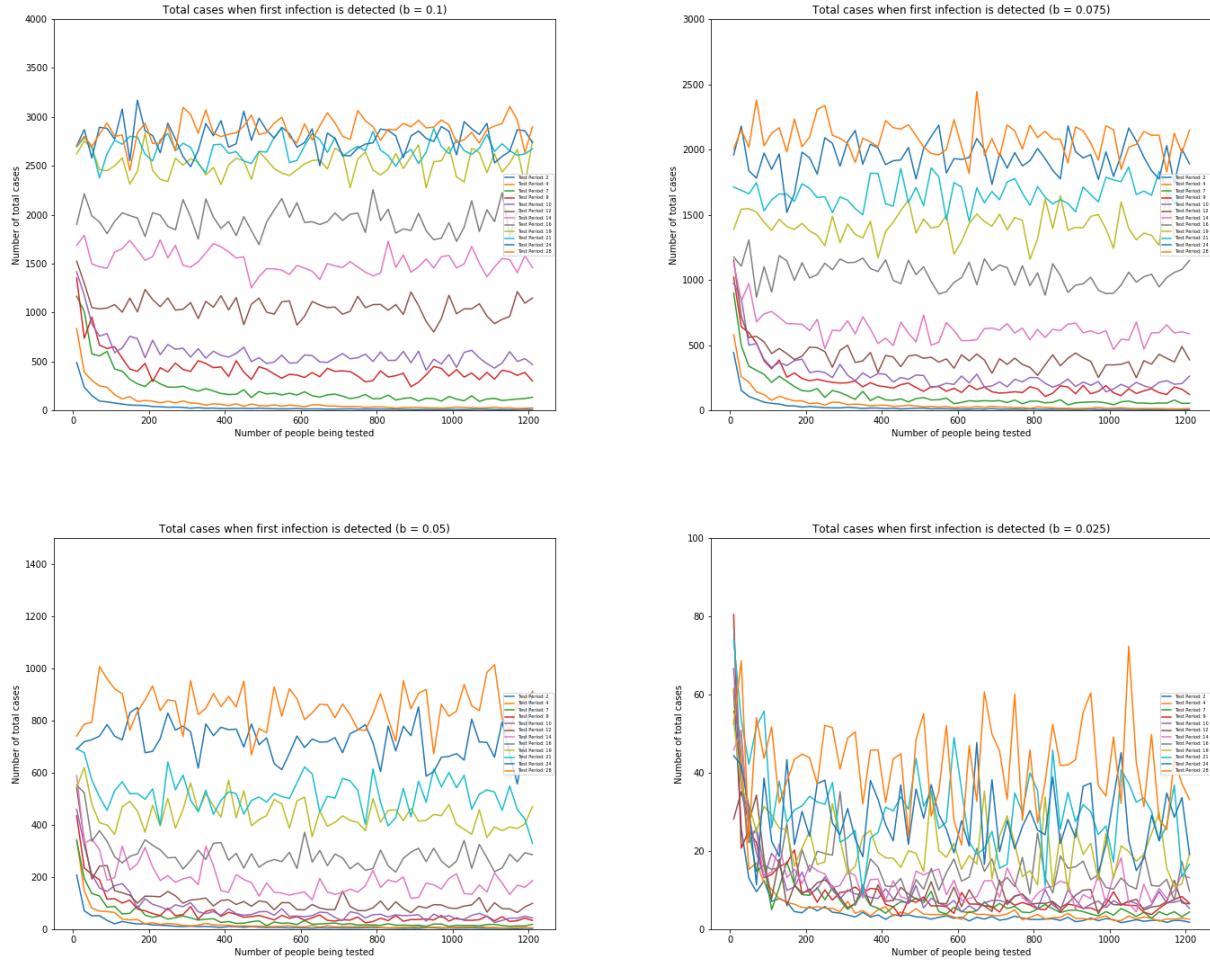


Figure 3: Number of total cases (infected + removed) present in the network at the time of the first positive test as a function of the number Q of people in the testing pool. Average over 100 realizations.

A second important information conveyed by Figure 3 is that the frequency of testing, instead, has a large impact on the total number of cases to be expected at the time of the first detection. Obviously, if too much time passes between a screening and the next, the disease has the possibility to spread undetected.

3.3 Impact of the connectivity of patient zero

An obvious factor affecting the likelihood of having a rapid spread of the disease is the number of edges. If the *patient zero* is a node with only few edges, there are sizeable chances that that node will transit to the “removed” state before being able to transmit the disease to others, especially if the transmission probability b has a low value. On the other hand, if *patient zero* is a node with tens or hundreds of edges, an outbreak may be inevitable.

Thus we ran a batch of simulations where the number k of edges of the first infected node was prescribed beforehand. These simulations use $b = 0.05$ and a size of the testing pool $Q = 200$. For each value of k , and for testing intervals of 7, 14, 21, and 28 days, figure 4 shows the probability of having a given number of total cases at the time of the first positive test.

At the relatively rapid testing interval of 7 days, the most likely number of cases present in the network at the time of first detection increases roughly proportionally with k . If k is small, then the probability distribution peaks at one (only *patient zero* is diseased) and

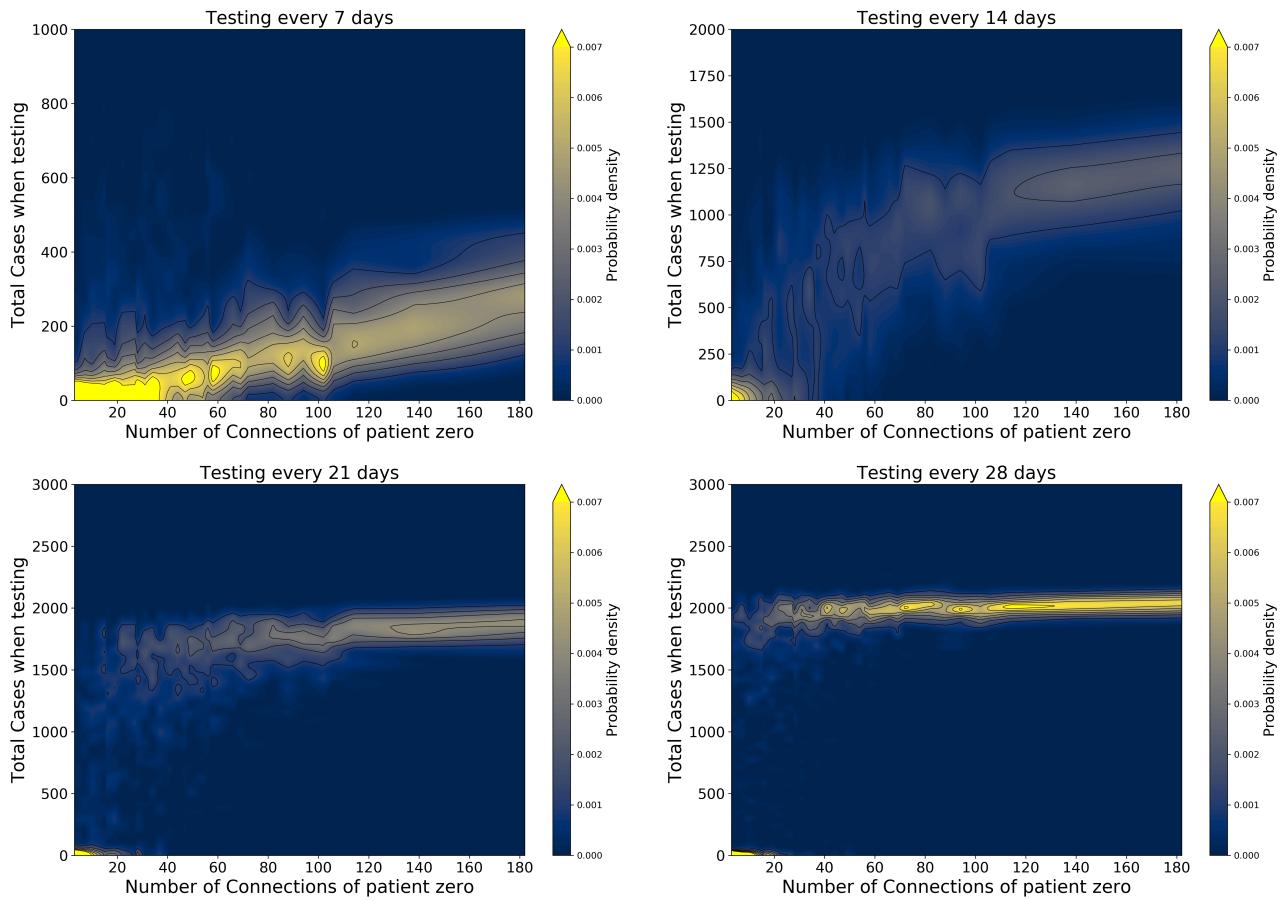


Figure 4: Probability density of having a given number of total cases (infected + removed) as a function of the number k of edges (connections) of patient zero at the time of the first positive test. Gaussian kernel density estimation over 100 realizations for each value of k , with $b = 0.05$, $Q = 200$.

decays for progressively larger numbers of infected, with a tail extending into the hundreds of cases. For larger k the distribution has a broad peak with two roughly symmetric tails.

When testing occurs every 14 days and k is small, then, as before, the probability of the number of cases at the time of first detection has a distribution that peaks at one, but with a longer tail. As k increases, the distribution becomes bimodal, with a first peak around one, and a second peak at 500 or more nodes, the further away from zero the higher the value of k . For even higher value of k the lower peak disappears.

For even longer testing intervals, at low values of k the distribution becomes more concentrated around one, corresponding to cases where the outbreak does not occur, and the cases where the outbreak occurs produce an extremely long and flat tail which raises to a second peak at values close to 2000 cases. For increasing values of k the distribution concentrates in the second peak, and the probability of having just a handful of cases drops to virtually zero when $k > 40$. Then the peak of the distribution approaches a value of about 2000 cases, which is the number of cases that one expects to have for an uncontrolled outbreak with this value of b (see the second panel of Figure 2).

4 Discussion

In this first report we have assumed a specific structure of the NYUAD interpersonal connectivity network: a Barabási-Albert network with $N = 4000$ nodes and $m = 3$ minimum edges. We have explored the properties of disease outbreaks produced by a single initial infected node for varying values of the transmission probability b . Albeit this is a grossly

simplified model of the NYUAD community, and of the dynamic of possible a COVID-19 outbreak, our simulations reveal a number of features that should mirror what may happen in reality, and that suggest concrete and actionable measures.

It is important to stress that the initial presence of a single infected member of the community may lead to essentially three kinds of outcomes (Figure 2).

First, the outbreak might not start. If *patient zero* is an individual who seldom comes into contact with other people, there are reasonable chances that the disease will not be transmitted before it runs its course. In fact, the presence of an infected individual conducting a relatively isolated life might very well go undetected.

Second, the outbreak might have a *slow fuse*. Starting from an individual with relatively few connections, all with other individuals with relatively low connections, it might take weeks for the disease to spread to a sizeable fraction of the community. If the transmission probability across the network is low, the time evolution of the outbreak may be particularly erratic, with completely random ups and downs in the number of the infected.

Third, the outbreak might spread exponentially rapid, right from the start. This is the expected outcome if the disease starts from a node with a large number of connections, or reaches such a node within a small number of hops. In this case, particularly for high values of b , the spread of the disease may very well prove to be quicker than our ability to react to it. It should not be forgotten that the average doubling time of the disease in the range of values of b that we have explored is always quite high (see table at the end of sec.3.1). Further lowering the value of b would cross the so-called *percolation threshold* below which an outbreak does not occur. For the network studied here the percolation threshold is $b \simeq 0.13$.

The average number of cases present in the network at the time of first detection is roughly independent of the size Q of the testing pool (Figure 3). This is a somewhat obvious outcome when the periodicity of the tests is much larger than the disease's doubling time: between one test and the next the disease may propagate to such a large fraction of the whole network that it doesn't take a large random sample to have a near certainty of picking at least one infected. For testing periods comparable with the doubling times this is a less obvious result and, in effect, for short testing periods the curves show some dependence on the size Q . We must stress that those shown in Figure 3 are averages, but except for very short testing periods the underlying distribution is bimodal with two very well defined peaks. Therefore, the total number of cases in a single realization may end up being much higher than that average.

Regardless of the underlying distribution, we feel safe by concluding that testing often is more important than testing extensively. In fact, if the volunteers pool is large, much will be gained by segmenting it into sub-pools to be tested in consecutive days. This is illustrated in Figure 5, where the same simulations of 3 are performed, except that the testing pool is divided in as many equal-size sub-pools, as is the testing period, and each day a different sub-pool is tested (results are assumed to be available the day after). For example, if the pool size is $Q = 1000$ and the testing period is $P = 10$, then each day is tested a distinct sub-pool of 100 nodes. Even for relatively small pool sizes, this sort of daily *staggered testing* appears to be much more effective at catching an outbreak in its earliest stage rather than the simultaneous testing of an even larger pool, spaced out by several days with no testing.

As mentioned above, (see also Figure 4) if the disease starts from (or reaches) nodes connected with a large number of edges, the expected outcome is a quick and potentially uncontrollable outbreak. Therefore, much care should be given, if at all possible, in monitoring specific individuals who are known to come into contact with many other individuals. In a Barabási-Albert network, these individuals would play the role of *hub nodes* through which the disease would be transmitted to a large number of other individuals.

5 Future work

This first report focuses on the properties of a Barabási-Albert network. Although this network structure should approximate actual human networks [8], this does not appear to

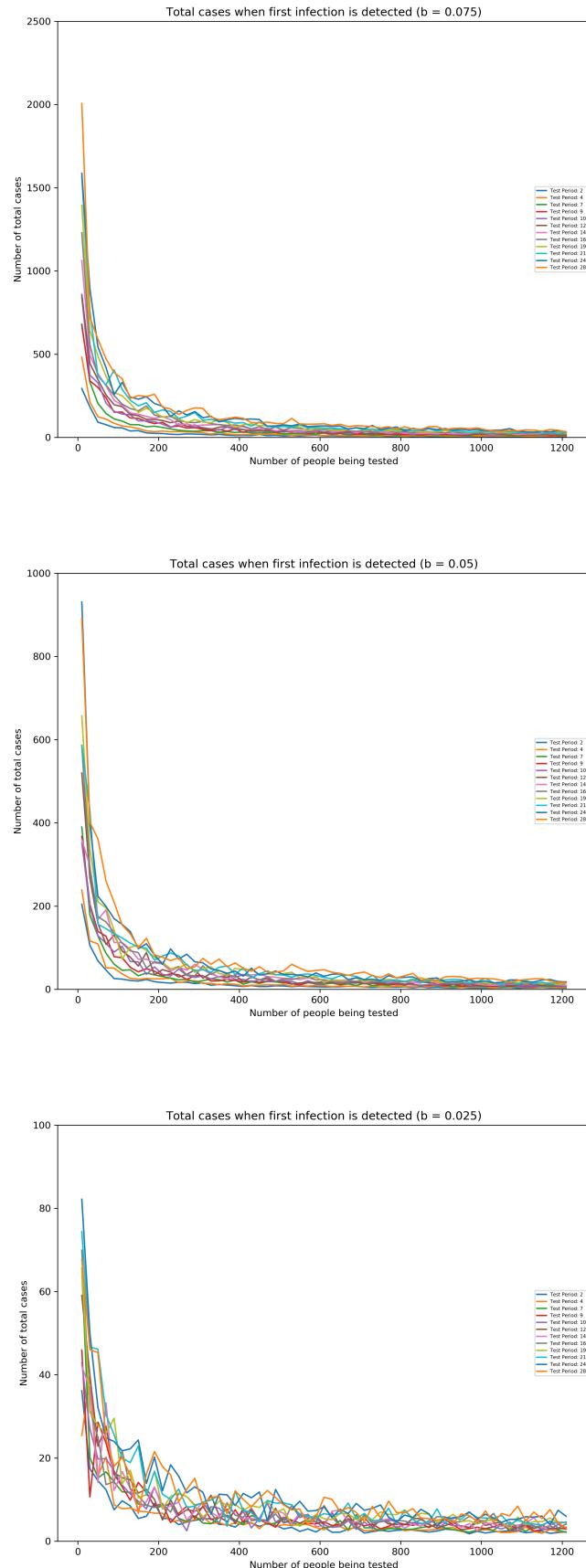


Figure 5: Number of total cases (infected + removed) present in the network at the time of the first positive test as a function of the number Q of people in the testing pool, assuming that every day a fraction Q/P of the is tested, where P is the test interval. Average over 100 realizations.

be a universal feature. In other cases, interpersonal contact networks have been reported to follow quite a different structure, where the degree distribution does not follow a power–law, but rather a Poisson–looking distribution [2]. Therefore, we feel necessary in a future report to explore the dynamics of the disease in a Erdős–Rényi network (a network structure in which the degree distribution follows a binomial distribution).

A reasonable enhancement of the model should be obtained by segmenting the network in two or three subnetworks, representing, e.g. contract workers, faculty and staff, students. The connectivity properties among the nodes, and the probability b of disease transmission may be different in each of the subnetworks.

Studying a network formed by distinct subnetworks would also allow us to study the impact of a non–randomly sampled testing pool. It may very well be that the Q volunteers may over–represent some category and under–represent another, with effects that we will aim at quantifying. Similarly, we’d like to contrast the situation where all or most of the hubs (nodes with the maximum number of connections) are in the testing pool, with that in which none of the hubs are in the testing pool.

In all of the present report we have always assumed that only one individual would contract the disease from external sources, and then spread it within the NYUAD community. At the moment of writing in the UAE less than one resident over 400 is infected, and the infection is waning. Even so, the probability that two or more NYUAD community members may become infected roughly at the same time should not be neglected. Thus we shall verify if and how this may change the results reported in the present document.

6 Appendix

The model parameters are specified in the table below.

Parameter	Symbol	Value
Daily probability of disease transmission along an edge	b	0.025, 0.05, 0.075, 0.1 days $^{-1}$
Duration of the “infected” status	D	5 days
Time step of the simulation	Δt	1 day
Number of nodes in the BA network	N	4000
Minimum number of edges of a node	m	3
Size of the test pool	Q	10, …, 1200 nodes
Testing interval	P	2, …, 28 days

In the limit $N \rightarrow \infty$ the degree distribution of a Barabási–Albert network with m minimum edges is

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \quad (1)$$

Figure 6 compares this asymptotic formula with the actual degree distribution of a realization of a network with $N = 4000$. The approximation is excellent, except for very high values of k , where the finite–size Barabási–Albert network shows a typical excess of nodes with high number of edges (hubs).

The theory of disease transmission over networks identifies two important non–dimensional parameters [3, 4, 7]. One is the *transmissibility*

$$T = 1 - (1 - b\Delta t)^{\frac{D}{\Delta t}} \quad (2)$$

which is the probability that a infected node transmits the disease to one of its neighbors (nodes connected through a single edge).

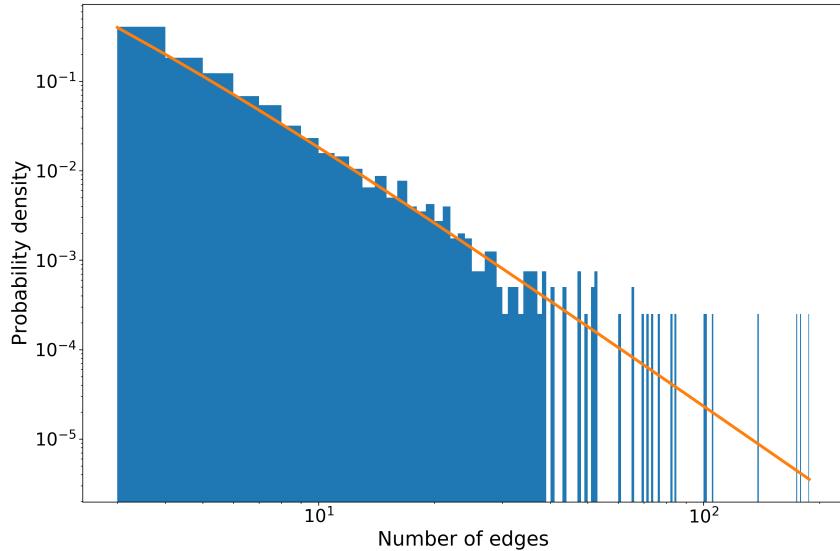


Figure 6: Degree distribution of one realization of the Barabási–Albert network with $N = 4000$ and $m = 3$ (blue bars) compared with the asymptotic formula (1) (orange line).

A second one is the *critical transmissibility*

$$T_c = \frac{\sum_k kp_k}{\sum_k k(k-1)p_k} \quad (3)$$

which was defined as the *percolation threshold* [4], that is, the minimal value of the transmissibility that can produce an outbreak.

Further analysis bring both quantities back to the familiar R_0 concept: the average number of nodes that are infected by an infected node at the onset of an outbreak [7]. The relation is quite simple:

$$R_0 = \frac{T}{T_c} \quad (4)$$

Observing that the second moment of (1) diverges, from (3) one finds that for a Barabási–Albert network $T_c = 0$ in the limit $N \rightarrow \infty$. For a network with a finite number of nodes this is not the case, because the degree distribution has a finite second moment, which can be evaluated numerically. This leads to the R_0 estimates reported in the table at the end of sec.3.1

References

- [1] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.
- [2] Hasan Guclu, Jonathan Read, Charles J Vukotich Jr, David D Galloway, Hongjiang Gao, Jeanette J Rainey, Amra Uzicanin, Shanta M Zimmer, and Derek AT Cummings. Social contact networks and mixing among students in k-12 schools in pittsburgh, pa. *PLoS One*, 11(3):e0151139, 2016.
- [3] Lauren Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.
- [4] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.
- [5] Lucas Daniel Valdez, Pablo Alejandro Macri, and Lidia Adriana Braunstein. Temporal percolation of the susceptible network in an epidemic spreading. *PLoS One*, 7(9):e44188, 2012.

- [6] Ayaz Virji, Youssef Idaghdour, Aisha Saeed Al Hamiz, Andrea Jabari, Fabio Piano, Kristin Gunsalus, Nizar Drou, Raghib Ali, and Sehamuddin Galadari. COVID-19 NYUAD Screening Pilot Study. Technical report, New York University Abu Dhabi, 05 2020.
- [7] Erik Volz and Lauren Ancel Meyers. Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Interface*, 6(32):233–241, 2009.
- [8] Eiko Yoneki, Pan Hui, and Jon Crowcroft. Wireless epidemic spread in dynamic human networks. In *Workshop on Bio-Inspired Design of Networks*, pages 116–132. Springer, 2007.